Sensitive Disclosures under Differential Privacy Guarantees

by

Chao Han

M.Sc., Dalian University of Technology, 2010

B.Sc., Dalian University of Technology, 2008

Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

in the
School of Computing Science
Faculty of Applied Science

SIMON FRASER UNIVERSITY

Spring 2016

Approval

Name:	Chao Han
Degree:	Doctor of Philosophy (Computing Science)
Title:	*Sensitive Disclosures under Differential Privacy Guarantees*
Examining Committee:	**Dr. Qianping Gu** (chair) Professor Computing Science Simon Fraser University
Dr. Ke Wang Senior Supervisor Professor Computing Science Simon Fraser University	_____
Dr. Anoop Sarkar Supervisor Associate Professor Computing Science Simon Fraser University	_____
Dr. Martin Ester Internal Examiner Professor Computing Science Simon Fraser University	_____
Dr. Li Xiong External Examiner Professor Department of Mathematics and Computer Science Department of Biomedical Informatics Emory University	_____
Date Defended:	15 Jan 2016

Abstract

Most syntactic methods consider *non-independent reasoning* (NIR) as a privacy violation and smooth the distribution of published data to avoid sensitive NIR, where NIR allows the information about one record in the data could be learned from the information of other records in the data. The drawback of this approach is that it limits the utility of learning statistical relationships. The differential privacy criterion considers NIR as a non-privacy violation, therefore, enables learning statistical relationships, but at the cost of potential disclosures through NIR.

In this thesis, we investigate the extent to which private information of an individual may be disclosed through NIR by query answers that satisfy differential privacy. We first define what a disclosure of NIR means by randomized query answers, then present a formal analysis on such disclosures by differentially private query answers. Our analysis on real life data sets demonstrates that while disclosures of NIR can be eliminated by adopting a more restricted setting of differential privacy, such settings adversely affects the utility of query answers for data analysis, and this conflict can not be easily resolved because both disclosures and utility depend on the accuracy of noisy query answers. This study suggests that under the assumption that the disclosure through NIR is a privacy concern, differential privacy is not suitable because it does not provide both privacy and utility.

The question is whether it is possible to (1) allow learning statistical relationships, yet (2) prevent sensitive NIR about an individual. In the second part of the thesis, we present a data perturbation and sampling method to achieve both (1) and (2). The enabling mechanism is a new privacy criterion that distinguishes the two types of NIR in (1) and (2) with the help of the law of large numbers. In particular, the record sampling effectively prevents the sensitive disclosure in (2) while having less effect on the statistical learning in (1). The data perturbation and sampling method are evaluated in real life data sets in terms of both sensitive disclosures and utility. Empirical results confirm that disclosures can be prevented with minor loss of utility.

Keywords: Data Privacy; Differential Privacy; Data Mining; Anonymization

Acknowledgements

Most of all, I would like to thank my senior supervisor Dr. Ke Wang for his insightful guidance. I am inspired by his passion on research. I also want to express my appreciation to my examining committee: Dr. Anoop Sarkar, Dr. Martin Ester, Dr. Li Xiong, thank you for your time and valuable suggestions.

To my former and current labmates: Zhihui Guo, Bo Hu, Peng Wang, Judy Yeh, Yongmin Yan, Zhensong Qian, Ryan Shea, Chenyi Zhang, Hongwei Liang, Weipeng Lin, Yue Wang, Zhilin Zhang, Jiaxi Tang, Yao Wu, Hao Wang, Beidou Wang, Xin Wang, Tong He, thank you for your constant support and encouragements.

Last, but certainly not least, I sincerely appreciate the constant love, support and understanding from my husband, Shaowei Wen, and my family!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Data Privacy

The burst of internet has brought us to an age of data. And the main data sources are the growing numbers of population, devices, and sensors connected by the internet. While data provides enormous value and benefits for the growth of global economy, it also brings significant privacy concerns. Privacy and security have been identified as a main challenge of publishing data [2]. For example, AOL released anonymized search logs for academic purposes, but searchers were easily re-identified by their queries, and AOL had to remove the data shortly due to the release [8]. A recent analysis of how companies are leveraging data analytics for marketing purposes showed that a retailer was able to identify that a teenager was pregnant even before her father knew [26]. Several other major privacy breaches have occurred in the past few years [61, 66, 86].

Nowadays, the field of data privacy has drawn many people's attention. What is data privacy? Dalenius defined the optimal data privacy in [22] as below:

> access to the published data should not enable the adversary to learn anything extra about any target victim compared to no access to the database, even with the presence of any adversary's background knowledge obtained from other sources.

Unfortunately, the above data privacy definition can not be achieved in real life due to boundless background knowledge that adversaries can obtain from all kinds of sources [27]. Background knowledge is the additional information that is obtained by adversaries from other resources other than the published data set. Generally speaking, the concept of data privacy is to provide data for analysis purpose while protecting individual's sensitive information from people with malicious purposes — defined as "adversaries" in this thesis. This is because one data user could dig useful

statistical relationships and uncover sensitive information of individuals in the data set at the same time. For example, if a collection of medical records is published to data researchers by a health care institution, one researcher could try to learn useful statistical patterns (e.g., smoking people tend to have lung cancer), and use such data to find individual's sensitive information (e.g., Bob's disease is HIV). The person located by adversaries is called as "target" or "victim" in the thesis. In previous example, Bob is the target and his disease information is learned by adversaries, therefore, his privacy is violated.

Numerous methods have been proposed to protect data privacy, see [35, 4, 10] for surveys. These approaches can be categorized into two types [19]: syntactic privacy methods and differential privacy method [27]. In syntactic methods, data is modified in ways such that some syntax conditions are satisfied. While different syntax reflects different ways of defining privacy, they either try to protect individuals from being recognized, or prevent their sensitive information from being learned by adversaries. Some popular syntactic methods will be introduced in Section 2.2.1. Differential privacy shifts the focus from limiting the occurrence of a disclosure to hiding the impact of a single individual on the occurrence. Differential privacy requires that no individual record could significantly change the result of statistical analysis. In its simplest form, it returns noisy query answers by adding some random noise to true answers, where the injected noise is carefully calibrated to achieve differential privacy while providing utility for analysis use. For example, while the true answer to the query "SELECT COUNT(*) FROM data set WHERE Gender = Female" is 200, the returned answer could be 221, where a noise of 21 is randomly drawn following some distribution. After differential privacy was proposed, it quickly became the gold standard privacy definition. A well-known claim is that differential privacy provides strong privacy guarantees against an adversary with strong background knowledge. In particular, it is claimed that even if the adversary knows all but one record in the data set, the privacy is not violated for the individual behind that record: the adversary can not observe a distinguishable difference with or without that specific one record on query results [17].

1.2 Non-independent Reasoning and Data Privacy

The most important difference between syntactic privacy methods and differential privacy method is whether they treat non-independent reasoning (NIR) as a privacy violation. NIR means that the information of one record could be learned from other records in the same data set, with the assumption that they share an identical underlying distribution. Suppose that in a data set of patient information, if the adversary knows that 80% of patients get HIV and Bob is in the data set, without

any other auxiliary information, the adversary has 80% certainty to claim that Bob gets HIV. Here the adversary assumes that all individuals in the data set have equivalent chance of having HIV. NIR is powerful, which is easy to apply and widely adopted in Data Mining and Machine Learning. For example, to accurately learn interesting patterns, usually, the pattern is learned from the training data and then evaluated on the validation data. The rationale behind this operation is that both data comes from the same source and shares the same underlying patterns/distributions.

Most syntactic methods treat NIR as a privacy violation, such as k-anonymity [75], l-diversity [60], t-closeness [57], β-likeness [14], Δ-growth [76] and ρ_1-ρ_2 privacy [32]. To limit the influence of NIR, these syntactic approaches focus on *smoothing* distributions of sensitive attribute values in a sub population. The operation of smoothing makes sure that the distribution of sensitive attribute values in a sub population is not far from the overall distribution in the raw data, and the overall distribution is treated as public information. Without smoothing, adversaries may find individuals in this sub population have a *different* probability of having some value sa on the sensitive attribute compared to the overall individuals, this different probability obtained by adversaries is considered to be a privacy violation in most literature [57, 60, 75]. For example, without smoothing adversaries may find that male engineers tend to have high salary (e.g., $> 50K$), this would easily make Bob, a male engineer, become the target of some financial fraud crimes. While the rule of "male engineers tend to have high salary" seems expected, it does demonstrate the potential risk of NIR on a real life data set. After all, truly sensitive data and findings are difficult to obtain and publish. One drawback of this smoothing strategy is that it limits the desired utility on learning interesting statistical patterns. Because what you can learn from the whole data set is almost what you can learn from any sub population.

Differential privacy, on the other hand, does not consider NIR to be a privacy violation because, as claimed by Blum et al. [12] (page 4), "We explicitly consider non-independent reasoning as a non-violation of privacy; information that can be learned about a row from sources other than the row itself is not information that the row could hope to keep private". Unlike most privacy methods, it shifts the focus from limiting the occurrence of a disclosure to hiding the impact of a single individual on the occurrence. Differential privacy requires that no individual record could significantly change the result of statistical analysis. This kind of constraint, however, is different from limiting the ability of the adversary to infer sensitive attribute values about individuals in the data set. Example 1 below shows how the disease information of one individual is learned by adversaries when differential privacy is applied.

Example 1. *Suppose that Bob is a male engineer and his record is contained in a table $\mathcal{D}$ (Gender, Job, Diseases), where Gender and Job are publicly known, and Diseases is a sensitive (private)*

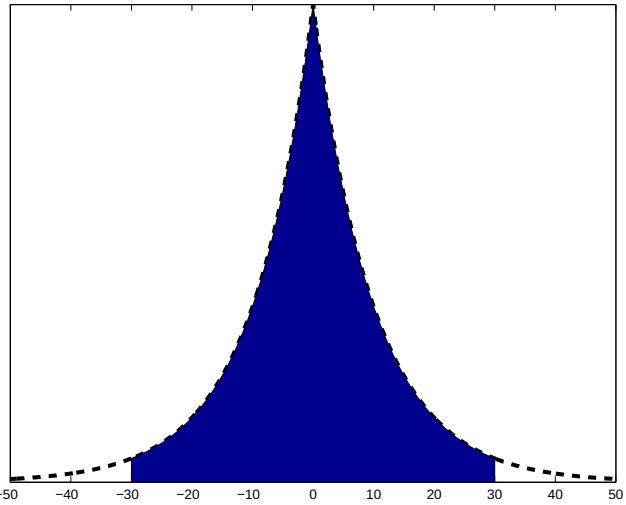


Figure 1.1: Laplace Probability Density ($b = 10$, $\mu = 0$)

attribute. To learn Bob's disease, an adversary issues two count queries, that are queries asking for the number of records satisfying the issued query.

Q_1*: "Gender=M AND Job=engineer",*

Q_2*: "Gender=M AND Job=engineer AND Diseases=HIV".*

To achieve differential privacy, a random Laplace noise ξ_i *(*$i = 1, 2$*) is injected to the true answer of query* Q_i *before returning to users. Suppose that noises are generated by the Laplace distribution with the scale* $b = 10$ *and the mean* $\mu = 0$*, and the noisy answers for* Q_1 *and* Q_2 *are* 1000 *and* 940*, respectively. The density function of a Laplace noise when* $b = 10$ *and* $\mu = 0$ *is illustrated in Figure 1.1, in which the shaded area shows the probability when the noise falls into the range of* $[-30, 30]$*. The cumulative distribution function of the Laplace distribution is:*

$$F(x) = \begin{cases} \frac{1}{2}\exp\left(\frac{x-\mu}{b}\right) & \text{if } x < \mu \\ 1 - \frac{1}{2}\exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases} \tag{1.1}$$

Based on Equation (1.1) the shaded area is $F(30) - F(-30) \approx 0.9$*. This means the probability that the injected Laplace noise* ξ_i *(*$i = 1, 2$*) falls into the range of* $[-30, 30]$ *is* 90%*. If both* ξ_1 *and* ξ_2 *fall into the range of* $[-30, 30]$ *(i.e., the true answer of* Q_1 *is in the rage of* $[970, 1030]$ *and the true answer of* Q_2 *is in the rage of* $[910, 970]$*), then at least* 90% *engineers have* HIV *as disease because* $910/1030 \approx 90\%$*. Since* ξ_1 *and* ξ_2 *are generated independently, both of them fall into the shaded area is around* $90\% \times 90\% = 81\%$*. Therefore, the adversary has the confidence of* more than 81% *to claim that more than* 90% *engineers have HIV. This is because there are other scenarios that the adversary could claim that more than* 90% *engineers have HIV. For example,*

when both ξ_1 and ξ_2 are 50, the true answers of Q_1 and Q_2 are 950 and 890, respectively. This means $890/950 \approx 93.7\%$ engineers have HIV. And the scenario of $\xi_1 = 50$ and $\xi_2 = 50$ does not belong to the 81% probability when both ξ_1 and ξ_2 fall into the range of $[-30, 30]$. Based on NIR, the disease of Bob could be learned from other male engineers in the same data set. Therefore, Bob's disease is revealed. Choosing the larger noise scale $b = 1000$ helps hide Bob's diseases in this case, but is not sufficient if the noisy answers are $1,000,000$ and $999,000$. On the other hand, a large scale renders query answers too noisy for meaningful data analysis. □

Motivated by above observations of syntactic privacy methods and differential privacy method, we raise the following two questions:

> (A). To which extent the sensitive information could be learned when some notion of differential privacy is applied?
>
> (B). Is it possible to (1) allow learning statistical relationships (e.g., smoking people tend to have lung cancer), and at the same time, (2) prevent disclosures on sensitive attribute values of individuals in the data set (e.g., Bob is likely to have cancer)? Syntactic privacy methods satisfy (2) but not (1), and differential privacy method satisfies (1) but not (2).

For (A) we analyze the negative results of differential privacy, i.e., sensitive disclosures when differential privacy is guaranteed. Our notion of disclosures refers to learning sensitive information (such as diseases) of an individual, independent of the participation in the data set, and is based on the utility claim of differential privacy. In other words, we show that the probability of disclosing sensitive information of a record increases whenever the differential privacy mechanism delivers a good accuracy of query answers. It is hard to eliminate such disclosures because they co-occur with a good utility for data analysis that the differential privacy mechanism aims to provide. For (B), the difficulty of achieving both (1) and (2) is that both learning statistical relationships and learning sensitive attribute values of individuals employ NIR. The key of achieving both (1) and (2) is to distinguish these two types of learning. We propose the notion of *reconstruction privacy*, which satisfies both (1) and (2), and we further discuss how to achieve reconstruction privacy.

1.3 Contributions

In this thesis, we have two main contributions by answering the two questions in Section 1.2. Neither of the two leading categories in data privacy, syntactic privacy methods and differential privacy method, is perfect. While syntactic privacy methods do not allow statistical relationships learning, the differential privacy method may release sensitive information of individuals in the data set.

The sensitive information disclosure in differential privacy has been investigated by some previous works [47, 20, 62]. All disclosures discussed in previous works have restricted requirements and details will be introduced in later part of this section. Question (A) is answered by showing that a disclosure may occur without restricted requirements in this thesis. More importantly, it is hard to eliminate such disclosures while allowing statistical relationships learning, because disclosures occurrences and allowing statistical relationships depend on the same thing — the accuracy of differentially private query answers. Intuitively, we ask the question (B): whether there is an approach which is able to overcome shortcomings of both syntactic privacy methods and differential privacy method.

1. Question (A) is answered in Chapter 3 through evaluating the sensitive information releasing when some notion of differential privacy is applied. In particular,

    - A notion of disclosures is formalized in terms of the probability of a small error in learning sensitive information through NIR (Section 3.2). The sensitive information in question (A) is calibrated through such disclosures.
    - A formal analysis is presented on disclosures through query answers that satisfy typical settings of differential privacy (Section 3.3). Specifically, we model the probability of the error of learning sensitive information through NIR by a ratio distribution of two Laplace variables. These variables represent the noisy answers of differential privacy. To our knowledge, this is the first study on the probability of ratio distribution for two Laplace variables. This modelling yields an efficient way of determining the disclosures of query answers produced by the differential privacy mechanism.
    - The above type of disclosures is studied on several real life data sets while a notion of differential privacy is satisfied, and the impact of eliminating such disclosures on data utility (Section 3.4). The study suggests that eliminating disclosures and retaining utility are a direct conflict because both disclosures and utility depend on the same type of information, i.e., noisy query answers. An implication of this study is that, under the assumption that NIR is a privacy violation, differential privacy does not provide both privacy and utility.

2. Question (B) is answered in Chapter 4. Reconstruction privacy is proposed to satisfy both (1) and (2) as in question (B). In particular,

    - The raw data has to be anonymized before publishing for protecting individual's sensitive information. To learn statistical relationships from the anonymized data set, some

estimation approaches have to be adopted. The procedure of estimation is called *reconstruction*. We define two types of reconstruction (Section 4.1): *personal reconstruction* which aims at the sensitive information of a particular individual, and *aggregate reconstruction* that is the source of learning statistical relationships. Besides, we propose an *inaccuracy requirement* on *personal reconstruction* for individuals as a new privacy criterion called *reconstruction privacy*. The division of the two types of reconstruction makes it possible to protect individual's sensitive information while providing useful statistical relationships through limiting the accuracy of personal reconstruction and preserving (as much as possible) the accuracy of aggregate reconstruction, and this further satisfies the two requirements in question (B).

- Reconstruction privacy imposes a minimum value for the *best* upper bound on the probability of having a larger error using F' to estimate f of an adversary, where f and F' are the actual and estimated frequency of a sensitive value in a personal reconstruction (Section 4.1). Note that the thing we try to limit is the error of the reconstruction for f, which should *not* be confused with the relative increase of the attacker's belief in previous works such as the β-likeness [14], t-closeness [57] and (ρ_1, ρ_2)-privacy [32]. Unlike these previous works, reconstruction privacy does not bound the maximum value of F' or f or require them to be close to the global distribution, making it suitable for learning statistical relationships through aggregate reconstruction. Also, reconstruction privacy avoids modeling the prior of an adversary, which can be tricky as shown in [27][12] but is necessary in these previous works.
- An efficient test of reconstruction privacy is presented (Section 4.2). First, we show a conversion between an upper bound for the tail probability of Poisson trials into an upper bound on the probability of having a larger error using F' to estimate f. Then, we obtain an efficient test of reconstruction privacy by adapting the notion of reconstruction privacy to an existing upper bound for Poisson trials, i.e., the Chernoff bound.
- An efficient algorithm for producing a perturbed version data set that satisfies a given specification of reconstruction privacy is presented (Section 4.3). The algorithm is highly efficient because it only needs to sort the records once and make another scan on the sorted data.
- Two claims are evaluated (Section 4.4). The first claim is that reconstruction privacy can be violated by real life data sets even after data perturbation. The second claim

is that the proposed method can preserve utility for statistical learning while providing reconstruction privacy.

1.4 The Differences from Previous Works

Sensitive disclosures under differentially private query answers have been examined by recent works [47, 20, 62, 85]. McClure et al. [62] proposed a way to generate binary synthetic data that satisfies differential privacy. It also calculated the posterior probability of adversaries on uncovering true sensitive values as the risk of statistical disclosure. The result in [62] showed that the level of differential privacy does not directly affect the extent of statistical disclosure risks. Unfortunately, the disclosure discussed in [62] depends on the simulation of data generation procedure. In this paper, our way of defining disclosures does not depend on data generalization, therefore, is more general and practical.

Kifer et al. [47] argued that in the presence of correlation on the sensitive information of records (e.g., if one member in a family gets flu, the other members in the family are likely to get flu as well), sensitive information of an individual could be learned from differentially private query answers. Intuitively, the absence of a record is no longer sufficient for hiding the sensitive information of a record because such information could be learned from correlated records. Bayesian differential privacy was proposed by Yang et al. [85] to evaluate and prevent disclosures through correlated data modelled by a Gaussian correlation model. The disclosures in this paper do not depend on record correlations. Our disclosures depend on only the utility of published answers: when a differential privacy mechanism promises a good utility, i.e., good accuracy of query answers, the probability of learning sensitive information of a record from such answers increases. Such disclosures are hard to prevent because they co-occur with good utility that a mechanism aims to provide.

Cormode et al. [20] demonstrated that a Bayes classifier could be built using differentially private query answers to predict the value of an individual on sensitive attributes. The Bayes classifier requires multiple queries for computing the joint probability of all attributes. This paper is different in the following aspects. First, we show that each disclosure requires only two queries, which increases the occurrence of disclosures by limiting the impact of noises to two queries. Second, our disclosures take into account the changes of confidence of learning sensitive information for a target individual relative to the confidence in the entire data set, which impose a larger threat. Third, we provide a theoretical explanation for the disclosures under differential privacy based on the convergence of the ratio distribution of two random Laplace variables. This result suggests that it is possible to learn the sensitive information of an individual from the answers published by a

differential privacy mechanism with an arbitrary accuracy provided that the true query answers are sufficiently large. In addition to the theoretical results, we also demonstrated that disclosures indeed occur on real life data sets, especially when a differential privacy mechanism delivers good utility.

Chapter 2

Preliminary

The raw data can not be published due to privacy concern, and the data has to be modified somehow to satisfy some notion of data privacy. Usually, the modification is carefully done through some anonymization operations. In this chapter we first introduce several leading categories of anonymization operations for achieving some notion of data privacy. Secondly, we introduce two main categories of data privacy definitions: syntactic methods and differential privacy. We also survey some classic approaches in syntactic methods.

At beginning people tried to protect privacy by removing *explicit identifiers* before releasing the data. Explicit identifiers are attributes that can be used alone to locate a person, such as names and SIN numbers. Unfortunately, this simple strategy failed to protect privacy [8, 66]. After explicit identifiers are removed, even single attribute could not uniquely identify a person, the combination of some attributes often singles out a person, such as $\{Age, Gender, Zipcode\}$. The combination of these attributes is called *Quasi-Identifier* (QI) [23]. In addition to QI, the data set $\mathcal{D}$ also contains one sensitive attribute SA (usually, the disease information in the medical record or the salary information in a census data set), which should be protected from adversaries. QI values are known to public, in other words, they are non-sensitive to adversaries.

In this thesis we assume all attributes other than SA are QI and we also define them as non-sensitive attributes (NA). In this thesis QI and NA are used interchangeably for referring attributes other than SA. In this thesis we define QI $group$ as a collection of records that agree on some QI value. And all records in the same QI $group$ share the same values on QI. It should be noted that adversaries can not distinguish the target from other individuals in the QI group that the target belongs to, because all individuals in one QI group share the same QI values and adversaries does not know the SA value of the target (otherwise there is not way to protect the target's SA values).

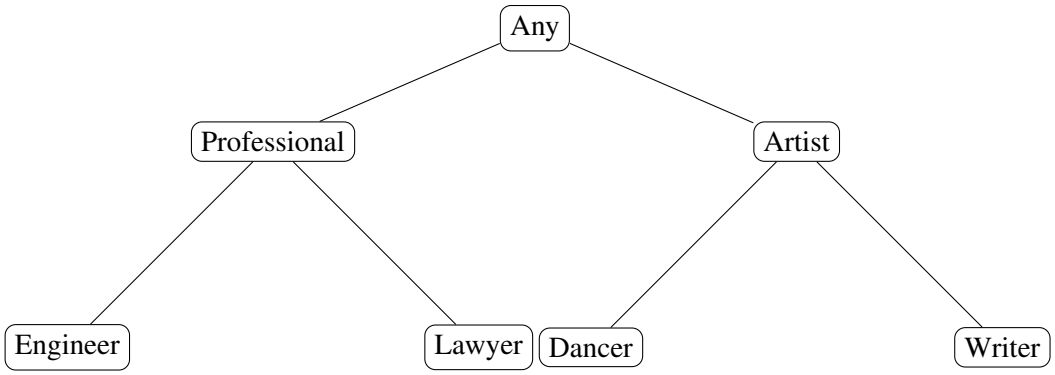
2.1 Anonymization Operations

The purpose of anonymization is to change data, before it is published, so that the owner of each record is hard to be identified, or the sensitive information of each record is hidden from adversaries. Some famous anonymization operations are: generalization, suppression, anatomization, permutation and perturbation [36]. Generalization and suppression try to make data less accurate for adversaries to learn sensitive information. For example, the job *production designer* may be replaced by *Artist*. Anatomization and permutation try to de-connect the relationship between QI and SA, so that adversaries can not infer accurate SA information of the target even she/he knows the QI information of the target. The good thing about anatomization and permutation is that neither QI values or SA values are modified, however, since the relationship between the two are de-associated it is not clear how regular Data Mining and Machine Learning tasks, such as clustering, classification, could be achieved. Perturbation tries to maintain the original data for some probability while changing the value of data for other cases. For example, the disease *HIV* may be replaced by Flu. In this chapter we introduce two most popular methods among the above five methods with details: generalization and perturbation.

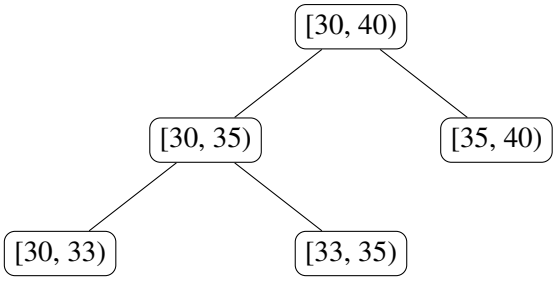
2.1.1 Generalization

The generalization operation replaces the original value with a less specific value. In particular, the replaced value has to be a super set of the original value. For categorical values, the value will be replaced by the taxonomy parent of itself. For numeric values, usually a specific value will be replaced by the interval which covers the original value, and the interval can be treated as a special type of categorical values. For example, in Figure 2.1 the taxonomy parent of *Engineer* is *Professional* and $[30, 35)$ is the taxonomy parent of $[30, 33)$ and $[33, 35)$. The purpose of generalization is to make data less accurate and make it harder for adversaries to learn sensitive information. The reverse operation of generalization is called *specialization*.

While generalization helps preventing adversaries from accurately learning sensitive information of targets, it also leads to an inaccurate statistical relationship learning because of less description values. It has been proved that it is NP-hard to find the best generalization that could give optimal utility while achieving privacy [64]. At beginning, generalization approaches [51, 72, 75] require that all generalized values have to be at the same level. For example, in Figure 2.1 if *Engineer* has to be generalized to *Professional* to satisfy some notion of data privacy, then *Lawyer* has to be generalized to *Artist* as well because *Artist* and *Professional* are in the same level. Later many works [9, 37, 38, 45, 52, 84] focus on improving utility using generalization while achieving the



(a) Jobs



(b) Age

Figure 2.1: Taxonomy Trees for $Jobs$ and Age

same level of privacy. And they do not require that generalized values have to be in the same level any more.

2.1.2 Perturbation

Perturbation has been considered to be a promising approach to achieve data privacy in the history of statistical disclosure because of its simplicity, efficiency and the ability to maintain statistical information [36, 4, 77, 80, 25]. In perturbation, some data values are replaced by some synthetic data with some probability so that adversaries can not easily link some record to its data owner.

A popular perturbation is called *Uniform Perturbation* [5]. For a given retention probability p, where $0 < p < 1$, for each record in $\mathcal{D}$, we toss a coin with head probability p. If the coin lands on head, the SA value in the record is retained; if the coin lands on tail, the SA value in the record is replaced with a value picked from the domain of SA with equal probability (i.e., $\frac{1-p}{m}$) at random. This perturbation operator is characterized by the following matrix $\mathbb{P}_{m\times m}$:

$$\mathbb{P}_{ji} = \begin{cases} p + \frac{1-p}{m} & \text{if } j{=}i \text{ (retain } sa_i\text{)} \\ \frac{1-p}{m} & \text{if } j{\neq}i \text{ (perturb } sa_i \text{ to } sa_j\text{)} \end{cases} \tag{2.1}$$

A proper choice of the retention probability p can ensure some privacy requirements, such as ρ_1-ρ_2 privacy [32][6]. The original proposal of ρ_1-ρ_2 privacy [32] considers the scenario where there is only the sensitive attribute SA, no other attributes. In this scenario, ρ_1-ρ_2 privacy holds if whenever the prior $\Pr[SA = sa]$ (i.e., the frequency of sa in the overall data set $\mathcal{D}$) is not more than ρ_1, the posterior $\Pr[SA = sa \mid SA^* = sa^*]$ (i.e., after observing the perturbed value sa^* the confidence of adversaries to claim that the original value is sa) is not more than ρ_2, where SA^* denotes the perturbed SA and sa^* denotes the perturbed value in a record in $\mathcal{D}^*$ [1]. $\Pr[SA = sa]$ is publicly known. Intuitively, ρ_1-ρ_2 privacy bounds the posterior of inferring the original SA value after observing the perturbed SA value in a record.

The most different thing of perturbation compared to generalization is that, the value of each record observed after perturbation can not reflect the authentic value, while the value observed after generalization, though not accurate, reflects the authentic value. In the view of statistical learning, the utility of the data set may be badly damaged due to large extent of generalization, while the statistical information may be better reserved after perturbation.

2.2 Data Privacy Definitions

2.2.1 Syntactic Methods

Usually syntactic methods protect data privacy through generalization and perturbation operations. In this section we introduce some popular and representative syntactic methods.

***k*-anonymity [75]**

k-anonymity is a well known syntactic privacy solution to prevent locating individuals by using their QI values. k-anonymity requires that in each QI $group$ there has to be at least k records, in other words, at least k records share the same QI values. Table 2.1 shows an example of 3-anonymity on $\{Job, Gender, Age\}$.

With k-anonymity the adversary may not precisely identify the record of the target, but she/he could infer the target's SA value from the published data because k-anonymity does not take care of SA values. For example, from Table 2.1 it can be simply observed that if the target is an artist then the probability that the target has HIV is 75% (three out of four artists have HIV).

***l*-diversity [60]**

[1] This corresponds to the upward ρ_1-ρ_2 privacy in [32], where the authors also defined the notion of downward ρ_1-ρ_2 privacy.

Table 2.1: Example of 3-anonymity

Job	Gender	Age	Disease
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	Hepatitis
Professional	Male	[35-40)	HIV
Artist	Female	[30-35)	Flu
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV
Artist	Female	[30-35)	HIV

l-diversity is proposed to address the limitation of k-anonymity by requiring that every QI group should contain at least l "well-represented" SA values. The simplest understanding of "well-represented" is to ensure there are at least l distinct SA values in each QI group.

When the overall distribution of SA values is skewed, however, the sensitive value of individuals could still be revealed, as called *skewness attack* in [57]. For example, in a patient data set where 99% people have Flu and the rest 1% people have HIV. Suppose that in a QI group g, distributions for these two values are both 50%, which satisfies 2-diversity. But it already reveals that individuals in g have more chance than the general population to have HIV.

t-closeness [57]

To avert skewness attack, t-closeness requires that the distribution of SA values in any QI group, P, must be within the maximum distance t from the distribution of SA values in the whole data set, Q. Here, Q is treated as the prior belief of adversaries and is public information. P, on the other hand, is considered to be the posterior belief of adversaries. The distance between P and Q, is treated as the information gain of adversaries in [57].

The distance between P and Q in t-closeness is the *cumulative* difference of these two distributions. This way of computing distance, though, does not protect those *less frequent* SA values, that are more vulnerable and deserve more protection in most cases (e.g., it is considered a more serious threat of releasing the disease is HIV, compared to revealing the disease is Flu, and HIV is more rare compared to Flu in general). Because the change of less frequent SA values are hidden by the change of more frequent SA values.

β-likeness [14]

β-likeness requires that the relative difference between the distribution of any sensitive value sa, r, in any QI group, and the overall distribution of sa in the whole data set, q, can not be greater than β, e.g., $\frac{r-q}{q} \leq \beta$. Unlike t-closeness, β-likeness treat all SA values equally and try to limit the change of each SA value within the threshold β.

ρ_1-ρ_2 privacy [32]

ρ_1-ρ_2 privacy requires that when the *prior* probability of a SA value sa (stereotypically, the overall distribution of sa), is no more than ρ_1, then its *posterior* probability is no more than ρ_2, and $\rho_1 < \rho_2$, where the posterior probability means the certainty of inferring the original SA is sa after observing the anonymized SA value.

Δ-growth [76]

Δ-growth can be treated as an extension version of ρ_1-ρ_2 privacy. It requires that the posterior for a SA value sa is within the maximum difference of Δ from the prior probability. By setting Δ to $\rho_2 - \rho_1$ no Δ-growth breach immediately guarantees no ρ_1-ρ_2 breach.

Remarks. k-anonymity and l-diversity fail to protect sensitive attribute values of individuals when the adversary has background knowledge. Background information can be obtained from various sources, such as well known facts, demographic information and observations of specific individuals. Suppose in a QI group consisting of two records with SA values of HIV and Flu. If the adversary knows that these two records belong to Bob and Alice, and Alice gets Flu, then the disease of Bob is revealed. ρ_1-ρ_2 privacy and Δ-growth may not protect SA values of targets in some scenarios. In the presence of the public attributes NA, the information on NA of a target individual t can be used to bias the selection of records for t for estimating the posterior. As a result, the posterior for t could be higher than (or lower than) $\Pr[SA = sa \mid SA^* = sa^*]$ (i.e., after observing the perturbed value sa^* the confidence of adversaries to claim that the original value is sa) derived in the absence of NA.

Other than the privacy issue, most syntactic methods also limit the desired utility for data researchers. For example, all t-closeness, β-likeness and ρ_1-ρ_2 privacy aim to limit the distance of posterior belief of adversaries (i.e., Q in t-closeness, r in β-likeness and ρ_2 in ρ_1-ρ_2 privacy) from the prior belief (i.e., P in t-closeness, q in β-likeness and ρ_1 in ρ_1-ρ_2 privacy). In this case it makes hard for data researchers to find useful pattern such as smoking people tend to have lung cancer, because what you can find in the overall data set is almost what you can find in any sub population.

The approach in this thesis, which will be introduced in detail in Chapter 4, does not bound the posterior belief of an adversary. Instead, it tries to bound the accuracy of estimating a SA value sa in a sub population after anonymization. Individual's sensitive information is protected because the SA distribution can not be accurately learned by adversaries. This strategy has two important benefits: it allows the room for learning statistical relationships (data user's posterior belief is not limited), and it frees the publisher of measuring the adversary's prior belief and specifying a threshold for posterior beliefs, which can be tricky [27][12].

2.2.2 Differential Privacy

Two data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ are neighbouring data sets if they have the same cardinality but differ in one record. The following notion of differential privacy is proposed in [27].

Definition 1. (ϵ-Differential Privacy) [27]. *A randomization mechanism $\mathcal{A}$ satisfies ϵ-differential privacy if for any output O of $\mathcal{A}$ and any neighbouring data sets $\mathcal{D}_1$ and $\mathcal{D}_2$,*

$$\Pr[\mathcal{A}(\mathcal{D}_1) = O] \leq exp(\epsilon) \times \Pr[\mathcal{A}(\mathcal{D}_2) = O]. \tag{2.2}$$

Typically ϵ is a small value close to zero and $exp(\epsilon)$ is a value close to 1. The inequality in Equation (2.2) implies that any neighbouring data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ will have the nearly equal probability for producing the output O. In other words, no single record could significantly affect the probability of the randomized output; the privacy of this record is protected because it can have any value as far as the published information is concerned. The parameter ϵ controls the privacy level. A smaller ϵ means a stronger privacy level because the two probabilities in Equation (2.2) are closer. In the literature [56, 43, 20, 24, 55, 81], ϵ is typically chosen from the range of $[0.01, 2]$. In this thesis, we follow this range.

For answering statistical queries, ϵ-differential privacy is achieved by a randomization mechanism that adds appropriately scaled random noise to the output of each query. The scale of the noise depends on the *sensitivity* of the class of queries, which captures the maximum possible change caused by a single record on the output of queries. Let Δ denote sensitivity, that is defined as below.

Definition 2. (Sensitivity) [27]. *The sensitivity of a sequence of queries, Q, denoted as Δ, is defined as:*

$$\Delta = \max_{\mathcal{D}_1, \mathcal{D}_2} ||Q(\mathcal{D}_1) - Q(\mathcal{D}_2)||_1, \tag{2.3}$$

for neighbouring data sets $\mathcal{D}_1$ and $\mathcal{D}_2$, where $Q(\mathcal{D}_i)$ is the vector of query answers on data set $\mathcal{D}_i$, $i = 1, 2$, and $||.||_1$ is the 1-norm of a vector.

In this thesis we focus on the core case when all queries are *count queries*, that are queries asking for the number of records satisfying the issued query. The function $Q(\mathcal{D}_i)$ maps the data set $\mathcal{D}_i$ to a vector of real number answers. For example, if the data set has one QI attributes: *Age*, and Q contains two queries: "Q_1*: SELECT COUNT(*) FROM data set WHERE Age*=20" and "Q_2*: SELECT COUNT(*) FROM data set WHERE Age* = 30". The sensitivity of Q is 1, because inserting or deleting a record (e.g., a record of a 30-year-old individual) would change only one query (i.e., Q_2) result by at most 1 and never change the other one query (i.e., Q_1) result. Note that if editing

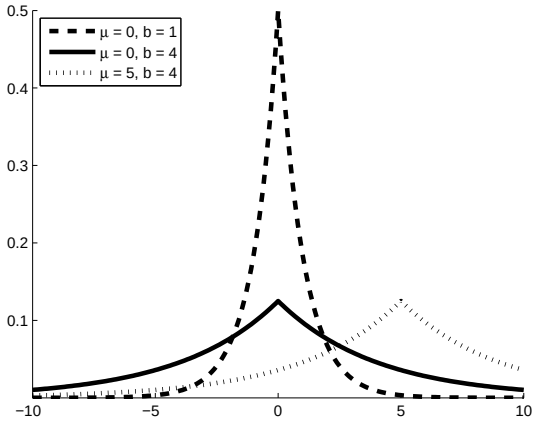


Figure 2.2: Laplace Probability Density Function

record is allowed then the sensitivity of Q is 2, because editing a record (e.g., changing a record from 30-year-old to 20-year-old) would change each query result by at most 1.

For answering statistical queries, ϵ-differential privacy is achieved by a randomization mechanism that adds appropriately scaled random noise to the output of each query. The injected noise may be drawn from a Laplace distribution [27], a Gaussian distribution [29, 63], or from some carefully designed matrix [55] to achieve differential privacy. In this thesis we focus on the Laplace mechanism, which has been shown in [27] to achieve differential privacy, because they are most studied in the literature.

We denote $Lap(b)$ as the Laplace random variable with mean $\mu = 0$ and scale b. Figure 2.2 shows the probability density function for Laplace distribution under various b and μ. The density function for Laplace distribution is

$$f(x) = \frac{1}{2b} exp\left(-\frac{|x-\mu|}{b}\right)$$

From Figure 2.2 we can observe that the generated random variable is concentratively distributed around the mean, 0. In other words, the probability of generating small noise is high while producing large noise is unlikely. This property supports Laplace mechanism as a good scheme to serve data privacy because the noise should not be neither too large to dominate the noisy answer nor too small to hide the true answer. The other observation we can obtain is that the density curve becomes flatter as b increases, which implies that large noise would be more likely to be added to the true counts, thus, the true counts are better hidden from the adversary. This property supports us to control the level of added noise by setting the value for b.

Theorem 1. (Laplace Mechanism [27]). *Let the set of queries $Q = \{Q_1, \cdots, Q_n\}$ be the sequence of statistical queries, where n is the total number of queries contained in Q. If an algorithm $\mathcal{A}$ adds i.i.d. Laplace noise with scale $b = \Delta/\epsilon$ and mean $\mu = 0$ to the result of each query in Q, where Δ is the sensitivity of Q, then $\mathcal{A}$ satisfies ϵ-differential privacy.*

Proof. Let $\mathcal{A}_Q$ be the Laplace mechanism for the set of queries Q. To satisfy ϵ-differential privacy, the Laplace mechanism, $\mathcal{A}_Q$, adds a Laplace variable with mean $\mu = 0$ and scale b to the result of each query in Q. For any output O of $\mathcal{A}$, the Laplace mechanism has been shown to satisfy ϵ-differential privacy [27].

$$\Pr[\mathcal{A}_Q(\mathcal{D}_1) = O] = \frac{1}{2b} exp\left(-\frac{||Q(\mathcal{D}_1) - O||_1}{b}\right) \tag{2.4}$$

$$\Pr[\mathcal{A}_Q(\mathcal{D}_2) = O] = \frac{1}{2b} exp\left(-\frac{||Q(\mathcal{D}_2) - O||_1}{b}\right) \tag{2.5}$$

Recall that to satisfy ϵ-differential privacy, we have to make sure the value of $\frac{\Pr[\mathcal{A}_Q(\mathcal{D}_1)=O]}{\Pr[\mathcal{A}_Q(\mathcal{D}_2)=O]}$ has an upper bound as $\exp(\epsilon)$. Replacing $\Pr[\mathcal{A}_Q(\mathcal{D}_1) = O]$ by the right term of Equation (2.4) and replacing $\Pr[\mathcal{A}_Q(\mathcal{D}_2) = O]$ by the right term of Equation (2.5)

$$\frac{\Pr[\mathcal{A}_Q(\mathcal{D}_1) = O]}{\Pr[\mathcal{A}_Q(\mathcal{D}_2) = O]} = \exp\left(\frac{||Q(\mathcal{D}_2) - O||_1}{b} - \frac{||Q(\mathcal{D}_1) - O||_1}{b}\right)$$

Applying the triangle inequality within the exponent

$$\frac{\Pr[\mathcal{A}_Q(\mathcal{D}_1) = O]}{\Pr[\mathcal{A}_Q(\mathcal{D}_2) = O]} \leq \exp\left(\frac{||Q(\mathcal{D}_2) - Q(\mathcal{D}_1)||_1}{b}\right) \tag{2.6}$$

Note the right term of Equation (2.6) is bounded by $\exp\left(\frac{\Delta}{b}\right) = \exp(\epsilon)$, which is required by the definition of the sensitivity of Q, and $b = \Delta/\epsilon$ as defined in Theorem 1. Thus the ϵ-differential privacy is achieved.

□

Many techniques have been proposed for applying differential privacy to specific data publishing and mining tasks. A survey by Dwork [28] provides a comprehensive review. For example, differential privacy has been applied to releasing query and click histograms from search logs [40, 48], recommender systems [63], publishing commuting patterns [59], publishing results of machine learning [13, 15, 46], clustering [33, 67], decision trees [34], mining frequent patterns [11], and aggregating distributed time-series [70].

An important assumption made by differential privacy is that NIR is not considered as a privacy violation. This is stated no more clear than in [12]: "We explicitly consider non-independent reasoning as a non-violation of privacy; information that can be learned about a row from sources other than the row itself is not information that the row could hope to keep private". In reality, however, NIR remains a real threat to privacy: if Bob's HIV status is *accurately* inferred from differentially private query answers and if Bob cares about keeping this status private, Bob's privacy is breached, even if Bob's record is not in the data. In fact, if Bob's HIV status can be accurately learned from other records, it makes no difference whether Bob's record participates in the learning process.

This paper treats NIR as a privacy violation and considers a different type of privacy, that is, whether some private information of an individual can be learned from the published information given that a target individual is known to participate in the database. While differential privacy ensures that the *presence (or absence)* of an individual can not be learned , we consider a notion of privacy ensuring that *sensitive information* of a target individual can not be learned regardless the presence or absence of an individual. Our work shows that the former does not guarantee the latter.

Some previous works on differential privacy focus on applying differential privacy to various data publishing scenarios. For example, [13, 42, 55, 81, 78, 3, 54] consider publishing query answers, [43, 83] consider publishing histograms, and [7, 24] consider publishing marginals. These works assume that the privacy concern is addressed by a specified ϵ-differential privacy, which ensures that an adversary can not distinguish whether an individual participates in the statistical database from published information. This thesis treats NIR as a privacy violation and considers a different type of privacy, that is, whether some private information of an individual can be learned from the published information given that a target individual is known to participate in the database. While differential privacy ensures that the *presence (or absence)* of an individual can not be learned, we consider a notion of privacy that ensures that *sensitive information* of a target individual can not be learned. Our work shows that the former does not guarantee the latter.

Some previous works focus on improving the utility of differential privacy. They can be divided into three categories: (1) data-aware algorithms [3, 21, 41, 82, 83, 55] that add different scales of noises to different data sets; (2) workload-aware algorithms [24] that add different scales of noises to different query workloads; (3) data- and workload-aware algorithms [54] that add different scales of noises to different data sets and different query workloads. Most of these works [3, 21, 83, 55, 54] focus on low-dimension data (i.e., histograms or marginals). They partition the data set to multiple bins of non-overlapped data and add carefully calibrated noise to each bin. The partition is also delicately designed for achieving better utility. These works improve the utility significantly. However, it is hard to extend these algorithms to adapt to high-dimension data because

the procedure of finding the best partition usually consumes enormous time (i.e., $O(n^3u^3)$ in [55] and $O(nu \log u + u^2)$ in [54], where n is the total number of queries in the query load and u is the product of all possible values in non-sensitive attributes and sensitive attributes).

Most of these algorithms add non-universal noise to query answers. In particular, they try to wisely add less noise to achieve better utility. In this paper, we aim to evaluate sensitive disclosures under differential privacy, not the utility. Therefore, it suffices to only consider the classic case when differential privacy is achieved through adding the fixed scale of noise. If a disclosure can be found when a fixed scale of noise is added, then it can almost always be found when less noise is added.

Chapter 3

Evaluating Disclosures under Differential Privacy

Non-independent reasoning (NIR) means that the information of one record could be learned from other records in the same data set, with the assumption that they share an identical underlying distribution. In Chapter 1 we pointed that differential privacy does not consider NIR to be a privacy violation, therefore, the SA values of one individual may be inferred by adversaries from records of other individuals through NIR even when differential privacy is applied.

In this chapter, we firstly illustrate one examples of disclosures under differential privacy in Section 3.1. This examples intuitively show how sensitive information could be uncovered when differential privacy is applied. Secondly, we formalize a notion of disclosures in terms of the probability of a small error in learning sensitive information through NIR in Section 3.2. Thirdly, we present a formal analysis on disclosures through query answers that satisfy typical settings of differential privacy in Section 3.3. Specifically, we model the probability of the error of learning sensitive information through NIR by a ratio distribution of two Laplace variables. And these variables represent for noisy answers returned by differential privacy mechanism. To our knowledge, this is the first study on the probability of ratio distribution for two Laplace variables. This modelling yields an efficient way of determining the disclosures of query answers produced by the differential privacy mechanism. Fourthly, we study the above type of disclosures on several real life data sets while a notion of differential privacy is satisfied, and the impact of eliminating such disclosures on data utility in Section 3.4. And finally, this chapter is summarized in Section 3.5.

3.1 Examples of Disclosure under Differential Privacy

Recall that Example 1 in Section 1.2 has been shown that NIR could be applied to infer the sensitive information of individuals in data sets with good accuracy when some notion of differential privacy is applied. One may assume that there should be no privacy issue in the above example if HIV is shared by so many records in $\mathcal{D}$. This needs clarification. Typically, $\mathcal{D}$ is collected for a targeted study, as such, a majority in $\mathcal{D}$ is not necessarily a majority in the whole population, hence, possessing a sensitive value is considered sensitive. For example, if $\mathcal{D}$ is collected for an HIV study, it is possible that 60% (i.e., a majority) of the males in $\mathcal{D}$ have HIV, but since the actual probability of HIV in a general population is much lower than 60%, learning someone having HIV with 60% probability is highly sensitive.

The above disclosure occurs to any noise distribution that has a *fixed* scale that is independent of the database size. To our knowledge, most differential privacy mechanisms use a fixed scale (such as Gaussian mechanism [29, 63] and Matrix mechanism [55]). While a fixed scale noise masks the impact of a single individual on the occurrence of a disclosure, it does not mask the fact that *a disclosure occurs to the individual*. This does not match legal definitions of privacy [50], i.e., individually identifiable data needs to be protected. Under differential privacy, even a single individual may not significantly affect the statistical analysis, the sensitive information of that individual may be uncovered by adversaries, which has been shown in Example 1.

The next example (Example 2) shows that disclosures could be prevented by applying stronger level of differential privacy, however, this comes with the cost of sabotaged utility.

Example 2. *Consider the* ADULT *data set [1] that contains 45,222 records (without missing values) from the 1994 Census database. Consider the five attributes Education, Occupation, Race, Gender, and Income. The Income attribute has two values, "$\leq$50K", for* 75.22% *of records, and "$>$50K", for* 24.78% *of records. We assume that learning the Income value for a record is sensitive. On the raw data, the following two count queries* Q_1 *and* Q_2 *return the answers* $ans_1 = 501$ *and* $ans_2 = 420$*, respectively:*

Q_1*: "Prof-school* $\wedge$ *Prof-specialty* $\wedge$ *White* $\wedge$ *Male",*
Q_2*: "Prof-school* $\wedge$ *Prof-specialty* $\wedge$ *White* $\wedge$ *Male* $\wedge$ *$>$50K".*

These answers imply the following rule with the confidence $Conf = \frac{ans_2}{ans_1} = 0.8383$*. The confidence indicates the certainty of the adversary to claim that any person satisfying* Q_1 *has more than* $50K$ *salary.*

$$\{\textit{Prof-school, Prof-specialty, White, Male}\} \rightarrow \textit{>50K.}$$

Since this confidence is significantly higher than the overall frequency 24.78% of the value ">50K", this rule may violate the privacy of the individuals matching the condition of Q_1. While this rule seems expected, it does demonstrate the potential risk of NIR on a real life data distribution. After all, truly sensitive data and findings are difficult to obtain and publish.

The differential privacy mechanism will return the noisy answers $ans'_i = ans_i + \xi_i$, $i = 1, 2$, where the noises ξ_i's follow some distribution, and an adversary has to gauge $Conf$ by $Conf' = \frac{ans'_2}{ans'_1}$. Consider the widely used Laplace noise distribution $Lap(b) = \frac{1}{2b}\, exp(-\frac{|\xi|}{b})$, where b is the scale factor. The setting of $b = \Delta/\epsilon$ would ensure ϵ-differential privacy for the sensitivity Δ of the query function. Let us set $\Delta = 2$ to account for the two count queries. Note that the effect of a larger Δ can be simulated by the effect of a smaller ϵ because $b = \Delta/\epsilon$.

Table 3.1: {Prof-school, Prof-specialty, White, Male} $\rightarrow$ >50K (Conf=83.83%)

	$\epsilon = 0.01$ ($b = 200$)		$\epsilon = 0.1$ ($b = 20$)		$\epsilon = 0.5$ ($b = 4$)	
	Mean	SE	Mean	SE	Mean	SE
$Conf'$	1.34392	1.36299	0.860966	0.0985138	0.832659	0.0645165
$\lvert ans_1 - ans'_1\rvert / ans_1$	0.614742	0.533185	0.0693353	0.0272098	0.0262412	0.0144438
$\lvert ans_2 - ans'_2\rvert / ans_2$	0.570118	0.983959	0.102247	0.0820627	0.069974	0.0636316

Table 3.1 shows the mean of $Conf'$ and the relative error $\frac{|ans_i - ans'_i|}{ans_i}$ of query answers over 10 trials of random noises, and the standard error (SE) of the mean [1] *and $SE = \frac{s}{\sqrt{10}}$, where s is the standard deviation and 10 is the total number of random trials. $Conf'$ measures the disclosure (in red) and $\frac{|ans_i - ans'_i|}{ans_i}$ measures the utility of query answers (in blue). At the higher privacy level $\epsilon = 0.01$, $Conf'$ deviates substantially from $Conf = 0.8383$, but the utility of the noisy answers is also poor because of the large relative errors and SE. At the lower privacy level $\epsilon = 0.5$, the utility of noisy answers improves significantly, but $Conf' = 0.8327$ is within 1% difference from $Conf$ with a small SE (i.e., 0.0645); in this case, any instances of ans'_1 and ans'_2 are sufficient to gauge the income level of an individual.* □

To ensure a good utility, a fixed (and small) scale b of noises is essential. Indeed, improving utility through reducing noises is a major focus of the work on differential privacy (see [53] for a list). As the query answer becomes larger, such noises become less significant, which improves the utility of noisy answers ans'_i, therefore, the accuracy of $Conf' = \frac{ans'_2}{ans'_1}$. Thus, the good utility of ans'_i comes together with the risk of disclosures. A general and quantitative analysis on this type of attack will be presented in Section 3.2. Choosing a large noise scale (i.e., a smaller ϵ) helps thwart such attacks, but it also hurts the utility for data analysis. In fact, as long as the noise scale stays fixed, the noises eventually become insignificant for large query answers.

[1] https://en.wikipedia.org/wiki/Standard_error

3.2 Disclosure of Noisy Answers

We now formalize the notion of disclosures through noisy answers and NIR. In this section, we consider the noisy answers generated by any noise distribution, not necessarily by the differential privacy mechanism. The case for noisy answers published by the differential privacy mechanism will be considered in Section 3.3.

3.2.1 Attacks

We consider a micro data set $\mathcal{D}$ containing one sensitive attribute (SA) and several other public non-sensitive attributes $NA = \{A_1, ..., A_{|NA|}\}$, where $|NA|$ is the number of distinct attributes in NA. For $1 \leq i \leq |NA|$, x_i denotes a domain value of A_i. Let $g = \mathcal{D}(x_1, ..., x_{|NA|})$ be a *personal group*, which is the subset of records that match x_i on every A_i. For example, in the data set with $NA = \{Gender, Age, Occupation\}$, the subset of records satisfying $\{Gender = Female, Age = 22, Occupation = Engineer\}$ is a personal group, and the subset of records agreeing on $\{Age = 22, Occupation = Engineer\}$ is the union of two personal groups $\{Gender = Female, Age = 22, Occupation = Engineer\}$ and $\{Gender = Male, Age = 22, Occupation = Engineer\}$. The total number of personal groups in the data set depends on the number of values in each NA. For example, if $NA = \{Gender, Zipcode\}$ and if $Gender$ has two values and $Zipcode$ has 100 values, there are 200 possible personal groups with each group containing the records in $\mathcal{D}$ that share the same values on $Gender$ and $Zipcode$. Notice that each record belongs to exactly one personal group.

Suppose that an adversary wants to learn the SA value of some target individual t. The adversary knows all values of the public NA of t and that t has a record in $\mathcal{D}$. One way to learn the SA value of t is to identify the personal group to which t belongs (i.e., the group that matches all t's NA values), say g, and to infer the information on SA of t from the distribution of SA values of the records in this group, assuming that t follows the same distribution on SA as those in its personal group through NIR. This assumption makes sense in that the records in the same personal group are indistinguishable by their NA. There may be other ways to learn the SA value of t, for example, using the combined set of the records from multiple personal groups that partially match t's NA values, provided that t's SA value follows the same distribution as those records in these groups. But considering disclosures via personal groups suffices for our purpose, which is showing disclosures, instead of eliminating disclosures.

To find t's value on SA, the adversary can estimate the probability of a particular SA value sa in g. Let g_{sa} denote the set of records in g that have sa on SA. This probability can be estimated by

$\frac{|g_{sa}|}{|g|}$, called the *true confidence* of sa in g, where $|\cdot|$ is the size of a set. Under our assumption, this confidence is the adversary's best bet on t's probability of having sa using the information on g.

Now consider a publishing mechanism where a user extracts information about $\mathcal{D}$ by getting answers to count queries. As a user, the adversary gets $|g|$ and $|g_{sa}|$ through the following two queries:

$$\begin{aligned} &Q_1 : \text{“SELECT COUNT(*) FROM } \mathcal{D} \text{ WHERE } NA = t[NA]\text{”} \\ &Q_2 : \text{“SELECT COUNT(*) FROM } \mathcal{D} \text{ WHERE } NA = t[NA] \text{ AND } SA = sa\text{”}. \end{aligned} \tag{3.1}$$

where $NA = t[NA]$ denotes the condition $A = t[A]$ for each NA attribute A in g and $t[A]$ is the value of t on A. Let ϕ and θ be the answers for queries Q_1 and Q_2, where $\theta \geq 0$, $\phi \geq \theta$ and $\phi > 0$. Note that $\phi = |g|$ and $\theta = |g_{sa}|$.

A privacy-preserving publishing mechanism will publish noisy answers by adding some noise to the answer to each query. Let $X = \phi + \xi_1$ and $Y = \theta + \xi_2$ be the noisy answers for Q_1 and Q_2 returned by some privacy mechanism, where ξ_i's are the noises. Note $\frac{\theta}{\phi} \leq 1$ and $\frac{\theta}{\phi}$ represents the chance that t has the sa value on SA. Note $\frac{Y}{X} = \frac{\theta+\xi_2}{\phi+\xi_1} = \frac{\theta/\phi+\xi_2/\phi}{1+\xi_1/\phi}$. The adversary gets the noisy answers X and Y, instead of the true answers ϕ and θ, and uses $\frac{Y}{X}$ to estimate the true confidence $\frac{\theta}{\phi}$. $\frac{Y}{X}$ is called the *noisy confidence* of sa in g.

The intuition that $\frac{Y}{X}$ may lead to a disclosure is as follows. For any ξ_i of a fixed scale, as the answer ϕ increases, ξ_2/ϕ and ξ_1/ϕ decrease and $\frac{Y}{X}$ approaches $\frac{\theta}{\phi}$. If $\frac{\theta}{\phi}$ is large enough (which is application specific), the adversary learns that t has the sensitive value sa with a high probability. This construction is general because it does not assume record correlation (record correlation is required in [47] for sensitive information disclosure analysis) and does not depend on the noise distribution except that the noises have a fixed scale. Below, we formalize this intuition. First, we show a lemma.

Lemma 1. *Let ϕ and θ be the true answers to Q_1 and Q_2, $\phi \neq 0$. Let $X = \phi + \xi_1$ and $Y = \theta + \xi_2$ be the noisy answers for Q_1 and Q_2 with the noises ξ_i having the zero mean and the variance V. Then*

$$E[\tfrac{Y}{X}] \simeq \tfrac{\theta}{\phi}(1 + \tfrac{V}{\phi^2}) \text{ and } Var[\tfrac{Y}{X}] \simeq \tfrac{V}{\phi^2}(1 + \tfrac{\theta^2}{\phi^2})$$

Proof. Note that $E[\frac{Y}{X}]$ is *not* equal to $\frac{E[Y]}{E[X]}$. Using the Taylor expansion technique [31, 74], $E[\frac{Y}{X}]$ and $Var[\frac{Y}{X}]$ can be approximated as follows:

$$E[\frac{Y}{X}] \simeq \frac{E[Y]}{E[X]} + \frac{cov[X,Y]}{E[X]^2} + \frac{Var[X]E[Y]}{E[X]^3}$$

$$Var[\frac{Y}{X}] \simeq \frac{Var[Y]}{E[X]^2} - \frac{2E[Y]}{E[X]^3}cov[X,Y] + \frac{E[Y]^2}{E[X]^4}Var[X]$$

The error of the approximation is the remaining terms of the Taylor expansion that are dropped. $E[X] = \phi$ and $E[Y] = \theta$ (because noises have the zero mean), $Var[X] = Var[Y] = V$, and the covariance $cov[X,Y] = cov[\phi + \xi_1, \theta + \xi_2] = cov[\xi_1, \xi_2]$. Since ξ_1 and ξ_2 are unrelated, $cov[\xi_1, \xi_2] = 0$. Substantiating these into the above equations and simplifying, we get $E[\frac{Y}{X}]$ and $Var[\frac{Y}{X}]$ as required. □

For any noise distribution with the zero mean and a fixed variance V, as the query answer ϕ increases, $\frac{V}{\phi^2}$ decreases, $E[\frac{Y}{X}]$ approaches $\frac{\theta}{\phi}$ and $Var[\frac{Y}{X}]$ approaches zero. In general, $E[\frac{Y}{X}]$ approaching $\frac{\theta}{\phi}$ does not entail $\frac{Y}{X}$ approaching $\frac{\theta}{\phi}$, for particular instances X and Y. However, if $Var[\frac{Y}{X}]$ approaches zero, the deviation of $\frac{Y}{X}$ from $E[\frac{Y}{X}]$ approaches zero, $\frac{Y}{X}$ approaches $\frac{\theta}{\phi}$. This is summarized in the next corollary.

Corollary 1. *Let ϕ and θ be the true answers to Q_1 and Q_2, $\phi \neq 0$. Let $X = \phi + \xi_1$ and $Y = \theta + \xi_2$ be the noisy answers for Q_1 and Q_2, where ξ_1 and ξ_2 are injected noises. For any noise distribution with the zero mean and a fixed variance V, as the query answer ϕ increases, $\frac{Y}{X}$ approaches $\frac{\theta}{\phi}$.*

To our knowledge, Corollary 1 covers all noise distributions employed by the differential privacy mechanism, including Laplace mechanism [27], Gaussian mechanism [30], and Matrix mechanism [55], because these distributions have a zero mean and a fixed variance. To see how large ϕ is needed for $\frac{Y}{X}$ to be accurate enough for $\frac{\theta}{\phi}$, let us consider the Laplace mechanism $Lap(b) = \frac{1}{2b}exp(-|\xi|/b)$, where b is the *scale factor*, and a similar analysis can be performed for other mechanisms. $Lap(b)$ has the zero mean and the variance $V = 2b^2$. The setting $b = \Delta/\epsilon$ ensures ϵ-differential privacy, where Δ is the *sensitivity* of the queries of interest, which roughly denotes the worst-case change in the query answer on changing one record in any possible database. Δ is a property of the queries, not a property of the database. Hence, V is fixed for a given query class and Corollary 1 applies to $Lap(b)$. Substituting $\frac{\theta}{\phi} \leq 1$ and $\frac{V}{\phi^2} = \frac{2b^2}{\phi^2} = 2\left(\frac{b}{\phi}\right)^2$ into Lemma 1 and simplifying, we get a simple bound on $|E[\frac{Y}{X}] - \frac{\theta}{\phi}|$ and $Var[\frac{Y}{X}]$ in terms of the scale factor b and the query answer ϕ (but not the privacy parameter ϵ or the sensitivity Δ of queries).

Corollary 2. *Let ϕ and θ be the true answers to Q_1 and Q_2, $\phi \neq 0$. Let $X = \phi + \xi_1$ and $Y = \theta + \xi_2$ be the noisy answers for Q_1 and Q_2, where ξ_1 and ξ_2 are injected noise. b is the scale factor of the injected Laplace noise. (i) $|E[\frac{Y}{X}] - \frac{\theta}{\phi}| \leq 2\left(\frac{b}{\phi}\right)^2$. (ii) $Var[\frac{Y}{X}] \leq 4\left(\frac{b}{\phi}\right)^2$.*

Thus, the value of $2\left(\frac{b}{\phi}\right)^2$ is an indicator of how close $\frac{Y}{X}$ is to $\frac{\theta}{\phi}$. Table 3.2 shows the values of $2\left(\frac{b}{\phi}\right)^2$ for various query answers ϕ and settings of b (the corresponding privacy parameter ϵ for the

Table 3.2: $2\left(\frac{b}{\phi}\right)^2$

b \ ϕ	5000	1000	500	200	100
$b = 10$ ($\epsilon = 0.2$)	**0.000008**	**0.0002**	**0.0008**	**0.005**	0.02
$b = 20$ ($\epsilon = 0.1$)	**0.000032**	**0.0008**	**0.0032**	0.02	0.08
$b = 40$ ($\epsilon = 0.05$)	**0.000128**	**0.0032**	0.0128	0.08	0.32
$b = 200$ ($\epsilon = 0.01$)	**0.0032**	0.08	0.32	2	8

setting of $\Delta = 2$ is shown within the brackets, which accounts for answering the two queries Q_1 and Q_2 in a row). The boldface highlights where $2\left(\frac{b}{\phi}\right)^2$ is small enough so that $\frac{Y}{X}$ is a good indicator of $\frac{\theta}{\phi}$. Take $(b = 20, \phi = 500)$ as an example where $2\left(\frac{b}{\phi}\right)^2 = 0.0032$. $|E[\frac{Y}{X}] - \frac{\theta}{\phi}| \leq 0.0032$ and $Var[\frac{Y}{X}] \leq 0.0032 \times 2 = 0.0064$. Indeed, Corollary 2 quantifies a condition of the occurrence of disclosures in terms of $\frac{b}{\phi}$: as a rule of thumb, a ratio $\frac{b}{\phi} \leq \frac{1}{20}$ would ensure that $\frac{Y}{X}$ is a good indicator of $\frac{\theta}{\phi}$ because $2\left(\frac{b}{\phi}\right)^2 \leq \frac{2}{400}$. In this case, if $\frac{\theta}{\phi}$ is high enough to be considered as sensitive, a sensitive disclosure would occur through accessing noisy answers X and Y. This condition also suggests that such disclosures can not be avoided by choosing a large scale factor b if the actual answer ϕ can be arbitrarily large.

$\frac{Y}{X}$ being close to $\frac{\theta}{\phi}$ is a necessary condition for inferring $\frac{\theta}{\phi}$, but accurate inference does not always lead to a disclosure of sensitive information. A disclosure also requires that $\frac{\theta}{\phi}$ is significantly higher than the *prior knowledge* in the absence of t's information on g. Otherwise, the accurate inference is not useful because prior knowledge is known to the adversary before she/he launches the attack. In this work, we consider the true confidence of sa in the entire data set $\mathcal{D}$ as the prior knowledge, that is, $\frac{|\mathcal{D}_{sa}|}{|\mathcal{D}|}$, where $|\mathcal{D}_{sa}|$ is the number of records in $\mathcal{D}$ containing sa on SA. When $\frac{\theta}{\phi}$ is significantly higher than $\frac{|\mathcal{D}_{sa}|}{|\mathcal{D}|}$, t's information on g has identified a group of individuals, one of them being t, that are more frequently associated with sa than general people in the data set. In this sense, we say that a disclosure occurs to t.

3.2.2 Definition of Disclosures

From the above discussion, a disclosure occurs when both conditions are satisfied: (A) the noisy confidence is close enough to the true confidence; (B) the true confidence is much larger than the prior. For (A), we introduce the *closeness probability*, CP_τ, to calibrate the closeness of the true confidence and the noisy confidence, as defined below.

Definition 3. (CP_τ, Closeness Probability) *For $\tau > 0$, the closeness probability for a SA value sa in a personal group g is*

$$CP_\tau = \Pr\left[\left|\frac{\frac{\theta}{\phi} - \frac{Y}{X}}{\frac{\theta}{\phi}}\right| \leq \tau\right]. \tag{3.2}$$

where θ and ϕ (X and Y) are true (noisy) answers to Q_1 and Q_2 as in Equation (3.1).

In the definition of CP_τ, τ is the parameter indicating the extent of the closeness between the true confidence and the noisy confidence. Since disclosure requires the noisy confidence to be close to the true confidence, we require $0 < \tau < 1$. In other words, when the absolute relative error of the two confidences is no smaller than 100% we consider the noisy confidence is too coarse to be adopted for estimating the true confidence accurately. A larger CP_τ and a smaller τ imply that the noisy confidence is a more accurate estimate of the true confidence. CP_τ is the probability that the relative error of the noisy confidence $\frac{Y}{X}$ compared to the true confidence $\frac{\theta}{\phi}$ is within the closeness parameter, τ. And a larger value of CP_τ suggests a disclosure could be more likely found. CP_τ is a probability, therefore, the value of CP_τ falls in the range of $(0, 1]$.

For (B), we introduce the variable *jump*, $\mathcal{J}$ to calibrate the distance from the true confidence, $\frac{\theta}{\phi}$, to the prior knowledge, f. In this thesis we always assume that the prior knowledge on some SA value sa is the global distribution of sa in overall data set (i.e., $f = \frac{|\mathcal{D}_{sa}|}{|\mathcal{D}|}$).

Definition 4. ($\mathcal{J}$, Jump) *For $\tau > 0$, the jump for a SA value sa in a personal group g is*

$$\mathcal{J} = \frac{\theta/\phi}{f} \tag{3.3}$$

where θ and ϕ are true answers to Q_1 and Q_2 as in Equation (3.1), f is the prior knowledge obtained by adversaries on sa and $f = \frac{|\mathcal{D}_{sa}|}{|\mathcal{D}|}$.

A larger value of $\mathcal{J}$ suggests that individuals in g have significantly larger probability of having sa as SA values compared to general individuals in the data set. Therefore, a disclosure may occur.

Definition 5. (Disclosure) *Given the thresholds $\tau > 0$, $\mathcal{K}_\mathcal{J} > 1$, $0 < \mathcal{K}_{CP} \leq 1$, and a personal group g, a value sa is* disclosed *in g wrt $(\tau, \mathcal{K}_\mathcal{J}, \mathcal{K}_{CP})$ if (1) $CP_\tau \geq \mathcal{K}_{CP}$ and (2) $\mathcal{J} \geq \mathcal{K}_\mathcal{J}$.*

$\mathcal{K}_{CP}$ and $\mathcal{K}_\mathcal{J}$ are given thresholds for CP_τ and $\mathcal{J}$, respectively. A smaller τ, a larger $\mathcal{K}_{CP}$, and a larger $\mathcal{K}_\mathcal{J}$ suggest a more severe disclosure, implying that the confidence can be estimated from the observed noisy confidence with better accuracy, and the confidence is much higher than its prior. The above definition of disclosures is consistent with the literature, e.g., β-likeness [14], in that they both try to limit the difference between the global distribution of some SA value sa and the distribution of such sa in a personal group g (i.e., condition (2)). Other than this constraint, we

Table 3.3: Notations in Chapter 3

Notations	Explanation
$\mathcal{D}$	the data set
$\mathcal{D}_{sa}$	the subset of $\mathcal{D}$ with sa on SA
t	a target individual
sa	a domain value of SA
g	a personal group
g_{sa}	the subset of g with sa on SA
$\phi = \|g\|, \theta = \|g_{sa}\|$	true answers for query Q_1 and Q_2 in (3.1)
X, Y	noisy answers corresponding to ϕ and θ
$\frac{\theta}{\phi}, \frac{Y}{X}$	true confidence and noisy confidence of sa in g

also consider the difference of confidence within a personal group before and after adding the noise, i.e., condition (1). This is because with the injected noise, the adversary is not able to get the true confidence of sa in g, instead, she/he has to use the noisy confidence to gauge the true confidence, therefore, a close noisy confidence is necessary for enabling a disclosure. Table 3.3 outlines the notation used in this section.

3.3 Disclosure of Differential Privacy

In this section, we study the risk of disclosures measured by Definition 5 with noisy answers published by a randomization mechanism satisfying a differential privacy guarantee.

3.3.1 Computing CP_τ

With the Laplace noise distribution $Lap(b)$, we can now derive for CP_τ in Definition 3. Recall that ϕ and θ are the actual query answers, and X and Y denote the variables for the noisy version of ϕ and θ after adding a Laplace noise. So, the noises $x - \phi$ and $y - \theta$ follow the Laplace distribution:

$$f_X(x) = \frac{1}{2b} \exp\left(-\frac{|x - \phi|}{b}\right) \tag{3.4}$$

$$f_Y(y) = \frac{1}{2b} \exp\left(-\frac{|y - \theta|}{b}\right) \tag{3.5}$$

Notice that ϕ is the mean of X and θ is the mean of Y because Laplace noises have the zero mean. Also, we assume that $0 < \theta \leq \phi$, $\phi > 0$ and $b > 0$.

It is easy to see that CP_τ is equal to

$$\Pr\left[\left|\frac{\theta}{\phi} - \frac{Y}{X}\right| \leq \tau'\right] \tag{3.6}$$

where $\tau' = \tau \times \frac{\theta}{\phi} > 0$. To compute the probability in (3.6), we define the following cumulative function for the ratio of two Laplace variables, $Z = \frac{Y}{X}$,

$$F_Z(z) = \Pr[Z \leq z]$$

Most studies on the ratio of two random variables [73][65] focus on situations when at least one variable follows some distribution other than Laplace. To our knowledge, this is the first study on the probability of ratio distribution of two Laplace variables. The disclosure considered here, the ratio of two random variables, G/H, has other applications. One example is the *stress-strength* model [49] in the context of *reliability*. In the simplest form, this model estimates the reliability of a component with *strength* under *stress*. G represents the stress encountered by the component and H stands for the *strength* of the component to beat stress. The component fails when the stress exceeds or equals to the strength and functions properly whenever $G < H$. Therefore, the reliability of the component is defined as the probability of not failing, i.e., $\Pr[G < H]$ or $\Pr[G/H \leq z]$ for some $z < 1$. In a similar spirit, the ratio probability can also be applied in the case of *demand-supply* where the two are modeled as random variables.

Let $F_Z^1(z)$ be $F_Z(z)$ for $z < 0$, let $F_Z^2(z)$ be $F_Z(z)$ for $0 < z \leq \frac{\theta}{\phi}$, and let $F_Z^3(z)$ be $F_Z(z)$ for $z > \frac{\theta}{\phi}$. Lemma 2 computes the cumulative density function of the ratio of two Laplace random variables. The proof of Lemma 2 can be found in Section A.

Lemma 2. *Assume $z \neq 0$ and $z \neq \pm 1$.*

For $z < 0$,

$$\begin{aligned} F_Z^1(z) = &\left[\frac{z^2}{2(1-z^2)}\right]\left[\exp\left(\frac{\theta - z\phi}{zb}\right)\right] - \frac{1}{2(z+1)}\left[\exp\left(\frac{-(\theta+\phi)}{b}\right)\right] \\ &+ \frac{1}{2}\exp\left(-\frac{\phi}{b}\right) - \frac{1}{2(z^2-1)}\left[\exp\left(\frac{z\phi - \theta}{b}\right)\right]; \end{aligned} \tag{3.7}$$

For $0 < z \leq \frac{\theta}{\phi}$,

$$\begin{aligned} F_Z^2(z) = &\left[\frac{z^2}{2(z^2-1)}\right]\left[exp\left(\frac{z\phi - \theta}{zb}\right)\right] - \frac{1}{2(z+1)}\left[exp\left(\frac{-(\theta+\phi)}{b}\right)\right] \\ &+ \frac{1}{2}exp\left(-\frac{\phi}{b}\right) - \frac{1}{2(z^2-1)}\left[exp\left(\frac{z\phi - \theta}{b}\right)\right]; \end{aligned} \tag{3.8}$$

For $z > \frac{\theta}{\phi}$,

$$F_Z^3(z) = \frac{z^2}{2(1-z^2)}\left[exp\left(\frac{\theta - z\phi}{zb}\right)\right] - \frac{1}{2(z+1)}\left[exp\left(\frac{-(\theta+\phi)}{b}\right)\right] + 1 + \frac{1}{2}exp\left(-\frac{\phi}{b}\right) - \frac{1}{2(1-z^2)}\left[exp\left(\frac{\theta - z\phi}{b}\right)\right]. \tag{3.9}$$

□

Theorem 2. *Assume that* $\theta > 0$, $\phi > 0$ *and* $\tau' > 0$.

1. $CP_\tau = \Pr\left[\left|\frac{\theta}{\phi} - \frac{Y}{X}\right| \leq \tau'\right]$ *is*

$$\begin{cases} F_Z^3\left(\frac{\theta}{\phi} + \tau'\right) - F_Z^1\left(\frac{\theta}{\phi} - \tau'\right) & \text{if } \frac{\theta}{\phi} - \tau' < 0 \\ F_Z^3\left(\frac{\theta}{\phi} + \tau'\right) - F_Z^2\left(\frac{\theta}{\phi} - \tau'\right) & \text{if } \frac{\theta}{\phi} - \tau' > 0 \end{cases} \tag{3.10}$$

2. $\lim_{\phi\to\infty} \Pr\left[\left|\frac{\theta}{\phi} - \frac{Y}{X}\right| \leq \tau'\right] = 1$.

Proof. 1) Let $Z = \frac{Y}{X}$ and the cumulative probability function of Z could be expressed as $F_Z^1(z)$, $F_Z^2(z)$ or $F_Z^3(z)$ when z falls into different ranges in Lemma 2. Rewrite $\left|\frac{\theta}{\phi} - Z\right| \leq \tau'$ into $\frac{\theta}{\phi} - \tau' \leq Z \leq \frac{\theta}{\phi} + \tau'$, and $CP_\tau = Pr\left[\left|\frac{\theta}{\phi} - Z\right| \leq \tau'\right] = F_Z\left(\frac{\theta}{\phi} + \tau'\right) - F_Z\left(\frac{\theta}{\phi} - \tau'\right)$. From the result of Lemma 2, because $\tau' > 0$, $\left(\frac{\theta}{\phi} + \tau'\right) > \frac{\theta}{\phi}$, $F_Z\left(\frac{\theta}{\phi} + \tau'\right)$ is always represented in the form of $F_Z^3\left(\frac{\theta}{\phi} + \tau'\right)$, $F_Z\left(\frac{\theta}{\phi} - \tau'\right)$ is expressed in the form of $F_Z^1\left(\frac{\theta}{\phi} - \tau'\right)$ when $\frac{\theta}{\phi} - \tau' < 0$, and $F_Z\left(\frac{\theta}{\phi} - \tau'\right)$ is expressed in the form of $F_Z^2\left(\frac{\theta}{\phi} - \tau'\right)$ when $\frac{\theta}{\phi} - \tau' > 0$. This shows Theorem 2, 1).

2) As ϕ approaches to ∞, all exp terms in F_Z^1, F_Z^2 and F_Z^3 in (3.10) approach 0 because the exponents are negative. Thus, $\Pr\left[\left|\frac{\theta}{\phi} - \frac{Y}{X}\right| \leq \tau'\right]$ approaches to 1. □

Theorem 2 provides the theoretical basis for why the noisy confidence is a good estimate of the true confidence when ϕ is large. It also provides a way to compute the probability in (3.6), thus, CP_τ. By replacing τ' with $\tau \times \frac{\theta}{\phi}$ in (3.10), we have the following computation of CP_τ.

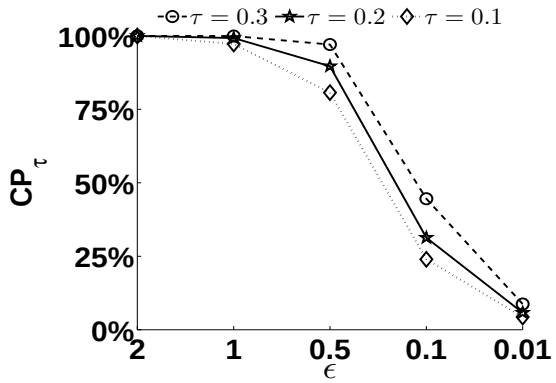
Corollary 3.

$$CP_\tau = \begin{cases} F_Z^3\left((1+\tau)\frac{\theta}{\phi}\right) - F_Z^1\left((1-\tau)\frac{\theta}{\phi}\right) & \text{if } \tau > 1 \\ F_Z^3\left((1+\tau)\frac{\theta}{\phi}\right) - F_Z^2\left((1-\tau)\frac{\theta}{\phi}\right) & \text{if } 0 < \tau < 1 \end{cases} \tag{3.11}$$

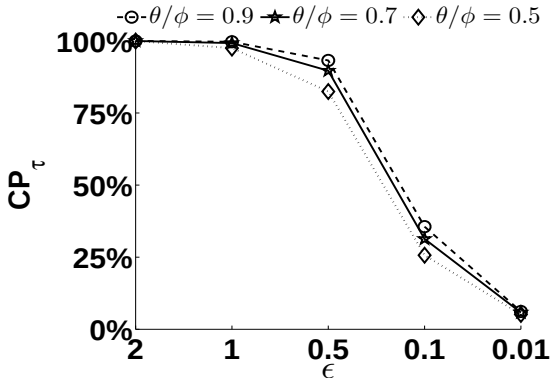
The value of CP_τ depends on τ, b, ϕ, and θ/ϕ. Let us evaluate CP_τ for typical values of these parameters. The settings for all parameters can be found in Table 3.4 with the default settings in bold face. Since our disclosure considers the distance from the true confidence θ/ϕ to the prior, it is unlikely to have a disclosure when true confidence is small. Thus, here we consider relatively large true confidece, i.e., $\{0.5, 0.7, 0.9\}$. The range of ϵ has been discussed in Section 2.2.2 and the same

Table 3.4: Parameter Table

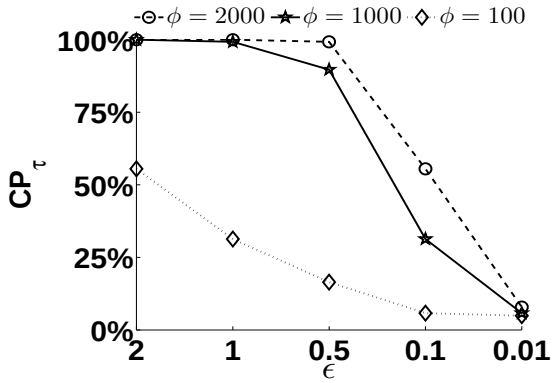
Parameters	Settings
$\frac{\theta}{\phi}$	0.5, **0.7**, 0.9
τ	0.1, **0.2**, 0.3
ϕ	100, **1000**, 2000
ϵ	2, 1, **0.5**, 0.1, 0.01
Δ	**25**



(a) vs. τ ($\frac{\theta}{\phi} = 0.7$, $\phi = 1000$)



(b) vs. $\frac{\theta}{\phi}$ ($\tau = 0.2$, $\phi = 1000$)



(c) vs. ϕ ($\tau = 0.2$, $\frac{\theta}{\phi} = 0.7$)

Figure 3.1: CP_τ ($\Delta = 25$)

range $[0.01, 2]$ is applied here. τ represents the closeness extent between the noisy confidence and the true confidence and a smaller τ means a more severe disclosure. We set τ for $\{0.1, 0.2, 0.3\}$ and we want to find out disclosures when the relative error of using noisy confidence to estimate the true confidence is $\{10\%, 20\%, 30\%\}$. Recall that the scale factor b of the Laplace noise is determined by both the sensitivity (i.e., Δ) and the differential privacy parameter ϵ. We fix Δ to 25 and change the value of ϵ. This choice of Δ is comparable with those derived from real life data sets in our experiments (Section 3.4).

Table 3.5: Attributes in Data Sets

Data Set	Attributes (Domain Size)
EDU	Gender (2), Occupation (50), Marital (6), Race (9), **Education (14)**
OCC	Gender (2), Education (14), Marital (6), Race (9), **Occupation (50)**
SALARY	Gender (2), Education (14), Marital (6), Race (9), Work-class (7), Country (69), **Salary (50)**

Figure 3.1 shows how CP_τ varies with respect to τ, $\frac{\theta}{\phi}$ and ϕ under various levels of ϵ-differential privacy. Several points can be observed: i) a smaller ϵ (i.e., adding more noises), which imposes a more restricted differential privacy, could help prevent the true confidence and the noisy confidence from being too close; ii) as τ increases, CP_τ increases as well, but τ does not affect the value of CP_τ much; iii) when the true confidence is larger, it is easier for the noisy confidence to be closer to the true confidence; iv) a larger ϕ results in a larger CP_τ; in other words, when ϕ is large, the true confidence and the noisy confidence tend to be closer. This is consistent with Theorem 2.

In sum, adding noises of a larger scale (by choosing a small ϵ) and having smaller ϕ are the two leading factors that could help prevent the noisy confidence from being too close to the true confidence. For a given data set, the values of ϕ and θ are fixed for given queries, and increasing the scale of noises is the only way to protect the true confidence. However, adding larger noise inevitably degrades the utility of data analysis because both noisy confidence and utility depend on noisy query answers. This suggests a limitation of differential privacy: a good utility leads to a bad disclosure, and elimination of disclosure leads to elimination of utility as well. We will investigate this relationship on real life data sets in the next section.

3.4 Reality Check

Section 3.3 modeled the disclosure when differential privacy is applied. In this section, we investigate the extent to which private information of an individual may be disclosed through NIR by query answers that satisfy differential privacy in *real life data sets*. We also study the impact on utility by simply using larger noises to remove such disclosures.

3.4.1 Experimental Setup

All experiments were implemented in Python and ran on an Intel Xeon(R) E5630 CPU 2.53GHZ PC with 12GB of RAM. All three data sets *SALARY*, *EDU* and *OCC* we used were extracted from the US census data [2] about personal information of American adults. *SALARY* was used in [79, 14], and both *EDU* and *OCC* were used in [79, 16]. All data sets have the same number of records (i.e.,

[2]Downloadable from `http://www.ipums.org`.

Table 3.6: The Data Set and Sample Marginals in Example 3

(a) A data set $\mathcal{D}$

Age	Gender	Occupation	Disease
23	F	Lawyer	Flu
35	F	Engineer	HIV
46	M	Engineer	Flu
30	M	Lawyer	HIV
50	M	Engineer	Flu
33	F	Lawyer	HIV

(b) A 2-D marginal

Gender	Occupation	Count
M	Lawyer	1
M	Engineer	2
F	Lawyer	2
F	Engineer	1

(c) A 2-D_{SA} marginal

Gender	Occupation	Disease	Count
M	Lawyer	Flu	0
M	Lawyer	HIV	1
M	Engineer	Flu	2
M	Engineer	HIV	0
F	Lawyer	Flu	1
F	Lawyer	HIV	1
F	Engineer	Flu	0
F	Engineer	HIV	1

500k). Table 3.5 shows the information of attributes with domain size in brackets, and the SA is marked in bold.

3.4.2 Publishing Scenarios

Many typical data analysis tasks make use of low dimensional marginals [7, 78], where each marginal corresponds to some subset of attributes with each row being the count for one combination of the attribute values. Therefore, a marginal consists of the answers for all the count queries over those attributes. For a given data set, we considered publishing all 2-D (two dimensional) marginals on the attributes in NA, and for each, publishing the corresponding marginal expanded with the extra attribute SA, denoted by 2-D_{SA}. There are $\binom{|NA|}{2}$ 2-D marginals and the same number of 2-D_{SA} marginals, where $|NA|$ is the number of distinct attributes in NA. $|\mathcal{M}| = 2 \cdot \binom{|NA|}{2}$ is the total number of marginals. Example 3 below illustrates instances of marginals on a raw data set $\mathcal{D}$.

Example 3. *Assume that a data set $\mathcal{D}$ has three NA:* Age*,* Gender *and* Occupation*, and one SA:* Disease*. Tables 3.6a, 3.6b, and 3.6c illustrate $\mathcal{D}$, a 2-D marginal on $\{Gender, Occupation\}$, and the corresponding 2-D_{SA} marginal on $\{Gender, Occupation, Disease\}$, respectively. $|NA|$ in*

this data set is 3*, thus, there are* $\binom{3}{2}$ *(i.e., three)* 2-D *marginals and three* 2-D_{SA} *marginals in total. The three* 2-D *marginals are* $\{Age, Gender\}$, $\{Age, Occupation\}$ *and* $\{Gender, Occupation\}$*, and the three* 2-D_{SA} *marginals are* $\{Age, Gender, Disease\}$, $\{Age, Occupation, Disease\}$ *and* $\{Gender, Occupation, Disease\}$*. Here the counts in the marginals are based on the raw data set before adding noises. The published marginals will contain noisy counts to satisfy differential privacy. Since noises are added independently for each count, the noisy counts in* 2-D *marginals will not be the aggregated results of corresponding* 2-D_{SA} *marginals.*

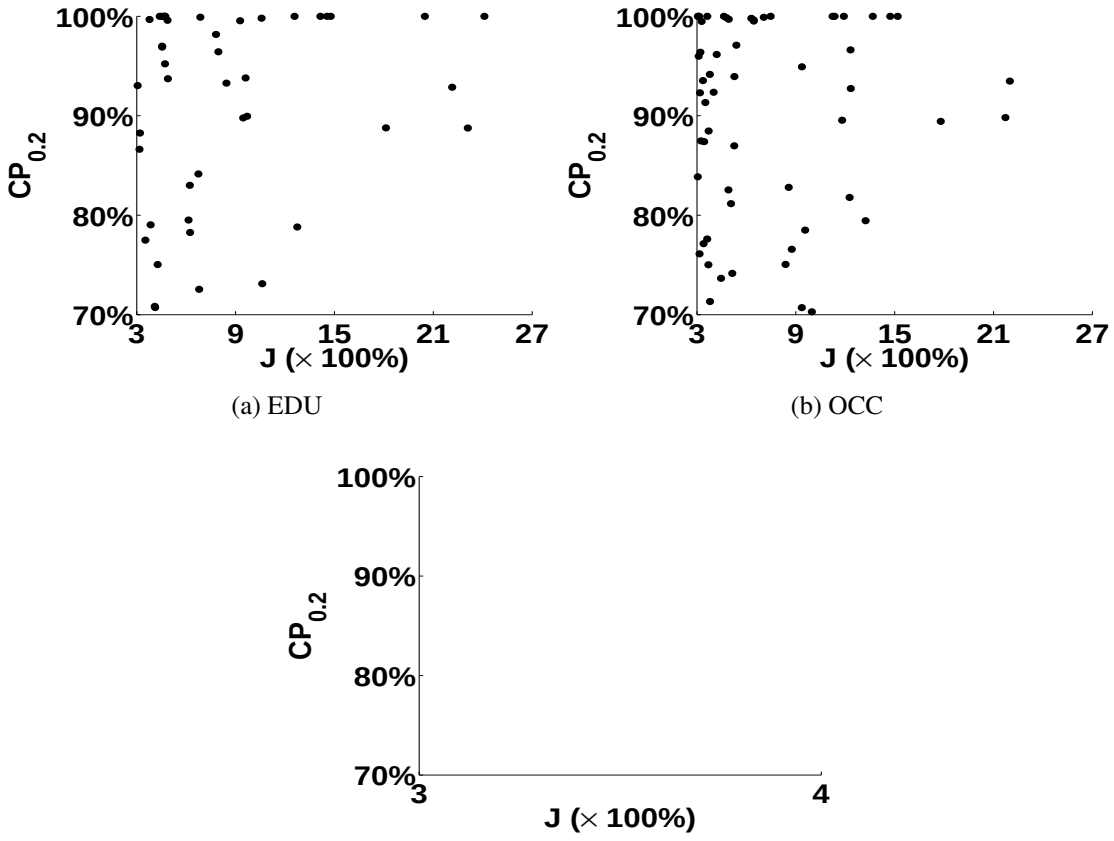
In the classic mechanism in [27], ϵ-differential privacy is achieved by adding noises (to a query answer) following the distribution $Lap\left(\frac{\Delta}{\epsilon}\right)$, where the sensitivity Δ is

$$\Delta = 2 \cdot |\mathcal{M}| = 4 \cdot \binom{|NA|}{2}. \tag{3.12}$$

This is because changing one record affects at most 2 counts in a marginal and there are $|\mathcal{M}|$ marginals. For example, in Table 3.6b, assume Bob is male lawyer in the data set $\mathcal{D}$. If his occupation is changed to Engineer, then two counts are affected: the count to $\{Male\ \&\ Lawyer\}$ becomes 0 and the count to $\{Male\ \&\ Engineer\}$ becomes 3. Note that if only deleting and adding a record are allowed in the data set then changing one record affects at most 1 count in a marginal, in which case our result could be adjusted by simply reducing the sensitivity by half. Both *EDU* and *OCC* have 4 distinct NA, therefore Δ is $4 \times \binom{4}{2} = 24$. *Salary* has 6 distinct NA and Δ in *Salary* is $4 \times \binom{6}{2} = 60$. Some methods [81, 79, 78] improve the utility by adding less noises while achieving the same level of privacy. Since our goal is to study the extent to which disclosures occur, if disclosures occur for the above classic mechanism, disclosures also occur for the improved methods that generate more accurate answers. For this reason, we shall focus on the classic mechanism.

3.4.3 Disclosures

Given thresholds $\mathcal{K}_{CP}$ and $\mathcal{K}_{CP}$, in a personal group g, a SA value sa is *disclosed* if the following two conditions are satisfied (as in Definition 5): (1) $CP_\tau \geq \mathcal{K}_{CP}$, where CP_τ is defined in Definition 3 as the probability that the noisy confidence is close to the true confidence, where the closeness is calibrated by parameter τ; (2) $\mathcal{J} \geq \mathcal{K}_{\mathcal{J}}$, and $\mathcal{J}$ is defined in Definition 4 as the distance from the true confidence to the prior knowledge. Condition (1) suggests that the probability that the noisy confidence is close to the true confidence is large (i.e., at least $\mathcal{K}_{CP}$). Condition (2) means that the true confidence in the personal group g is much higher than the prior (i.e., the true confidence is at least $\mathcal{K}_{\mathcal{J}}$ times higher than the prior).



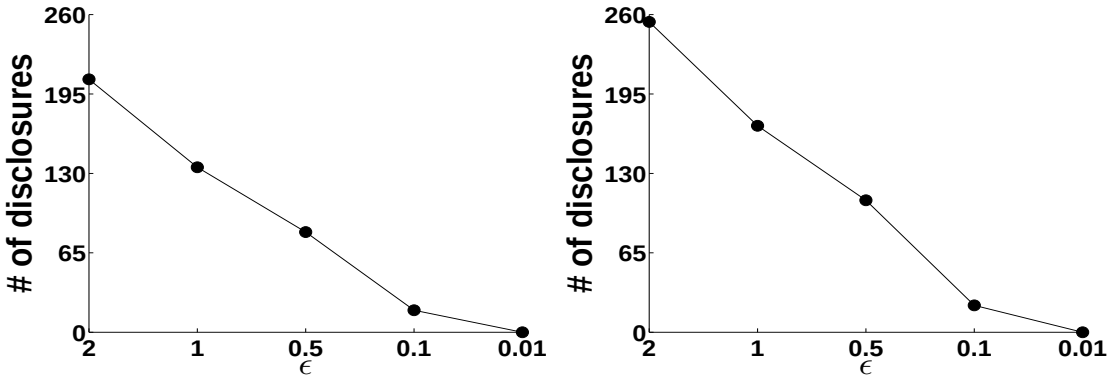
(a) EDU

(b) OCC

(c) SALARY

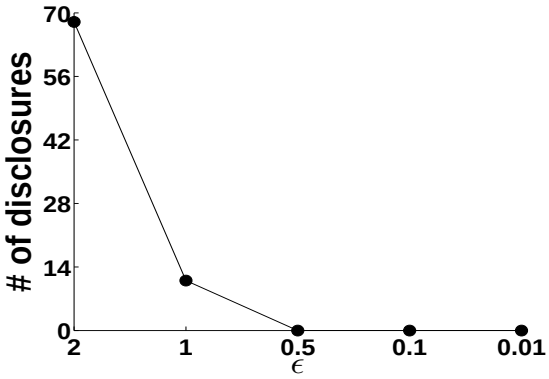
Figure 3.2: Disclosures in Terms of CP_τ and $\mathcal{J}$ ($\tau = 0.2, \mathcal{K}_{CP} = 0.7, \mathcal{K}_\mathcal{J} = 3, \epsilon = 0.5$)

Figure 3.2 shows the disclosures found in the three data sets for the privacy setting $\epsilon = 0.5$, $\tau = 0.2$, $\mathcal{K}_{CP} = 0.7$, and $\mathcal{K}_\mathcal{J} = 3$. One disclosure means that under 0.5-differential privacy, for some SA value sa in a personal group g, the probability that the noisy confidence is close to the true confidence (with a relative error smaller than 20%) is at least 70%, and the true confidence is at least three times larger than the prior of sa to the true confidence. Each point in this figure stands for one disclosure for some SA value sa with the values of $\mathcal{J}$ and CP_τ being represented by the x-axis and y-axis, respectively. As we can see, both *EDU* and *OCC* suffer from many disclosures. In fact, for several disclosures, $\mathcal{J}$ is more than 20 and $CP_{0.2}$ is nearly 100%. In these cases, the target individual belongs to a personal group, identified by her/his known NA, and it is 20 times more likely to have some SA value in this group than in the entire data set and the noisy query answers can be used to learn this information with a relative error of at most 20% in nearly 100% of cases (i.e., $CP_{0.2} \approx 100\%$).



(a) EDU

(b) OCC

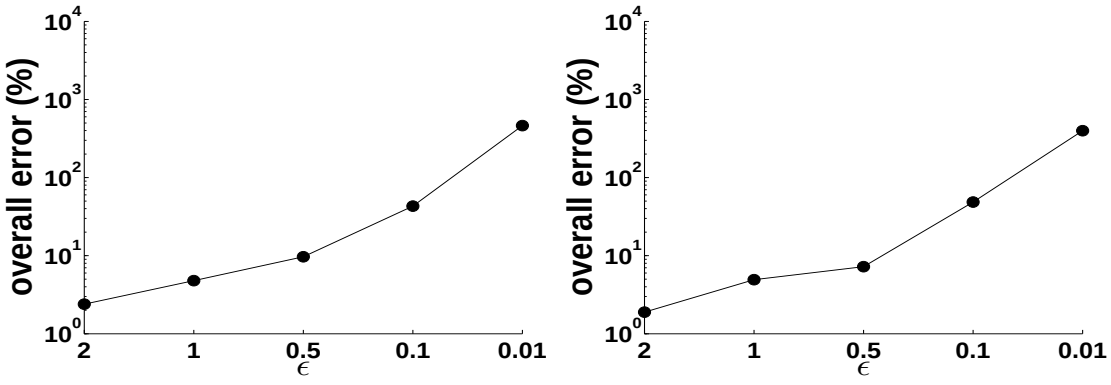


(c) SALARY

Figure 3.3: The Number of Disclosures vs ϵ ($\tau = 0.2$, $\mathcal{K}_{CP} = 0.7$, $\mathcal{K}_{\mathcal{J}} = 3$)

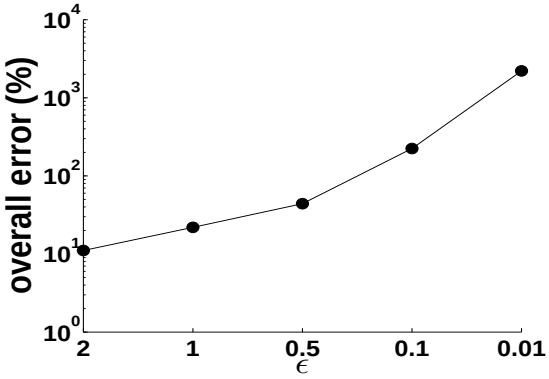
In contrast, for the above settings of parameters, no disclosure is observed on *SALARY* because noises of a larger scale are added to query answers, which reduces the occurrence of disclosures. *SALARY* has more NA attributes, thus, a larger $|\mathcal{M}|$ and a larger sensitivity Δ according to Equation (3.12), which leads to a larger scale factor b for a given privacy setting ϵ.

Figure 3.3 shows the number of disclosures for the three data sets when $\tau = 0.2$, $\mathcal{K}_{CP} = 0.7$ and $\mathcal{K}_{\mathcal{J}} = 3$ for various differential privacy settings ϵ. The number of disclosures in a personal group g is counted as the number of disclosed SA values in g, and the total number of disclosures of the data set is counted as the sum of the disclosures in all personal groups. Under a more relaxed privacy setting for ϵ, i.e., > 0.5, disclosures are still observed for *SALARY*. By reducing ϵ to 0.01, no disclosure is found because the increased noise scale reduces the closeness probability CP_τ, therefore, it is harder to find a disclosure. And this is consistent with the findings in Figure 3.1. However, as we shall see shortly, the increased scale of noises renders a large error in query answers, which degrades the utility for data analysis.



(a) EDU

(b) OCC



(c) SALARY

Figure 3.4: Overall Error

3.4.4 Utility

The study in Section 3.4.3 suggests that setting the differential privacy parameter ϵ to a small value helps prevent disclosures. While this addresses the privacy issue, it affects adversely the utility of published query answers. We consider all queries corresponding to the rows in all the 2-D and 2-D_{SA} marginals defined in Section 3.4.2, and we measure the utility of a noisy query answer by the *relative error* of the answer to such queries. For each query, suppose that ans denotes the true answer and noi denotes the noisy answer. The relative error is defined as $err = \frac{|noi-ans|}{\max\{ans,\delta\}}$, where δ is the *sanity bound* specified by the user to eliminate the effect of unreasonably small query results. As in [78], we set $\delta = 10^{-4} \times |\mathcal{D}|$, where $|\mathcal{D}|$ is the number of records in the data set $\mathcal{D}$. Since $|\mathcal{D}| = 500K$ in our experiments, $\delta = 50$. The utility for all queries is evaluated by the *overall error*, defined as the averaged relative error.

It is interesting to cross-examine the overall error in Figure 3.4 and the number of disclosures in Figure 3.3. While a smaller ϵ helps prevent disclosures, it degrades the utility by a larger overall

error. For example, at $\epsilon = 0.01$, no disclosure is found in any data set, but the overall error is awfully large, i.e., about 400% for *EDU* and *OCC*, and about 2200% for *SALARY*. Such excessively noisy query answers are useless for data analysis. For the overall error to be no more than 10%, which we consider necessary for useful utility, ϵ must be relaxed to 0.5 for *EDU* and *OCC* and to 2 for *SALARY*. But in this case, a substantial number of disclosures occur, as shown in Figure 3.3.

3.5 Summary

Both the theoretical analysis in Theorem 2 and the above empirical study on real life data sets suggest that, good utility of published query answers (i.e., a small relative error) under differential privacy comes with disclosures of private information via NIR, which allows inference of private information of an individual through noisy query answers, whereas eliminating such disclosures comes with the cost of poor utility for data analysis. This dilemma is rooted from the fact that both utility and disclosures are based on the same type of information: accurate query answers. When utility is retained, disclosures are permitted, and when disclosures are eliminated, utility is compromised. An implication of this study is that it is important to consider what kind of privacy one wishes to protect. If privacy is about hiding one's participation in the database, differential privacy achieves this goal. If privacy is about concealing one's sensitive information, differential privacy does not do the job unless one is willing to give up the utility for data analysis. Understanding this limitation of differential privacy is important to avoid unexpected disclosures while enjoying the good utility of differential privacy.

Chapter 4

Reconstruction Privacy

The question we study in this chapter is how to (A) allow learning statistical relationships (such as "smokers tend to have lung cancer"), and at the same time, (B) prevent learning sensitive information about an individual (such as "Bob likely has HIV"). As discussed above, syntactic privacy criteria provide (B) but not (A) (Chapter 2), whereas the differential privacy criterion provides (A) but not (B) (Chapter 3). The difficulty of providing both (A) and (B) is that both learning statistical relationships and learning sensitive information make use of NIR, one for utility and one for privacy violation. The key to our approach lies at distinguishing these two types of learning.

We formally define the problem and present our approach in Section 4.1. In particular, we define *reconstruction privacy* that protects individual's sensitive information while providing utility for statistical analysis. The way to test reconstruction privacy is proposed in Section 4.2. How to achieve reconstruction privacy is presented in Section 4.3. Experimental results are shown in Section 4.4.

4.1 Our Approach

In this section we first show our approach through an example (Example 4), and we define our model of data perturbation, privacy criterion, and the problems we will study.

Example 4. *Consider a table $\mathcal{D}(Gender, Job, Disease)$, where Gender and Job are public and Disease is sensitive. Assume that Disease has 10 possible values. To hide the Disease value, for each record in $\mathcal{D}$, uniform perturbation [5] for a given retention probability, say 20%, will retain the Disease value in the record with 20% probability and replace it with a value chosen uniformly from the 10 possible values of Disease at random with the remaining 80% probability. This can be implemented by tossing a biased coin with head probability 20%. Let $\mathcal{D}^*$ denote the perturbed*

data. The operation of perturbation is for protecting individual's sensitive information, because the observed Disease value may not be the real Disease value.

$\mathcal{D}^*$ can be utilized to reconstruct the distribution of Disease in a given subset of records. The operation of reconstruction is for utility. It helps data researchers to find real statistical interesting patterns in terms of distribution of SA. Consider any subset S of $\mathcal{D}$, the counterpart S^* for $\mathcal{D}^*$, and any Disease value d. Let f_d denote the (actual) frequency of d in S, f_d^* denote the (observed) frequency of d in S^*, and $E[F_d^*]$ denote the expectation of f_d^* (over all coin tosses). All frequencies are in fraction. The observed frequency of d come from two sources: (1) when the actual Disease is d, d is retained (i.e., $0.2f_d$), or when d is placed by one Disease value in the whole domain and d is chosen again (i.e., $(0.8/10)f_d$); (2) when the actual Disease is not d and d is chosen to replace the actual Disease (i.e., $(0.8/10)(1 - f_d)$). The following equation follows from the perturbation operation applied to the data:

$$E[F_d^*] = (0.2 + 0.8/10)f_d + (0.8/10)(1 - f_d) \tag{4.1}$$

Approximating the unknown $E[F_d^*]$ by the observed f_d^*, we get an estimate of f_d as $\frac{f_d^* - 0.08}{0.2}$. This estimate is the maximum likelihood estimator (MLE) *[5] computed using the perturbed S^*.*

Given the published $\mathcal{D}^*$, suppose that an adversary tries to learn the likelihood that Bob, a male engineer with a record in $\mathcal{D}$, has breast cancer or BC for short. One way is considering the subset S_1 for all male engineers in $\mathcal{D}$, and another is considering the subset S_2 for all engineers in $\mathcal{D}$. Let M_d^1 and M_d^2 denote the MLE for a disease d in S_1 and S_2, respectively. Two questions can be asked.

Question 1: Which of M_{BC}^1 and M_{BC}^2 should be used to quantify the risk to Bob? *S_1 contains exactly the records that match all Bob's public information, whereas S_2 contains additional records that partially match Bob's public information. Without further information, S_1 is more relevant to Bob than S_2, so M_{BC}^1 should be used to evaluate the risk to Bob. Most likely M_{BC}^2 is not useful for inferring whether Bob has breast cancer, assuming that the additional records for female engineers follow a different distribution on BC from those for male engineers. The case when the additional records have the same distribution as Bob will be discussed in Section 4.1.3. On the other hand, the frequency M_d^2 for some disease d may be useful for data analysis, such as learning the statistical relationship that career engineers tend to have cervical spondylosis.*

Question 2: How to limit the accuracy of M_{BC}^1 while preserving the accuracy of M_d^2 for data analysis? *The errors of M_d^1 and M_d^2 were caused by approximating the unknown $E[F_d^*]$ with the observed f_d^* in Equation (4.1). From the law of large numbers, f_d^* is closer to $E[F_d^*]$ when more*

records are randomized (i.e., more coin toss). Since S_2^* contains more records than S_1^*, M_d^2 is more accurate for estimating the frequency of d in S_2 than M_d^1 for estimating the frequency of d in S_1. We can leverage this gap to limit the accuracy of M_{BC}^1 while preserving the accuracy of M_d^2. □

This example illustrates two types of reconstruction for MLEs. The reconstruction of M_{BC}^1 based on S_1 is called *personal reconstruction* because it aims at a particular individual by matching all public attributes of Bob; the reconstruction of M_d^2 based on S_2 is called *aggregate reconstruction* because it aims at a large population without specifically targeting any individual. We argue (in Section 4.1.2) that personal reconstruction is the source of privacy concerns whereas aggregation reconstruction is the source of utility. The law of large numbers suggests that these two types of reconstruction respond differently to the reduction of record perturbation. We leverage this gap to limit the accuracy of personal reconstruction while preserving the accuracy of aggregate reconstruction. In the rest of the paper, we present an approach to avoid the disclosures of NIR in a data perturbation approach. This effort can be considered as an action on the part of the data publisher.

4.1.1 Data Perturbation

As in [14, 76, 16], we consider a table $\mathcal{D}$ that has one sensitive (private) attribute denoted by SA and several pubic attributes denoted by $NA = \{A_1, \cdots, A_n\}$. We assume that the domain of SA has $m > 2$ sensitive values, $sa_1, \cdots, sa_m$.

Assumptions. To hide the SA information of a record, we perturb the SA value but keep the attributes in NA unchanged in a record. We assume that an adversary has no prior knowledge on positive correlation between NA and SA; otherwise, the public information on NA already discloses the information about SA. The adversary can have prior knowledge on correlation among the attributes in NA, which presents no problem because we never modify the attributes in NA. We also assume that an adversary has no prior knowledge about correlation among SA of *different* records. This assumption can be satisfied by including exactly one record from a set of correlated records, as suggested in [71].

Prior knowledge on negative correlation [58] deserves some more explanations. Consider the negative correlation "females do not have prostate cancer". This correlation tells that the observed prostate cancer is not the original SA value for a female, but does not tell what is the original value because each of the remaining $m - 1$ values has an equal probability. For this reason, we assume that m is larger than 2 (or even larger) so that guessing a remaining value has enough uncertainty. We should emphasize that this situation is not unique for data perturbation, and differentially private answers have similar issues: if the noisy answer for the query on "Female and Prostate Cancer" is -5

(or more generally, too small according to prior knowledge), the above negative correlation would disclose a small range of the noise added, i.e., -5 or less, after observing the noisy answer, which invalids the Laplace distribution assumption. In general, if too much information is leaked through prior knowledge, no mechanism will work.

One criticism on distinguishing SA and NA is that such distinction can be tricky sometimes. This deserves some clarification as well. One approach that does not make such distinction is treating all attributes as sensitive attributes and randomizing a record over the Cartesian product of the domains of all attributes [6][71]. Unfortunately, this approach is vulnerable to undoing the randomization by removing "infeasible" records added during randomization. An example of infeasible records is (Age=1, Job=$prof$, Disease=HIV) since a 1-year child can not possibly be a professor, so the adversary can easily tell that this record was added by randomization. Treating Age and Job as public attributes and randomizing only Disease can avoid this problem. In general, treating and randomizing more attributes like sensitive ones when they are actually public attributes would introduce more vulnerabilities to the removal of "infeasible" records. In this sense, randomizing only the truly sensitive attribute actually provides more protection.

The perturbed version $\mathcal{D}^*$ is produced of $\mathcal{D}$ by applying *uniform perturbation* [6, 5, 32] on SA as follows. For a given retention probability p, where $0 < p < 1$, for each record in $\mathcal{D}$, we toss a coin with head probability p. If the coin lands on head, the SA value in the record is retained; if the coin lands on tail, the SA value in the record is replaced with a value picked from the domain of SA with equal probability (i.e., $\frac{1-p}{m}$) at random. Note that the same value could be picked. Therefore, the probability that one SA value is unchanged is $p + \frac{1-p}{m}$. This perturbation operator is characterized by the following matrix $\mathbb{P}_{m\times m}$:

$$\mathbb{P}_{ji} = \begin{cases} p + \frac{1-p}{m} & \text{if } j{=}i \text{ (retain } sa_i\text{)} \\ \frac{1-p}{m} & \text{if } j{\neq}i \text{ (perturb } sa_i \text{ to } sa_j\text{)} \end{cases} \tag{4.2}$$

A proper choice of the retention probability p can ensure some privacy requirements, such as ρ_1-ρ_2 privacy [32, 6]. We end this section with a comparison between differential privacy approach and perturbation approach in the current work. In the differential privacy approach, a noise is added to the query answer and the noisy answer is used *as is*. For this reason, a *small* and *fixed* noise scale is essential for good utility. As discussed in Section 3.3, as the data size increases, such noises are vulnerable to NIR. In data perturbation, the SA value in each record is perturbed independently and the original distribution of SA must be *reconstructed* from the perturbed records by taking into account the perturbation operation performed. As the data size increases, the number of record perturbation increases *proportionally*, which is less vulnerable to NIR. In addition, data

perturbation is more amendable to record insertion because each record is perturbed independently and the reconstruction is performed by the user himself. In contrast, updating (published) noisy query answers can be tricky because a new record could affect multiple queries and a correlated change of query answers can be exploited by the adversary to learn the information about the new record.

In this thesis the goal is to provide promising utility for statistical relationships learning while protecting individual's sensitive information. Recall that in Section 2.1.2 the operation of perturbation has been introduced with the ability of retaining general statistical information while making individual record fail to reflect its authentic value, thus, it is a good fit for achieving our goal.

4.1.2 Types of Reconstruction

We adopt the following notation. Let $NA = \{A_1, \cdots, A_n\}$. For $1 \leq i \leq n$, let x_i be either a domain value of A_i or a wildcard, denoted by $-$, that matches every domain value of A_i. $\mathcal{D}(x_1, \cdots, x_n)$ denotes the subset of records in $\mathcal{D}$ that match x_i on every A_i, and $\mathcal{D}^*(x_1, \cdots, x_n)$ denotes the corresponding subset for $\mathcal{D}^*$. If, for $1 \leq i \leq n$, x_i is a non-wildcard, $\mathcal{D}(x_1, \cdots, x_n)$ is a *personal group*. If at least one x_i is a wildcard, $\mathcal{D}(x_1, \cdots, x_n)$ is an *aggregate group*. For example, for $NA = \{Gender, Job\}$, $\mathcal{D}(male, eng)$ is a personal group and $\mathcal{D}(-, eng)$ is an aggregate group. Intuitively, a personal group contains all records that can not be distinguished by any information other than SA. For example, even if an adversary may know the age of Bob, this information is not helpful to distinguish any record in the personal group $\mathcal{D}(male, eng)$ because all records in the personal group are exactly identical on NA. Without confusion, we call both $\mathcal{D}(x_1, \cdots, x_n)$ and $\mathcal{D}(x_1, \cdots, x_n)^*$ a personal or aggregate group.

In Example 4, we argued that the personal group $\mathcal{D}^*(male, eng)$ should be used to quantify the risk of inferring the disease breast cancer for the male engineer Bob, instead of the aggregate groups $\mathcal{D}^*(-, eng)$, $\mathcal{D}^*(male, -)$, or $\mathcal{D}^*(-, -)$. The rationale is that unless further information is available, it is to the adversary's advantage not to use a record that is *known* not belonging to Bob. On the other hand, the adversary can not rule out any record in $\mathcal{D}^*(male, eng)$ because they all match Bob's public attributes. In this sense, $\mathcal{D}^*(male, eng)$ is the most relevant subset of records for learning Bob's information on SA. An analogy is short-listing the suspect of a robbery: if the eyewitness has reported that the suspect was a blonde Caucasian male (i.e., the public attributes), it makes sense to focus on the subset of blonde Caucasian males in the police database, instead of examining all Caucasian males records. The above observation motivates the following two types of reconstruction.

Definition 6. *A* personal reconstruction *refers to estimating the frequencies of the SA values in a personal group g based on the perturbed g^*. An* aggregate reconstruction *refers to estimating the frequencies of the SA values in an aggregate group G based on the perturbed G^*.* □

We consider a personal reconstruction as the source of privacy concern because it aims specifically at an individual by matching all the individual's public information. In contrast, we consider an aggregate reconstruction as the source of utility because it aims at a larger population without specifically targeting a particular individual. In the presence of both a person group and an aggregate group that match an individual's public attributes, the person group overrides the aggregate group as far as quantifying the privacy risk is concerned. For example, to learn the Diseases of the male engineer Bob, the personal reconstruction based on $\mathcal{D}^*(male, eng)$ overrides the aggregate reconstruction based on $\mathcal{D}^*(male, -)$, $\mathcal{D}^*(-, eng)$, and $\mathcal{D}^*(-, -)$. These different roles of reconstruction are stated in the next principle.

Definition 7 (Split Role Principle)**.** *A personal reconstruction aims specifically at a particular individual and is responsible for privacy violation. An aggregate reconstruction aims at a larger population and is responsible for providing utility. As far as privacy protection is concerned, it suffices to ensure that personal reconstruction is not accurate.* □

Remarks. The Split Role Principle provides only a relative privacy guarantee because some disclosure can still occur to an individual through aggregate reconstruction in the name of utility, such as "females tend to have breast cancer (compared to males)". But our principle assures the individual that such disclosures are not specifically targeting him or her, and those that do (i.e., personal reconstruction) have been made unreliable. There is one case when an aggregate reconstruction can help adversaries to learn the sensitive information of a particular individual: when the sensitive values in some other personal groups have the same distribution of the personal group containing the target. In this case the SA distribution of a union of several personal groups could reflect the SA distribution of each individual personal group. More details of using aggregate reconstructions to learn sensitive information of a particular individual will be discussed in Section 4.1.4. In fact, any statistical database with any non-trivial utility incurs some amount of disclosure [27]. Our principle assures that only a limited amount of disclosure is incurred by enabling non-trivial utility.

4.1.3 Reconstruction Privacy

Our goal is to limit the accuracy of personal reconstruction and preserve (as much as possible) the accuracy of aggregate reconstruction. At first glance, this may sound like an impossible goal: if the personal reconstruction based on $\mathcal{D}^*(female, j)$ for every job j is made inaccurate, how could

the aggregate reconstruction based on $\mathcal{D}^*(female, -)$, which is the aggregation of $\mathcal{D}^*(female, j)$, be kept accurate? How is this possible when the aggregate group $\mathcal{D}^*(female, -)$ is the union of all personal groups $\mathcal{D}^*(female, j)$ for all jobs j? The general idea of achieving this goal is as follows. The perturbation on SA in each record is a random trial, i.e., a coin toss. $\mathcal{D}^*(female, -)$ has more random trials than each $\mathcal{D}^*(female, j)$ individually, where j is a value in the domain of Job. The law of large numbers implies that, if we reduce the size of $\mathcal{D}^*(female, j)$ by sampling, we expect that the personal reconstruction based on $\mathcal{D}^*(female, j)$ will be affected more adversely than the aggregate reconstruction based on $\mathcal{D}^*(female, -)$. This is analogous to estimating the head probability of a coin: if ten people each toss the coin five times and combine their coin tosses together, the estimation is much more accurate than each person estimating the head probability using their own five coin tosses. This observation motivates a new privacy definition in terms of the accuracy of personal reconstruction.

This observation motivates a sampling based algorithm to limit the accuracy of personal reconstruction while enabling the accuracy of aggregate reconstruction. The above different responses to sampling can be explained by the law of large numbers: the average of the results obtained from a large number of trials is close to the expected value, and will tend to become closer as more trials are performed.

Under the Split Role Principle, our privacy guarantee is that all personal reconstructions are not effective for learning the information about SA. To formalize this guarantee, consider a personal group g^* and g, and a particular SA value sa. Let f denote the frequency of sa in g and let F' denote the estimate of f obtained from the personal reconstruction based on g^* (the way to compute F' will be introduced in Section 4.2.1). Note that F' is a random variable because $\mathcal{D}^*$ is a result of coin tosses. $\frac{F'-f}{f}$ is the relative error of F'. A larger $\frac{F'-f}{f}$ means that an adversary faces more uncertainty in using F' to gauge of the likelihood of sa for an individual. The next definition formalizes an "inaccuracy requirement" on $\frac{F'-f}{f}$.

Definition 8 (Reconstruction Privacy)**.** *Let $\lambda > 0$ and $\delta \in [0, 1]$. sa is* (λ, δ)-reconstruction-private *in a personal group g^* if $\Pr\left[\frac{F'-f}{f} > \lambda\right] < U$ or $\Pr\left[\frac{F'-f}{f} < -\lambda\right] < L$, for some U and L, implies $\delta \leq min\{U, L\}$. A personal group g^* is* (λ, δ)-reconstruction-private *if every sa is* (λ, δ)-reconstruction-private *in g^*. $\mathcal{D}^*$ is* (λ, δ)-reconstruction-private *if every personal group g^* is* (λ, δ)*-reconstruction-private. (All probabilities are taken over the space of coin tosses during the perturbation of SA values.)* □

Corollary 4. *If $\mathcal{D}^*$ is (λ, δ)-reconstruction-private, for any target individual t and for any SA value sa, the adversary can not prove either $\Pr\left[\frac{F'-f}{f} \geq \lambda\right] < \delta$ or $\Pr\left[\frac{F'-f}{f} \leq -\lambda\right] < \delta$, where f is*

the frequency of sa in the personal group g which contains t and F' is the random variable for the estimate of f based on g over all coin tosses.

Note that reconstruction privacy is a property of the perturbation matrix $\mathbb{P}$, not a property of a particular instance of $\mathcal{D}^*$. In plain words, (λ, δ)-reconstruction-privacy ensures that the *smallest upper bound* is not less than δ; in this sense, the adversary has difficulty to get an accurate estimate of f, and the larger λ or δ is, the greater this difficulty is. As an example, violating $(0.3, 0.3)$-reconstruction-privacy by g^* means that the adversary can get a smaller-than-0.3 upper bound on $\Pr\left[\frac{F'-f}{f} > 0.3\right]$ or $\Pr\left[\frac{F'-f}{f} < -0.3\right]$. This implies at least one of the following:

$$\Pr\left[\frac{F'-f}{f} \leq 0.3\right] \geq 70\%, \text{ where } F' > f$$
$$\Pr\left[\frac{F'-f}{f} \geq -0.3\right] \geq 70\%, \text{ where } F' < f$$

Our definition considers such a high probability of a small error as a potential risk.

Remarks. Note that in Chapter 3 we defined a disclosure as in Definition 5. Based on the disclosure definition, either preventing the noisy confidence to be too close to the true confidence, or preventing the true confidence to be much larger than the prior could help thwart the attack. The second approach has been discussed in many previous works and most syntactic methods focus on this area, they try to limit the change of the true confidence from the prior, but this also limits the utility. Because what you can learn from any sub-population is almost what you can learn from the whole data set, and details have been discussed in Section 1.2. Reconstruction privacy takes a different approach, i.e., it bounds the maximum value of F' or f by imposing a large error requirement on estimating f using F'. In this way, the utility will not be limited.

$F' - f$ should not be confused with the change in the posterior belief of an adversary. In fact, f is the probability of sa in the personal group g and F' is the estimate of f based on the personal reconstruction for g^*, and $\frac{F'-f}{f}$ is the relative error of the estimate. Our definition considers a small estimation error as a privacy risk, regardless of the absolute value of f, on the basis that any accurate person reconstruction is potentially a risk because it discloses the actual distribution of SA that aims at a target individual. The choice of the relative error, instead of the absolute error, is necessary because a larger actual frequency f requires a larger absolute error for protection. Bounding the accuracy of estimating f, instead of bounding the posterior belief of an adversary, has two important benefits: it allows the room for learning statistical relationships (through aggregate reconstruction), and it frees the publisher of measuring the adversary's prior belief and specifying a threshold for posterior beliefs, which can be tricky [27][12]. Finally, the choice of smallest upper bounds, rather than lower bounds, on $\Pr\left[\frac{F'-f}{f} > \lambda\right]$ and $\Pr\left[\frac{F'-f}{f} < -\lambda\right]$, allows us to leverage the literature on upper bounds for random variables to estimate $\Pr\left[\frac{F'-f}{f} > \lambda\right]$.

Consequently, it is possible that the actual probabilities are smaller than such upper bounds, but the consideration of *smallest* upper bounds ensures that it is difficult for an adversary to get a tighter bound. As we shall see in Corollary 5, such upper bounds increase exponentially as the size of a personal group decreases, which suggests that sampling effectively cripples personal reconstruction while preserving aggregate reconstruction. The choice of the minimum constraint on the best upper bounds of an adversary allows us to leverage the extensive research on upper bounds of tail probabilities in the literature. We will return to this point in Section 4.2.2.

(λ, δ)-reconstruction-privacy is a constraint on the error for personal reconstruction, but it places no constraint on aggregate reconstruction. We shall leverage this difference to achieve (λ, δ)-reconstruction-privacy while preserving the accuracy of aggregate reconstruction. Perturbation, specified by a retention probability p, provides uncertainty about the SA value in a record (such as ρ_1-ρ_2 privacy). (λ, δ)-reconstruction-privacy further ensures that the distribution of SA within a personal group g for any target individual can not be accurately reconstructed from the randomized data $\mathcal{D}^*$.

Definition 9 (Enforcing Privacy)**.** *Given a database $\mathcal{D}$, a retention probability p ($1 > p > 0$) for perturbing SA, and privacy parameters λ and δ, devise an algorithm that enforces (λ, δ)-reconstruction-privacy on $\mathcal{D}^*$ while preserving aggregate reconstruction as much as possible.* □

Indeed, our privacy criterion depends on the uncertainty of SA in a record in $\mathcal{D}^*$ introduced by a retention probability $p > 0$. In this sense, reconstruction privacy can be considered as an *additional* protection on top of other privacy criteria, such as ρ_1-ρ_2 privacy. The retention probability p is needed to satisfy traditional privacy definitions. On the contrary, the value of p also affects the utility of the data for statistical analysis. Intuitively, the more original data values are retained, the better the data set utility is. For example, [32, 6] give the maximum p for ensuring ρ_1-ρ_2 privacy.

4.1.4 Generalized Personal Groups

As mentioned in Section 4.1.2, there is one case when an aggregate reconstructions can be used to learn the sensitive information of an individual: if several personal groups have the same SA distribution, then the SA distribution in the union of these personal groups learned by an adversary reflects the SA distribution in each individual personal group.

Consider two personal groups $g^* = \mathcal{D}(male, eng)$ and $g'^* = \mathcal{D}(female, eng)$. Our reconstruction privacy limits the accuracy of reconstruction for each personal group, but does not limit the accuracy of reconstruction for the combined $g^* \cup g'^*$, i.e., the aggregate group $\mathcal{D}^*(-, eng)$, because the reconstruction for $g^* \cup g'^*$ is not relevant to an individual, assuming that males and

females have a different distribution on SA, such as on breast cancer. However, this argument may be invalid if the adversary has further knowledge about the distribution of NA values on SA values. For example, suppose that $FavoriteColor$ is another public attribute and that the value of $FavoriteColor$ of an individual has nothing to do with the diseases, the adversary may do reconstruction after aggregating all personal groups that differ only in the values on $FavoriteColor$, and such reconstruction is more accurate than the reconstruction based on a single personal group because it uses more randomized records. In this case, aggregate groups disclose sensitive information.

To address this issue, for each public attribute A_i, if a set of A_i's domain values have the same impact on SA, we will merge all values in the set into a single generalized value, and we define personal groups based on such generalized values. With this preprocessing, every generalized value of A_i now has a different impact on SA, thus, has a different distribution on SA. Then our previous argument that an aggregate group does not provide a representative statistics for a target individual remains valid, because such groups combining several sub-populations follow different distributions on SA.

For example, in Example 4 it does not help adversaries for considering the subset $S = \mathcal{D}(-, eng)$ for all genders of engineers to infer whether Bob gets breast cancer. This is because gender plays a significant role in estimating whether an individual gets breast cancer (females are more likely to get breast cancer). On the contrary, if the data set contains another NA, $FavoriteColor$ (the value of $FavoriteColor$ has nothing to do with the disease), then adversaries may do reconstruction after aggregating all values in $FavoriteColor$ for a better reconstruction accuracy. After aggregating NA values, reconstruction privacy limits reconstruction accuracy within new generalized personal groups, for the reconstruction on the original personal groups, the accuracy would even worsen because it involves less independent random trials. In the rest of the paper, personal groups refer to personal groups after aggregating NA values based on their impacts on SA.

Now the question is how to identify the values of A_i that have the same impact on SA. We first show how to tell whether two domain values x_i and x_i' of A_i have the same impact on SA. Then we show how to divide all values of A_i to various groups where all values in one group have the same impact on SA and any two values from different groups have different impacts on SA.

The well studied χ^2-*test* [68] tells if two data sets are from different distributions. For two domain values x_i and x_i' of A_i, let o_{ij} (resp. o_{ij}') be the number of records in $\mathcal{D}$ satisfying $A_i = x_i$ (resp. $A_i = x_i'$) and $SA = sa_j$, $1 \leq j \leq m$, where m is the total number of SA values. Let $O_i = [o_{i1}, \cdots, o_{im}]$ and $O_i' = [o_{i1}', \cdots, o_{im}']$, that represent the distributions of SA conditioned on x_i and x_i'. In proper statistical language, can we disprove, to a certain required level of significance,

the *null hypothesis* that the two data sets O_i and O'_i are drawn from the same population distribution function? Disproving the null hypothesis in effect implies that the data sets are from different distributions. Failing to disprove the null hypothesis only shows that the data sets can be consistent with a single distribution function. In this paper we assume that NA values x_i and x'_i have the same impact on SA if O_i and O'_i come from the same distribution.

Since $|O_i| = \sum_{j=1}^{m} o_{ij}$ and $|O'_i| = \sum_{j=1}^{m} o'_{ij}$ are not necessarily equal, our case is that of two binned distributions with unequal number of data points. In this case, the degree of freedom is equal to m and the χ^2 value is computed as in [69]:

$$\chi^2 = \sum_{j=1}^{m} \frac{\left(\sqrt{|O'_i|/|O_i|}o_{ij} - \sqrt{|O_i|/|O'_i|}o'_{ij}\right)^2}{o_{ij} + o'_{ij}} \tag{4.3}$$

The value of χ^2 tells whether O_i and O'_i are from the same distribution through evaluating the sum of the difference of corresponding pairs (e.g., o_{ij} and o'_{ij}) in O_i and O'_i. The more difference that O_i is from O'_i, the larger the value of χ^2 is, because the sum of the difference of all corresponding pairs tend to be larger. The parts of $\sqrt{|O'_i|/|O_i|}$ and $\sqrt{|O_i|/|O'_i|}$ are for cancelling the impact that the sums of O_i and O'_i are different.

Then we obtain the expected value of χ^2 by checking the chi-square distribution with two parameters, the degree of freedom (e.g., m) and the value of *significance*, the maximum probability that the computed χ^2 from Equation (4.3) could be greater than the expected χ^2. We set the conventional setting of 0.05 for significance. If the value computed by Equation (4.3) is greater than this expected value of χ^2, we can disprove the null hypothesis that the two data sets O_i and O'_i are drawn from the same population distribution function because the probability for this is less than 5% (i.e., the significance). Otherwise, we consider that the two data sets are consistent with a single distribution function.

We already know how to find out whether two domain values x_i and x'_i have the same impact on SA. The next step is to merge all values of A_i that have the same impact. To achieve this, we represent the χ^2-test results for all pairs (x_i, x'_i) of values of A_i using a graph. Each value x_i of A_i is a vertex in the graph and we connect two vertices x_i and x'_i if the χ^2-test on (x_i, x'_i) fails to disprove the null hypothesis that the two data sets O_i and O'_i are drawn from the same population distribution function. Finally, for each connected component of the graph, we merge all the values in the component into a single generalized value. The problem of finding all groups of NA values with the same impact can be expressed as the problem of finding *connected components* from a graph, which is a well-known problem with solutions provided in [44]. This method ensures that

any two values x_i and x'_i from different components have a different impact on SA in that O_i and O'_i are drawn from different population distribution functions.

In the rest of the paper, we assume that the domain values of each public attribute A_i are generalized values produced by the above merging procedure and that the personal and aggregate groups defined in Section 4.1.2 are based on such generalized domain values.

4.2 Testing Privacy

An immediate question is how to test (λ, δ)-reconstruction-privacy. From Definition 8, this requires to obtain the smallest upper bounds U and L on $\Pr\left[\frac{F'-f}{f} > \lambda\right]$ and $\Pr\left[\frac{F'-f}{f} < -\lambda\right]$. The following discussion refers to a subset S of $\mathcal{D}$ and the corresponding subset S^* of $\mathcal{D}^*$. $|S|$ denotes the number of records in S. Let $(f_1, \cdots, f_m)$ be the frequencies of SA values $(sa_1, \cdots, sa_m)$ in S, $(O_1^*, \cdots, O_m^*)$ be the variables for the observed counts of $(sa_1, \cdots, sa_m)$ in S^*, and $(F'_1, \cdots, F'_m)$ be the variables for an estimate of $(f_1, \cdots, f_m)$ reconstructed using S^*. These vectors are also written as column-vectors $\overleftarrow{f}$, $\overleftarrow{O^*}$, and $\overleftarrow{F'}$. When no confusion arises, we drop the subscripts i from f_i, O_i^*, F'_i. Table 4.1 summarizes the notations used in this chapter.

Table 4.1: Notations in Chapter 4

Notations	**Explanation**
$\mathcal{D}, \mathcal{D}^*$	the raw data and perturbed version
S, S^*	a subset of records and perturbed version
g, g^*	a personal group and perturbed version
m	the domain size $\|SA\|$
t	a target individual
sa_i	a domain value of SA
f_i	the frequency of sa_i in S
o_i^*	the count of sa_i in S^*
O_i^*	the variable for o_i^*
F'_i	the variable for the estimate of f_i
$\overleftarrow{f}$, $\overleftarrow{F'}$, $\overleftarrow{O^*}$	the column-vectors of f_i, F'_i, O_i^*
$\mathbb{P}$	the perturbation matrix in Equation (4.2)
p	the retention probability
(λ, δ)	privacy parameters

4.2.1 Computing F'

First of all, let us examine the computation of F'. Example 4 illustrates the basic idea of computing the estimate F' of f for a particular SA value sa based on the perturbed data. Generalizing that

idea to the vectors $\overleftarrow{F'}$ and $\overleftarrow{f}$, our perturbation operation implies the equation $\mathbb{P} \cdot \overleftarrow{f} = \frac{E[\overleftarrow{O^*}]}{|S|}$, where $\mathbb{P}$ is the perturbation matrix in Equation (4.2). Approximating $E[\overleftarrow{O^*}]$ by the observed counts $\overleftarrow{O^*}$, we get the estimate of $\overleftarrow{f}$ given by $\mathbb{P}^{-1} \cdot \frac{\overleftarrow{O^*}}{|S|}$, where $\mathbb{P}^{-1}$ is the inverse of $\mathbb{P}$. This estimate is called the *maximum likelihood estimator* (MLE).

Theorem 3 (Theorem 2, [5])**.** *$\mathbb{P}^{-1} \cdot \frac{\overleftarrow{O^*}}{|S|}$ is the MLE of $\overleftarrow{f}$ under the constraint that its elements sum to 1. Let $\overleftarrow{F'}$ denote this MLE.* □

The next lemma gives an equivalent computation of $\overleftarrow{F'}$ without referring to $\mathbb{P}^{-1}$.

Lemma 3. *For any subset S of $\mathcal{D}$ and any SA value sa, (i) $E[O^*] = |S|(fp + (1-p)/m)$, (ii) $F' = \frac{O^*/|S| - (1-p)/m}{p}$, and (iii) $E[F'] = f$.*

Proof. (i) O^* comes from two sources of records in S: those that have the SA value sa and are retained, and those that have a SA value other than sa and are perturbed to sa. The expected number of the records in the first source is $|S|f(p+(1-p)/m)$, and the expected number of the records in the second source is $|S|(1-f)((1-p)/m)$. Summing up the two gives $E[O^*] = |S|(fp+(1-p)/m)$. This shows (i).

(ii) From Theorem 3, $\overleftarrow{F'} = \mathbb{P}^{-1} \cdot \frac{\overleftarrow{O^*}}{|S|}$. Let $\frac{\overleftarrow{1-p}}{m}$ denote the column-vector of the constant $\frac{1-p}{m}$ of length m. We have

$$\frac{\overleftarrow{O^*}}{|S|} = \mathbb{P} \cdot \overleftarrow{F'} = p\overleftarrow{F'} + \frac{\overleftarrow{1-p}}{m}$$

Thus, $F' = \frac{O^*/|S| - (1-p)/m}{p}$, as required for (ii).

(iii) Taking the mean on both sides of the last equation, $E[F'] = \frac{E[O^*]/|S| - (1-p)/m}{p}$. Substituting $E[O^*]$ in (i) and simplifying, we get $E[F'] = f$. This shows (iii). □

Lemma 3(iii) implies that F' is an unbiased estimator of f. Lemma 3(ii) gives a computation of F' in terms of the known values O^*, $|S|$, p, m without referring to $\mathbb{P}^{-1}$. In the rest of the paper, we adopt this computation of F' in the definition of reconstruction privacy (Definition 8).

4.2.2 Bounding $\Pr\left[\frac{F'-f}{f} > \lambda\right]$ and $\Pr\left[\frac{F'-f}{f} < -\lambda\right]$

Recall that $F' = \frac{O^*/|S| - (1-p)/m}{p}$ from Lemma 3(ii). To bound $\Pr\left[\frac{F'-f}{f} > \lambda\right]$ and $\Pr\left[\frac{F'-f}{f} < -\lambda\right]$, we first obtain the upper bounds for the error of *observed* O^* and then convert them into the upper bounds for the error of *reconstructed* F'. The next theorem gives the conversion between these bounds.

Theorem 4 (Bound Conversion)**.** *Consider any subset S of $\mathcal{D}$ and any SA value sa with the frequency f in S. Let O^* be the observed count of sa in S^* and let F' be the MLE of f. Let $\mu = E[O^*]$.*

For any functions $U(\omega,\mu)$ and $L(\omega,\mu)$ of ω and μ, and for a comparison operator $\bigoplus$ that is either $<$ or $>$,

1. $\Pr\left[\frac{O^*-\mu}{\mu} > \omega\right] \bigoplus U(\omega,\mu)$ *if and only if* $\Pr\left[\frac{F'-f}{f} > \lambda\right] \bigoplus U(\omega,\mu)$;
2. $\Pr\left[\frac{O^*-\mu}{\mu} < -\omega\right] \bigoplus L(\omega,\mu)$ *if and only if* $\Pr\left[\frac{F'-f}{f} < -\lambda\right] \bigoplus L(\omega,\mu)$.

where $\lambda = \frac{\omega\mu}{|S|pf}$.

Proof. We show 1) only because the proof for 2) is similar. From $F' = \frac{O^*/|S| - (1-p)/m}{p}$ (Lemma 3(ii)), $O^* = |S|(F'p + (1-p)/m)$, and from Lemma 3(i), $\mu = |S|(fp + (1-p)/m)$. So $\frac{O^*-\mu}{\mu} > \omega$ $\Leftrightarrow O^* - \mu > \omega\mu \Leftrightarrow |S|p(F'-f) > \omega\mu \Leftrightarrow \frac{F'-f}{f} > \frac{\omega\mu}{|S|pf}$. 1) follows by letting $\lambda = \frac{\omega\mu}{|S|pf}$. □

According to Theorem 4, if we have the smallest upper bounds on $\Pr\left[\frac{O^*-\mu}{\mu} > \omega\right]$ or $\Pr\left[\frac{O^*-\mu}{\mu} < -\omega\right]$, we immediately have the smallest upper bounds on $\Pr\left[\frac{F'-f}{f} > \lambda\right]$ or $\Pr\left[\frac{F'-f}{f} < -\lambda\right]$. This conversion does not hinge on the particular form of the bound functions U and L, and applies to both upper bounds (when $\bigoplus$ is $<$) and lower bounds (when $\bigoplus$ is $>$). Therefore, finding the smallest upper bounds for F' is reduced to that for O^*. The latter can benefit from the literature on upper bounds for tail probabilities of Poisson trials. In particular, the coin toss for each record is an independent Poisson trial and several upper bounds for Poisson trials are known. Markov's inequality and Chebyshev's inequality are some early upper bounds, for example. The Chernoff bound, due to [18], is a much tighter bound as it gives exponential fall-off of probability with distance from the error. The following is a simplified yet tight form of the Chernoff bound.

Theorem 5 (Chernoff Bounds, [18])**.** *Let $X_1, \cdots, X_n$ be independent Poisson trials such that for $1 \le i \le n$, $X_i \in \{0,1\}$, $\Pr[X_i = 1] = p_i$, where $0 < p_i < 1$. Let $X = X_1 + \cdots + X_n$ and $\mu = E[X] = E[X_1] + \cdots + E[X_n]$. For $\omega \in (0,\infty)$,*

$$\Pr\left[\frac{X-\mu}{\mu} > \omega\right] < U(\omega,\mu) = exp(-\frac{\omega^2\mu}{2+\omega}) \tag{4.4}$$

and for $\omega \in (0,1]$,

$$\Pr\left[\frac{X-\mu}{\mu} < -\omega\right] < L(\omega,\mu) = exp(-\frac{\omega^2\mu}{2}).\Box \tag{4.5}$$

The observed count O^* of sa in S^* is equal to $X = X_1 + \cdots + X_n$, where X_i is the indicator variable whether the i-th row in S^* has the value sa. If the i-th row has sa prior to perturbation, $p_i = p + (1-p)/m$, otherwise, $p_i = (1-p)/m$. $E[O^*] = |S|(fp + (1-p)/m)$ (Lemma 3). To

obtain the upper bounds for F', we instantiate the upper bounds U and L for O^* in Equations (4.4) and (4.5) into Theorem 4. This gives the next corollary.

Corollary 5 (Upper Bounds for F')**.** *Let* $\omega = \frac{\lambda|S|pf}{\mu}$ *and* $\mu = |S|(fp+(1-p)/m)$*. For* $\omega \in (0, \infty)$*,*

$$\Pr\left[\frac{F'-f}{f} > \lambda\right] < U(\omega, \mu) = exp(-\frac{\omega^2\mu}{2+\omega}) \tag{4.6}$$

and for $\omega \in (0, 1]$*,*

$$\Pr\left[\frac{F'-f}{f} < -\lambda\right] < L(\omega, \mu) = exp(-\frac{\omega^2\mu}{2}).\square \tag{4.7}$$

Note that $\omega = \frac{\lambda pf}{pf+(1-p)/m}$ and $\mu = |S|(fp + (1-p)/m)$. λ, p, f, m are constants. Reducing $|S|$ decreases μ, which increases the upper bounds U and L *exponentially*. Thus, reducing $|S|$ effectively thwarts the attacker from bounding $\Pr\left[\frac{F'-f}{f} > \lambda\right]$ and $\Pr\left[\frac{F'-f}{f} < -\lambda\right]$ by a small upper bound. Our enforcement algorithm presented in the next section is based on this observation.

A remaining question is whether $U = exp(-\frac{\omega^2\mu}{2+\omega})$ and $L = exp(-\frac{\omega^2\mu}{2})$ in Corollary 5 derived from the Chernoff bound for O^* are the smallest upper bounds for F', as required by the definition of (λ, δ)-reconstruction-privacy. Suppose U and L are not the smallest upper bounds. There would exist a smaller upper bound U_2 on $\Pr\left[\frac{F'-f}{f} > \lambda\right]$ or a smaller upper bound L_2 on $\Pr\left[\frac{F'-f}{f} < -\lambda\right]$. Then Theorem 4 implies that U_2 and L_2 are better bounds than the Chernoff bounds U and L for O^*. However, the fact that the Chernoff bound remained in use in the past 60 years suggests that finding smaller upper bounds is difficult. Until the Chernoff bound is improved, we assume that the upper bounds U and L in Corollary 5 are the best upper bounds for F' of an adversary. This assumption is not a real restriction because Theorem 4 allows us to "plug in" any better bound for O^* for a better bound for F'. When the adversary finds a better bound than the Chernoff bound and the data publisher still uses the Chernoff bound, if the better bound is a general result and the publisher refuses to "plug in" it, the responsibility is with the publisher. Otherwise, under our assumptions about prior knowledge in Section 4.1.1, getting a better bound requires knowledge about the random coin tosses in the perturbation process. Like all randomized mechanisms, we assume that actual results of random trails are not available to the adversary.

4.2.3 Testing

With the upper bounds L and U in Corollary 5, it is straightforward to test whether (λ, δ)-reconstruction-privacy holds by testing $\delta \leq min\{L, U\}$. We can further simplify this test. For ω in the range $(0, 1]$, it is easy to see $L < U$, therefore, $\delta \leq min\{L, U\}$ degenerates into $\delta \leq L$. Substituting the expres-

sions for ω and μ in Corollary 5 into $L(\omega, \mu)$, we get $L = exp(-\frac{(\lambda pf)^2|S|}{2(fp+(1-p)/m)})$, where λ is in the range $(0, 1 + \frac{(1-p)/m}{pf}]$, which corresponds to the range $(0, 1]$ for ω. Substituting the expression for L into $\delta \leq L$ gives rise to the following test of (λ, δ)-reconstruction-privacy.

Corollary 6. *Let sa be a SA value, g be a personal group, and f be the frequency of sa in g. For $\lambda \in (0, 1 + \frac{(1-p)/m}{pf}]$ and $\delta \in [0, 1]$, sa is (λ, δ)-reconstruction-private in g^* if and only if*

$$|g| \leq \frac{-2(fp + (1-p)/m) \ln \delta}{(\lambda pf)^2} \square \tag{4.8}$$

Given the data set $\mathcal{D}$, the personal groups g and the frequencies f for all SA values in g can be found by sorting the records in $\mathcal{D}$ in the order of all attributes in NA followed by SA. Therefore, all the quantities in Equation (4.8) are either given (i.e., λ, δ, p, m) or can be computed efficiently (i.e., f and $|g|$). A larger $|g|, f, p$ makes this inequality less likely hold, thus, makes (λ, δ)-reconstruction-privacy more likely violated. In fact, under these conditions there are either more random trials or more retention of the SA value, which leads to a more accurate reconstruction.

4.3 Enforcing Privacy

If reconstruction privacy is not satisfied, we can restore reconstruction privacy by satisfying the condition in Equation (4.8) for every SA value and every personal group. Observe that the right-hand side of Equation (4.8) decreases as f increases. Therefore, a personal group g^* satisfies reconstruction privacy if and only if $|g| \leq s_g$, where

$$s_g = \frac{-2(fp + (1-p)/m) \ln \delta}{(\lambda pf)^2} \tag{4.9}$$

and f is the maximum frequency for any SA value in g. Another interpretation is that s_g is the maximum number of independent trials if g^* is to satisfy reconstruction privacy. If $|g| > s_g$, reconstruction privacy is violated (because of too many independent trails). To fix this, one approach is increasing s_g to the current group size $|g|$ by reducing f or p (note that m, λ, δ are fixed). This approach is not preferred because reducing f will distort the data distribution and reducing p has a global effect of making the perturbed data too noisy. Our approach is reducing $|g|$ to the size s_g by *sampling* a subset g_1 of the size s_g and *perturbing* g_1 instead of g. This sampling essentially reduces the excessive number of independent random trials. To ensure $s_{g_1} = s_g$, g_1 must preserve the (relative) frequency of every SA value in g (to the right-hand side of Equation (4.9) unchanged after sampling). Preserving frequencies also helps minimize the distortion to data distribution. After per-

Table 4.2: (a) The personal group g before SPS. (b) $Sampling(g, s_g)$ produces a sample g_1 of g with $\tau = s_g/|g| = 0.75$. (c) $Perturbing(g_1, p, m)$ produces the randomized version of g_1, g_1^*, with $p = 0.8$. (d) $Scaling(g_1^*, |g|)$ generates g_2^* through scaling up g_1^* to the size $|g|$ with $\tau' = |g|/|g_1^*| = 20/15 = 1.33$.

(a) g

NA	SA
$\cdots$	sa_1
$\cdots$	sa_1
$\cdots$	sa_1
$\cdots$	sa_1
$\cdots$	sa_1
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2

(b) g_1

NA	SA
$\cdots$	sa_1
$\cdots$	sa_1
$\cdots$	sa_1
$\cdots$	sa_1
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2

(c) g_1^*

NA	SA
$\cdots$	sa_1
$\cdots$	sa_2
$\cdots$	sa_1
$\cdots$	sa_1
$\cdots$	sa_2
$\cdots$	sa_1
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_1
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_1

(d) g_2^*

NA	SA
$\cdots$	sa_1
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_1
$\cdots$	sa_1
$\cdots$	sa_1
$\cdots$	sa_2
$\cdots$	sa_1
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_1
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_2
$\cdots$	sa_1
$\cdots$	sa_1

turbing the sample g_1, a *scaling* step is needed to scale the perturbed g_1^* back to the original size $|g|$ to minimize the impact on the global distribution. Below, we present an algorithm named *Sampling-Perturbing-Scaling (SPS)* to meet both the group size requirement and the frequency preservation requirement.

The algorithm based on the above idea is described in Algorithm 1. The input is a database $\mathcal{D}$, the retention probability p ($0 < p < 1$), the domain size m of SA, and the privacy parameters λ and δ. The output is a modified version of $\mathcal{D}^*$ that satisfies (λ, δ)-reconstruction-privacy. For each personal group g, if $|g| \leq s_g$, g_2^* is equal to g^*. Otherwise, g_2^* is produced by the three steps on Lines 7-9 described above. $\mathcal{D}_2^*$ contains all g_2^*.

Example 5 illustrates how the SPS algorithm is performed on one personal group g.

Example 5. *Suppose that a personal group g contains 5 records for one SA value sa_1 and 15 records for another SA value sa_2. $|g| = 20$, $|g_{sa_1}| = 5$, $|g_{sa_2}| = 15$. Assume $s_g = 15$. Table 4.2 illustrates how each step of SPS is operated on g.* □

Remarks. Several points are worth noting. First, *Sampling* kicks in only if $|g|$ exceeds the maximum size s_g; otherwise, all records in g will be used for perturbation. Therefore, if the data set is

Algorithm 1 Sampling-Perturbing-Scaling (SPS)

Input: the data set $\mathcal{D}$, retention probability p, the number of SA values m,the privacy parameters λ, δ
Output: Randomized $\mathcal{D}_2^*$ that is (λ, δ)-reconstruction-private

$\mathcal{D}_2^* \leftarrow$ ‰
Sort $\mathcal{D}$ to get all personal groups g
for all personal groups g in $\mathcal{D}$ **do**
    compute s_g using Equation (4.9)
    if $|g| \leq s_g$ **then**
        $g_2^* \leftarrow Perturbing(g, p, m)$
    else
        $g_1 \leftarrow Sampling(g, s_g)$
        $g_1^* \leftarrow Perturbing(g_1, p, m)$
        $g_2^* \leftarrow Scaling(g_1^*, |g|)$
    end if
    add g_2^* to $\mathcal{D}_2^*$
end for
return $\mathcal{D}_2^*$

$Sampling(g, s_g)$:

$temp \leftarrow$ ‰
$\tau \leftarrow s_g/|g|$
for all SA value x occurring in g **do**
    $g_x \leftarrow$ the set of records in g having x
    add to $temp$ any $\lfloor |g_x|\tau \rfloor$ records from g_x
    add to $temp$ one additional record from g_x with probability $|g_x|\tau - \lfloor |g_x|\tau \rfloor$
end for
return $temp$

$Perturbing(g_1, p, m)$:

$temp \leftarrow$ ‰
for all record r in g_1 **do**
    let r^* be r with SA perturbed with retention probability p
    add r^* to $temp$
end for
return $temp$

$Scaling(g_1^*, |g|)$:

$\tau' \leftarrow |g|/|g_1^*|$
$temp \leftarrow$ ‰
for all record r^* in g_1^* **do**
    add to $temp$ $\lfloor \tau' \rfloor$ duplicates of r^*
    add to $temp$ one additional duplicate of r^* with probability $\tau' - \lfloor \tau' \rfloor$
end for
return $temp$

small enough to have such a poor accuracy that already satisfies reconstruction privacy, our algorithm will behave like the standard uniform perturbation without performing sampling. In this case, the poor accuracy is not caused by our sampling, but by the inadequate amount of data. Second, the duplication in *Scaling* does not introduce new random trials because it is performed *after* the

perturbation in g_1^*. The adversary may notice some duplicate records in g_2^*, but this is not a problem because privacy is actually achieved on g_1^* before the scaling step. Third, we compute the maximum number of records each personal group could hold without violation reconstruction privacy, therefore, in the step of bringing noise (i.e., sampling) we only add the minimum amount of noise for achieving reconstruction privacy. In other words, our method also provides the best utility under reconstruction privacy.

Complexity analysis. Let $|\mathcal{D}|$ denote the number of records in $\mathcal{D}$. The sorting step takes $|\mathcal{D}|log|\mathcal{D}|$ time to generate all personal groups. Subsequently, each of the steps $Sampling$, $Perturbing$, and $Scaling$ takes one data scan. A more efficient implementation, however, is to perform these three steps in a single data scan: as a record r is sampled, immediately we perturb the SA value of r and then duplicate the perturbed record a certain number of times as described, and add the duplicates to g_2^*. In total, the algorithm takes $(|\mathcal{D}|log|\mathcal{D}|)$ time.

4.3.1 Analysis

We prove two claims about the output $\mathcal{D}_2^* = \cup g_2^*$. The first claim is on privacy guarantee: each g_2^* in $\mathcal{D}_2^*$ is (λ, δ)-reconstruction-private. The second claim is on utility: for any subset S consisting of one or more personal groups and the corresponding subset S_2^* in $\mathcal{D}_2^*$, F'_{g_2} is an unbiased estimator of f, where f is the frequency of a particular SA value in S and F'_{g_2} is the estimate of f reconstructed from S_2^*, respectively. We first present some facts.

Let g be a personal group. Assume $|g| > s_g$. Let g_1, g_1^*, g_2^* be computed for g and let $O_g^*, O_{g_1}^*, O_{g_2}^*$ be the observed count for a particular SA value sa in g^*, g_1^*, g_2^*, respectively. Let f_g and f_{g_1} be the frequency of sa in g and g_1. Let F'_g, F'_{g_1}, F'_{g_2} be the MLE reconstructed from g^*, g_1^*, g_2^*. We avoid to use f_1, F'_1, F'_2 as these symbols have been used as the frequencies for SA values sa_1 and sa_2. Let $u \simeq v$ denote that u and v are equal modulo the random trial for the additional record in *Scaling* and *Sampling*.

- Fact 1: $f_{g_1} \simeq f_g$ and $|g_1| \simeq s_g$. This is because *Sampling* preserves the frequency of sa in g and the sample g_1 has the size s_g.
- Fact 2: $O_{g_2}^*/|g_2^*| \simeq O_{g_1}^*/|g_1^*|$. This is because *Scaling* from g_1^* to g_2^* preserves the frequency of sa.
- Fact 3: $F'_{g_1} \simeq F'_{g_2}$. This follows from $F'_{g_i} = \frac{O_{g_i}^*/|g_i^*| - (1-p)/m}{p}$, $i = 1, 2$ (Lemma 3(ii)) and Fact 2.

- Fact 4: $E[O^*_{g_2}] \simeq E[O^*_g]$. From Lemma 3(i), $E[O^*_{g_1}] = |g_1|(f_{g_1}p + (1-p)/m) \simeq s_g(f_{g_1}p + (1-p)/m)$ (Fact 1). Since *Scaling* duplicates each record in g^*_1 by $\frac{|g|}{s_g}$ times, $E[O^*_{g_2}] \simeq \frac{|g|}{s_g} \times E[O^*_{g_1}] = |g|(f_{g_1}p + (1-p)/m)$. From Lemma 3(i), $E[O^*_g] = |g|(f_g p + (1-p)/m)$. Then $f_{g_1} \simeq f_g$ (Fact 1) implies $E[O^*_{g_2}] \simeq E[O^*_g]$.

Theorem 6 (Privacy)**.** *For each personal group g, g^*_2 returned by the SPS algorithm is (λ, δ)-reconstruction-private.*

Proof. If $|g| \leq s_g$, $g^*_2 = g^*$, by Corollary 6, g^*_2 is (λ, δ)-reconstruction-private. We assume $|g| > s_g$. In this case, g^*_1 is (λ, δ)-reconstruction-private because $|g_1| \simeq s_g$ (Fact 1). We claim $\frac{F'_{g_2} - f_g}{f_g} \simeq \frac{F'_{g_1} - f_{g_1}}{f_{g_1}}$, which implies that F'_{g_2} has the same tail probability for error as F'_{g_1}; therefore, g^*_2 is (λ, δ)-reconstruction-private because g^*_1 is. This claim follows from $f_{g_1} \simeq f_g$ (Fact 1) and $F'_{g_1} \simeq F'_{g_2}$ (Fact 3). □

Theorem 7 (Utility)**.** *Let S be a set of records for one or more personal groups in $\mathcal{D}$, S^* be the corresponding set for $\mathcal{D}^*$, and S^*_2 be the corresponding set for $\mathcal{D}^*_2$. Let f be the frequency of a SA value sa in S, and let F' and F'_{S_2} be the estimates of f reconstructed from S^* and S^*_2. Then $E[F'_{S_2}] \simeq f$.*

Proof. Let $O^*_2 = \sum O^*_{g_2}$, $O^* = \sum O^*_g$, $|S^*| = \sum |g^*|$, and $|S^*_2| = \sum |g^*_2|$, where $\sum$ is over the personal groups g for S. $|S^*| \simeq |S^*_2|$. From Lemma 3(ii), $E[F'] = \frac{E[O^*]/|S^*| - (1-p)/m}{p}$ and $E[F'_{S_2}] = \frac{E[O^*_2]/|S^*_2| - (1-p)/m}{p}$. From Fact 4, $E[O^*] \simeq E[O^*_2]$. Thus, $E[F'] \simeq E[F'_{S_2}]$. From Lemma 3(iii), $E[F'] \simeq f$, thus, $E[F'_{S_2}] \simeq f$. □

Intuitively, Theorem 7 says that the estimate reconstructed using the corresponding records in $\mathcal{D}^*_2$ is an unbiased estimator of the actual frequency.

4.4 Experimental Studies

Two claims are evaluated in this section. The first claim is that reconstruction privacy could be violated on real life data sets. The second claim is that the proposed SPS algorithm eliminates personal reconstruction with minor sacrifice on the utility of aggregate reconstruction.

4.4.1 Experimental Setup

We implemented the proposed SPS algorithm as described in Section 4.3 in C++ and ran all experiments on an Intel Xeon(R) E5630 CPU 2.53GHZ PC with 12GB of RAM. We utilized two publicly

available data sets. The first one is the $ADULT$ data set [1]. This data set has 45,222 records (without missing values) extracted from the 1994 Census database with the attributes Education, Occupation, Race, Gender, and Income. We chose Income as SA and the remaining attributes as the public attributes NA. The second data set is the OCC data previously used in [79][16]. This data set contains personal information about 500K American adults with 6 discrete attributes Age, Gender, Education, Marital, Race, and Occupation. We chose Occupation as SA and the remaining attributes as NA. We considered five samples of OCC of sizes $100K$, $200K$, $300K$, $400K$, $500K$. These data sets have different characteristics: $ADULT$ represents a small data set with very few SA values (with Income having only two values), whereas OCC represents a large data set with a large number of balanced distributed SA values (with Occupation having 50 values). We want to see how these characteristics would affect the evaluation of our claims.

We consider the generalized personal groups by applying the aggregation in Section 4.1.4 to NA. Tables 4.3 and 4.4 show the impacts on the domain size of each NA, the total number of personal groups (i.e., $|\mathcal{G}|$), and the averaged personal groups size (e.g., $|\mathcal{D}|/|\mathcal{G}|$ with $|\mathcal{D}|$ as the total number of records) of $ADULT$ and OCC $300K$. In the rest of this section, we use the generalized values of public attributes.

Table 4.3: NA Aggregation Impact on $ADULT$

	Domain Size of NA				$\|\mathcal{G}\|$	$\|\mathcal{D}\|/\|\mathcal{G}\|$
	Education	Occupation	Race	Gender		
Before Aggregation	16	14	5	2	2240	20
After Aggregation	7	4	2	2	112	404

Table 4.4: NA Aggregation Impact on OCC $300K$

	Domain Size of NA					$\|\mathcal{G}\|$	$\|\mathcal{D}\|/\|\mathcal{G}\|$
	Age	Gender	Education	Marital	Race		
Before Aggregation	77	2	14	6	9	116424	3
After Aggregation	1	2	14	6	9	1512	331

The utility of the published data is evaluated by the accuracy of answering count queries of the form:

$$SELECT\ COUNT\ (*)\ FROM\ \mathcal{D}\ WHERE\ A_1 = a_1 \wedge \cdots \wedge A_d = a_d \wedge SA = sa_i \quad (4.10)$$

where $A_j \in NA$, $a_j \in dom(A_j)$, and $sa_i \in dom(SA)$. The answer to the query, ans, is the number of records in $\mathcal{D}$ satisfying the condition in the WHERE clause. Such answers can be used to learn statistical relationships between the attributes in NA and SA. Given the perturbed data

$\mathcal{D}^*$, ans is approximated by $est = |S^*| * F'$, where S^* is the set of records in $\mathcal{D}^*$ satisfying $A_1 = a_1 \wedge \cdots \wedge A_d = a_d$, $|S^*|$ is the size of S^*, and F' is the MLE given by Lemma 3(ii) based on S^*. The *relative error* of est is defined as $\frac{|est-ans|}{ans}$. A smaller relative error means a larger accuracy and better utility. Queries on only NA are not considered because such queries have zero relative error.

Data mining and analysis typically focuses on low dimensional statistics, such as 1-D or 2-D marginals with a size above a sanity bound [78]. We generated a pool of 5,000 count queries with the query dimensionality d in $\{1, 2, 3\}$ and with the *selectivity*: $ans/|\mathcal{D}| \geq 0.1\%$. For each query, we selected d from $\{1,\ 2,\ 3\}$, selected d attributes from NA without replacement, selected a value $a_i \in dom(A_i)$ for each selected attribute A_i, and finally selected a value $sa_i \in dom(SA)$. All selections are random with equal probability. If the query's selectivity is 0.1% or more, we added it to the pool. We then replaced all NA values in the query with the generalized values obtained from Section 4.1.4 to make sure all NA values have different impacts on SA. We report the average of relative error over all queries in this pool. In addition, since $\mathcal{D}^*$ is randomly generated in each run, we reported the average of 10 runs to avoid the bias of a particular run.

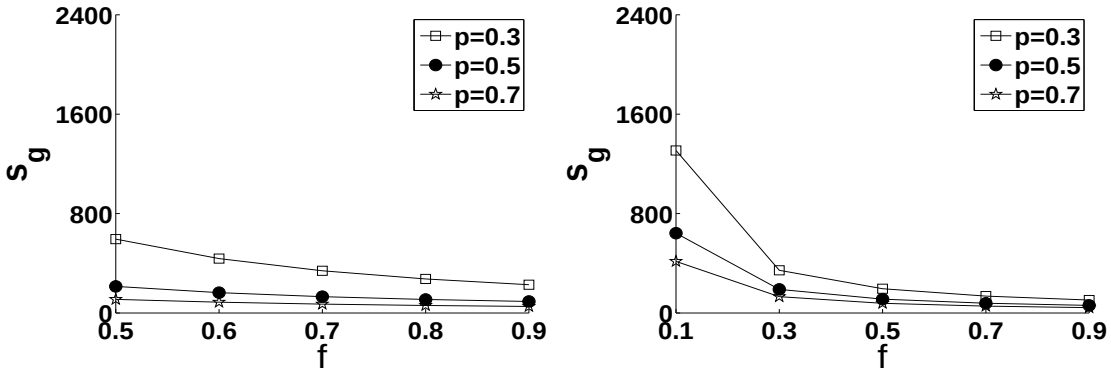
The uniform perturbation, denoted by UP, as described in Section 4.1.1 has been used as a privacy mechanism in [6, 5, 32]. But these privacy mechanisms do not address the disclosure of personal reconstruction. Our method addresses this disclosure by applying UP to sampled data. Although the method proposed by Yang et al. [85] also tried to prevent sensitive disclosures under differential privacy, it focused on a different scenario that sensitive disclosures occurred due to data correlation. When data is correlated, the sensitive information of one record could be referred from other records in the data set. For example, if one family member gets flu, then the rest members in this family are highly possible to get flu as well. We consider a different privacy concern of sensitive disclosures when data has not to be correlated. Therefore, we do not need to compare with Yang's method in this thesis.

Our evaluation has two parts. First, we evaluate how often reconstruction privacy is violated by the perturbed data $\mathcal{D}^*$ produced by UP. Then, we evaluate the cost of achieving reconstruction reconstruction by our SPS algorithm. This cost is measured by the increase in the relative error for queries answered using $\mathcal{D}_2^*$ produced by SPS, compared to the relative error of queries answered using $\mathcal{D}^*$ produced by UP. The same retention probability p is used for both UP and SPS. Table 4.5 shows the settings of p, λ, and δ with the default settings in boldface.

Below, a group means a personal group. First, we study the condition $|g| \leq s_g$ for testing whether a group g^* satisfies reconstruction privacy as described in Section 4.3, where s_g is the

Table 4.5: Parameter Table

Parameters	Settings
p	0.1, 0.3, **0.5**, 0.7, 0.9
λ	0.1, 0.2, **0.3**, 0.4, 0.5
δ	0.1, 0.2, **0.3**, 0.4, 0.5



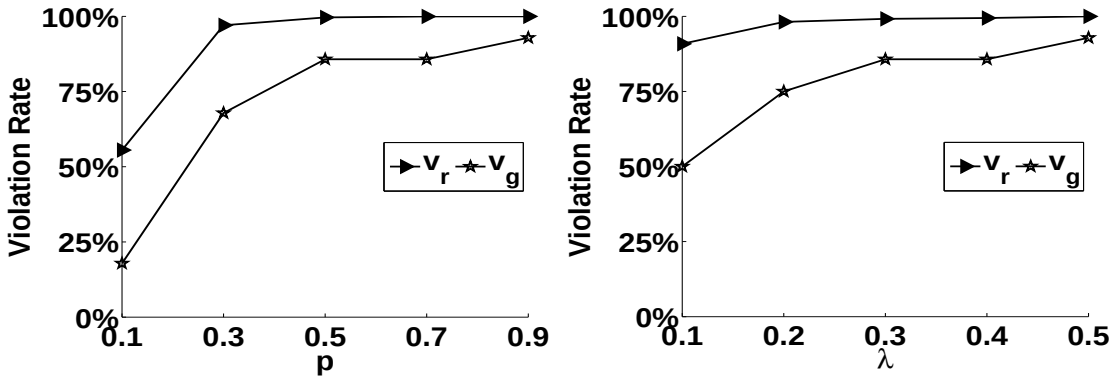
(a) ADULT (b) OCC

Figure 4.1: Maximum Group Size s_g vs. Maximum Frequency f

maximum threshold on the group size defined as

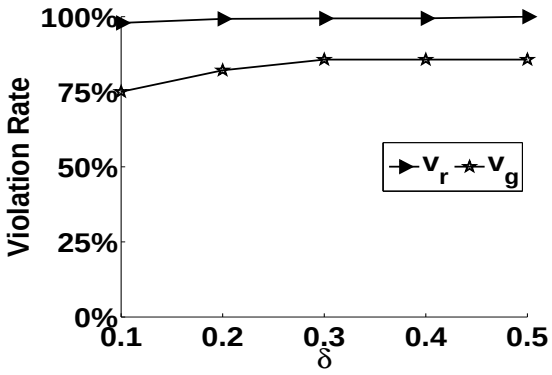
$$s_g = \frac{-2(fp + (1-p)/m)\ln\delta}{(\lambda pf)^2} \tag{4.11}$$

f is the maximum frequency of any SA value occurring in g. Figure 4.1 plots the relationship between s_g and f (for the default settings of λ and δ). Note that the range of f is $[0.5, 0.9]$ for $ADULT$, but is $[0.1, 0.9]$ for OCC. This is because $ADULT$ contains only 2 distinct SA values, as a result, f is at least 50% in all personal groups. Each curve corresponds to a setting of p. For each curve in Figure 4.1, the region above the curve represents the area where this condition fails, that is, $|g| > s_g$ for a given f. The large area above these curves suggests that the maximum group size s_g can be easily exceeded, and thus, there is a good chance of violating reconstruction privacy. Observing both Figure 4.1 and Equation (4.11) we get that, when parameters: λ, δ and p are given, the value of m and f have opposite effects on the value of s_g, particularly, f becomes the dominant factor when f is small (e.g., when $f \leq 0.3$ in Figure 4.1). The value of s_g boosts when f is smaller, implying that personal groups with smaller f tend to be reconstruction private because it is easier for them to satisfy the condition of $|g| \leq s_g$. We will confirm this observation on the two real life data sets shortly.



(a) vs. p

(b) vs. λ



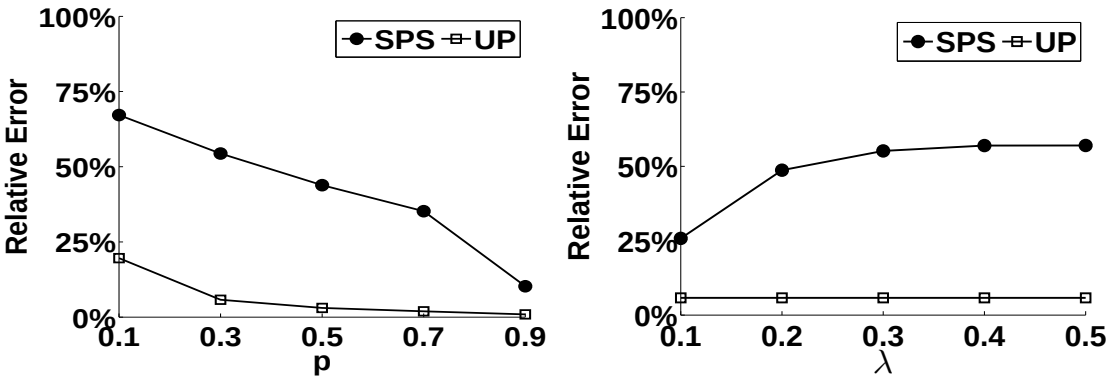
(c) vs. δ

Figure 4.2: ADULT: Privacy Violation

4.4.2 ADULT Data Set

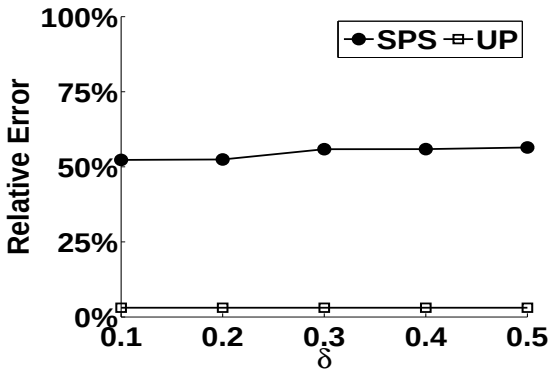
Violation. Figure 4.2 shows the extent to which reconstruction privacy is violated on the perturbed $ADULT$ data set $\mathcal{D}^*$ produced by UP. This *extent* is measured at two levels. v_g represents the percentage of groups that violate reconstruction privacy. v_r represents the percentage of records contained in all violating personal groups, i.e., the coverage of the violating groups in terms of the number of individuals affected. We consider this coverage because all the records in a violating group are under the same risk of accurate personal reconstruction.

Both violations in terms of v_r and v_g are obvious. Take the default setting of $p = 0.5$, $\lambda = 0.3$ and $\delta = 0.3$ as an example. The 85% of all groups are violating and covering more than 99% of the records. This privacy risk is interpreted as follows: with probability of $1 - \delta = 70\%$, the estimate F' of some SA value is within a relative error of $\lambda = 30\%$, and this case covers more than $v_r = 99\%$ of all individuals. The large coverage is expected because a larger group more likely violates reconstruction privacy (Figure 4.1).



(a) vs. p

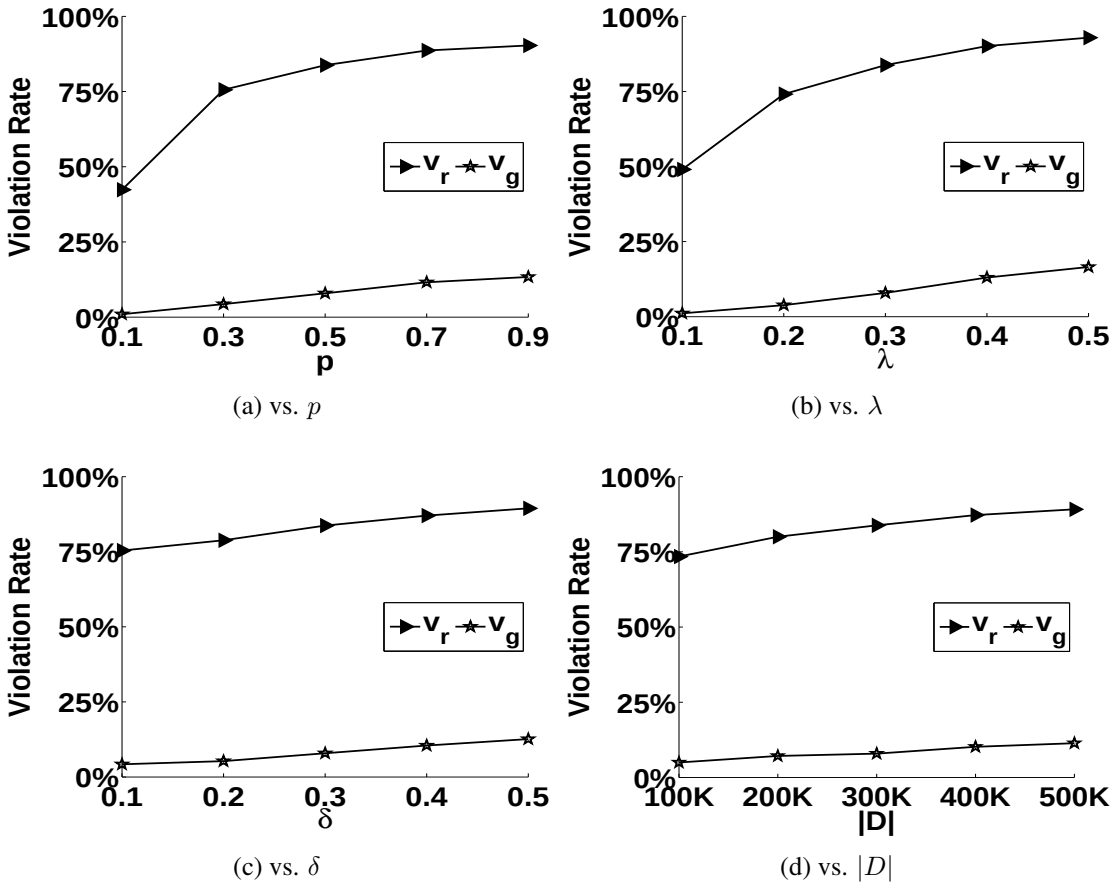
(b) vs. λ



(c) vs. δ

Figure 4.3: ADULT: Relative Error

Cost. Figure 4.3 shows the increase of relative error due to the sampling of SPS. Compared to UP, the relative error for SPS increases about 50% in the *worst case*. This increase is due to the sampling required to eliminate the violation of reconstruction privacy. Considering the large coverage of the violation (i.e., v_r in Figure 4.2), having such increase of error is reasonable. We emphasize that this increase is due to the large f in personal groups in $ADULT$. Recall that f is no less than 50% and when f is larger personal groups tend to violate reconstruction privacy (Figure 4.1). Note that $ADULT$ is not general in real life in terms of very few number of SA values, for other data sets with more SA values, the increased error would be reduced, which will be confirmed soon on the OCC data set. Choosing a small p helps eliminate violation, but also quickly increases the relative error for both UP and SPS (Figures 4.2a and 4.3a). Indeed, a too small p makes the perturbed data become nearly pure noises. This study confirms our discussion at the beginning of Section 4.3 that the approach of reducing p does not preserve utility.



(a) vs. p (b) vs. λ

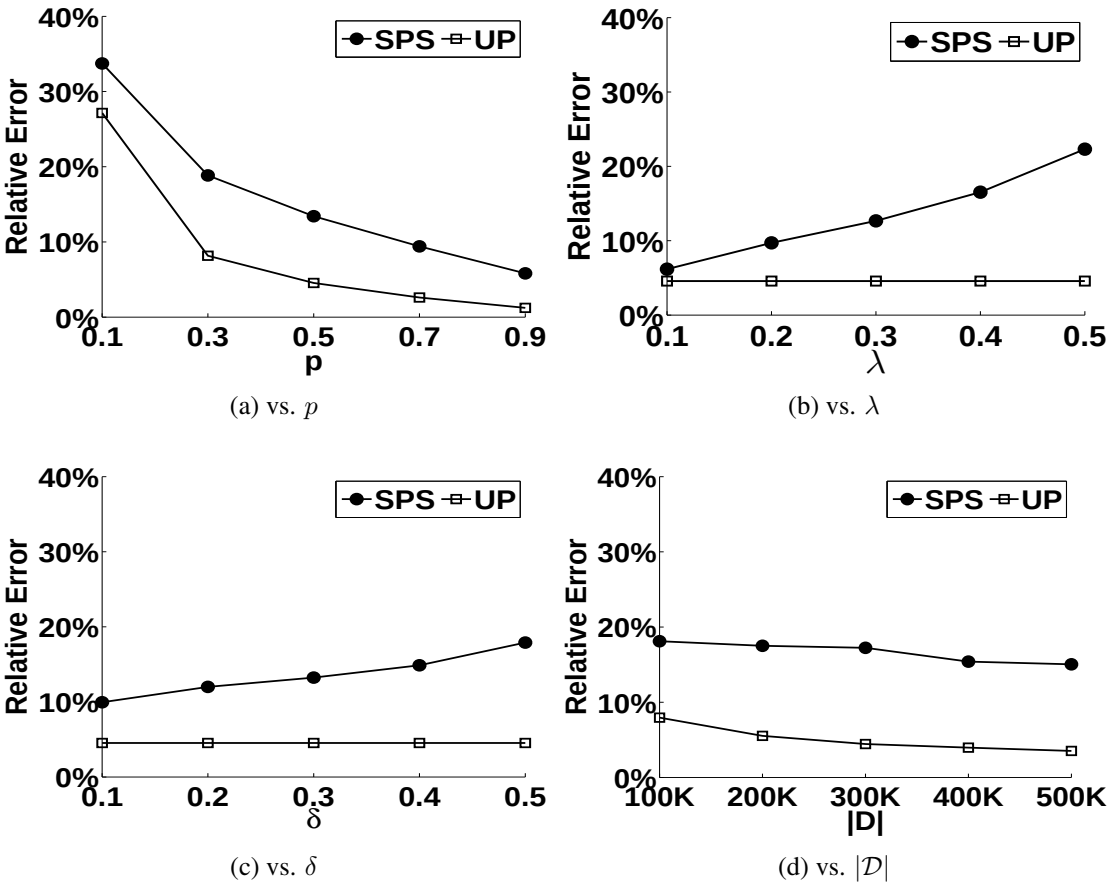
(c) vs. δ (d) vs. $|D|$

Figure 4.4: OCC: Privacy Violation

4.4.3 OCC Data Set

Violation. OCC is a larger data set with a much larger number of balanced distributed SA values. We are curious how this characteristic change would affect our claims. Figure 4.4 shows the extent to which reconstruction privacy is violated. The default data size is 300K when $|\mathcal{D}|$ is not specified. Compared to the $ADULT$ data set, the frequency f of a SA value is much smaller; consequently, the value of s_g is much larger (Figure 4.1). The larger s_g makes it easy to satisfy the condition of $|g| \leq s_g$, therefore, it is less likely that groups in OCC would violate reconstruction privacy, which explains the much smaller v_g and also confirms our claim on Figure 4.1 that smaller f may lead to less reconstruction violations. Besides, the larger s_g implies that violation groups must have larger g because $|g| > s_g$, which explains the small number of violation groups covering the most records in the data set.

Cost. Figure 4.5 compares the relative error of UP and SPS. An obvious difference from the $ADULT$ data set is that there is less increase in the relative error (e.g., less than 10% for most of settings) for SPS compared to the relative error for UP across all settings of parameters. This is



(a) vs. p

(b) vs. λ

(c) vs. δ

(d) vs. $|\mathcal{D}|$

Figure 4.5: OCC: Relative Error

a consequence of the smaller percentage r_g of the violating groups discussed above. In this case, most of the groups do not need sampling because they satisfy reconstruction privacy and only the small number of violating groups will be sampled. Even for such groups, a small reduction in the number of record perturbation is sufficient to increase the error of personal reconstruction to the level required by our privacy criterion.

Another interesting point is that even though a larger data size $|\mathcal{D}|$ causes more violations of reconstruction privacy (Figure 4.4d), it actually decreases the relative error for SPS (Figure 4.5d). As explained above, for this data set, eliminating violation incurs little additional error beyond that of UP. Therefore, as the data size increases, the relative error of UP gets smaller, so does the relative error of SPS. This finding suggests that the proposed SPS algorithm could be more effective on a larger data set.

4.5 Summary

Our empirical studies supported the claim that reconstruction attack could occur on real life data sets, whether they are small or large and whether the number of sensitive attribute is small or large. The studies also supported the claim that the proposed privacy criterion and the sampling method are effective to preserve the utility for data analysis while eliminating such attacks. This effectiveness is more observed on larger data sets with a large number of balanced distributed sensitive attributes.

Chapter 5

Conclusion

Differential privacy has emerged as the gold privacy method for the past two decades. Unlike syntactic methods, differential privacy provides strong privacy guarantees against an adversary with strong background knowledge. In particular, it is claimed that even if the adversary knows all but one record in the data set, the privacy is not violated for the individual behind that one record: the adversary can not observe a distinguishable difference with or without that specific one record on query results [17].

Differential privacy does not treat NIR as a privacy concern. Personal sensitive information may be released through NIR even when some notion of differential privacy is applied. In this thesis we evaluated how likely and how accurately the sensitive information could be released. Syntactic methods use the smoothing operation to prevent disclosures through NIR. Unfortunately, the smoothing operation is also a handicap to statistical relationships learning. We proposed reconstruction privacy, a new privacy criterion with data perturbation and sampling, to achieve both preventing sensitive personal information disclosures and allowing statistical relationships learning.

To summarize, in this thesis we answered two questions:

> (A). To which extent the sensitive information could be learned when some notion of differential privacy is applied?
>
> (B). Is it possible to (1) allow learning statistical relationships (e.g., smoking people tend to have lung cancer), and at the same time, (2) prevent disclosures on sensitive attribute values of individuals in the data set (e.g., Bob is likely to have cancer)? Syntactic privacy methods satisfy (2) but not (1), and differential privacy method satisfies (1) but not (2).

Question (A) is answered in Chapter 3. Disclosures in differential privacy have been investigated in [47, 20, 62] with restricted requirements. For example, the disclosure in [47] requires data

correlation, the disclosure discussed in [62] depends on the simulation of data generation procedure, and [20] needs a Bayes classifier to predict the sensitive attribute value of a target. In this thesis, we proposed a general way of defining disclosures in terms of the probability of a small error in learning sensitive information through NIR. In particular, we only need two queries to launch an attack. A key finding of this study is that this type of disclosures can be used to learn the sensitive information of an individual (such as diseases) with good accuracy whenever differentially private answers have good utility because both disclosure and utility are based on the accuracy of published query answers. For this reason, it would be difficult to prevent such disclosures while providing good utility under differential privacy. Our study suggests that it is important to consider what kind of privacy one wishes to protect. If privacy is about hiding one's participation in the database, differential privacy achieves the goal. If privacy is about protecting individual's sensitive information, differential privacy does not do the job unless one is willing to give up the utility for data analysis. Understanding this limitation of differential privacy is important to avoid unexpected disclosures while enjoying the good utility of differential privacy.

Question (B) is answered in Chapter 4. The key to protecting individual's sensitive information while providing promising utility for statistical learning is separating two types of learning, i.e., learning individual's sensitive information and learning statistical relationships. Our essential insight is to distinguish between reconstruction that aims at a target individual and reconstruction that aims at a larger group of individuals. The former is more relevant to a target individual, thus, should be the focus of protection. We presented a data perturbation approach to prevent sensitive NIR while enabling statistical learning. We achieved these goals through a property implied by the law of large numbers, which allows us to separate these two types of learning by their different responses to reduction in random trials. Based on this idea, we used record sampling to reduce the random trials in data perturbation, which mostly affects NIR specific to an individual while having only a limited effect on statistical learning.

We end this chapter by providing several interesting and promising directions for future work. We briefly discuss these directions:

- **Multiple sensitive attributes**

  In this thesis we assume that the data set $\mathcal{D}$ contains one SA. This assumption is consistent with most of privacy criteria, such as k-anonymity and l-diversity [39]. In real life, however, many research data sets contain multiple SA. For example, in a collection of patient records, both the attribute *Disease* and *Treatment* are sensitive. If these multiple SA are correlated, for

example, the disease can be somehow inferred given the treatment, how this scenario could be protected. This brings in new problems and further research on multiple SA is needed.

- **Adding new records**

  Our approach for achieving reconstruction privacy works properly for static published data. For streaming data, however, some improvement is needed. Based on Equation (4.9) the inserted data may affect the maximum frequency in a personal group, that further affects the maximum number of random trials a personal group could retain, and finally affects the whole perturbation procedure. Thus, finding a way to allow adding new records while providing the same reconstruction privacy is an interesting direction for future work.

- **Customized way to generalize personal groups**

  In Section 4.1.4 the way for producing generalized personal groups is introduced through merging NA values that have the same impact on SA. This is for preventing adversaries to use generalized personal groups to get better estimate accuracy. We only considered the impact on SA from a single NA. For example, the age of 20 and 21 may be integrated to $[20, 21]$. In real life, however, some sets of multiple NA values may have the same impact on SA. For example, a senior female Caucasian and a senior female Asian may have the same probability to suffer Alzheimer's disease. Therefore, the ages of $60 \sim 70$ are aggregated to $[60, 70]$, and at the same time, the race of $Caucasian$ and $Asian$ have to be aggregated to $\{Caucasian, Asian\}$. The impacts of NA values on SA is application specific, thus, further study is required for allowing a customized way to generalize personal groups.

Bibliography

[1] Adult data set. http://archive.ics.uci.edu/ml/datasets/Adult.

[2] Privacy in the age of big data: A time for big decisions. *Standford Law Review*, 2012.

[3] Gergely Acs, Claude Castelluccia, and Rui Chen. Differentially private histogram publishing through lossy compression. In *ICDM*, pages 1–10. IEEE Computer Society, 2012.

[4] Nabil R Adam and John C Worthmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys (CSUR)*, 21(4):515–556, 1989.

[5] Rakesh Agrawal, Ramakrishnan Srikant, and Dilys Thomas. Privacy preserving olap. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 251–262. ACM, 2005.

[6] Shipra Agrawal and Jayant R Haritsa. A framework for high-accuracy privacy-preserving mining. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 193–204. IEEE, 2005.

[7] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *PODS*, pages 273–282. ACM, 2007.

[8] Michael Barbaro and Tom Zeller Jr. A face is exposed for aol searcher no. 4417749. *The New York Times Magazine*, 2006.

[9] Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 217–228. IEEE, 2005.

[10] Chen Bee-Chung, Kifer Daniel, LeFevre Kristen, and Machanavajjhala Ashwin. Privacy-preserving data publishing. *Foundations and Trends in Databases*, 2(1-2):1–167, 2009.

[11] Raghav Bhaskar, Srivatsan Laxman, Adam Smith, and Abhradeep Thakurta. Discovering frequent patterns in sensitive data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 503–512. ACM, 2010.

[12] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The sulq framework. In *PODS*, pages 128–138. ACM, 2005.

[13] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *STOC*, pages 609–618. ACM, 2008.

[14] Jianneng Cao and Panagiotis Karras. Publishing microdata with a robust privacy guarantee. *Proc. VLDB Endow.*, 5(11):1388–1399, July 2012.

[15] Kamalika Chaudhuri and Claire Monteleoni. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*, pages 289–296, 2009.

[16] Rhonda Chaytor and Ke Wang. Small domain randomization: Same privacy, more utility. *Proc. VLDB Endow.*, 3(1-2):608–618, September 2010.

[17] Bee-Chung Chen, Daniel Kifer, Kristen Lefevre, and Ashwin Machanavajjhala. *Privacy-Preserving Data Publishing*. Foundations and Trends(r) in Databases. Now Publishers, 2009.

[18] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[19] Chris Clifton and Tamir Tassa. On syntactic anonymity and differential privacy. *Trans. Data Privacy*, 6(2):161–183, August 2013.

[20] Graham Cormode. Personal privacy vs population privacy: learning to attack anonymization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1253–1261. ACM, 2011.

[21] Graham Cormode, Cecilia Procopiuc, Divesh Srivastava, Entong Shen, and Ting Yu. Differentially private spatial decompositions. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 20–31. IEEE, 2012.

[22] Tore Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15(429-444):2–1, 1977.

[23] Tore Dalenius. Finding a Needle In a Haystack or Identifying Anonymous Census Records. *Journal of Official Statistics*, 2(3):329–336, 1986.

[24] Bolin Ding, Marianne Winslett, Jiawei Han, and Zhenhui Li. Differentially private data cubes: optimizing noise sources and consistency. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 217–228. ACM, 2011.

[25] Wenliang Du and Zhijun Zhan. Using randomized response techniques for privacy-preserving data mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 505–510. ACM, 2003.

[26] Charles Duhigg. How companies learn your secrets. *The New York Times Magazine*, 16, 2012.

[27] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming. ICALP 2006*, pages 1–12. Springer, 2006.

[28] Cynthia Dwork. Differential privacy: A survey of results. In *Theory and applications of models of computation*, pages 1–19. Springer, 2008.

[29] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006*, pages 486–503. Springer, 2006.

[30] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the 24th Annual International Conference on The Theory and Applications of Cryptographic Techniques*, EUROCRYPT'06, pages 486–503, Berlin, Heidelberg, 2006. Springer-Verlag.

[31] Regina C Elandt-Johnson and Norman Lloyd Johnson. *Survival models and data analysis*. Wiley, 1990.

[32] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222. ACM, 2003.

[33] Dan Feldman, Amos Fiat, Haim Kaplan, and Kobbi Nissim. Private coresets. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 361–370. ACM, 2009.

[34] Arik Friedman and Assaf Schuster. Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–502. ACM, 2010.

[35] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, June 2010.

[36] Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and S Yu. Philip. *Introduction to privacy-preserving data publishing: concepts and techniques*. CRC Press, 2010.

[37] Benjamin C. M. Fung, Ke Wang, and Philip S. Yu. Top-down specialization for information and privacy preservation. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 205–216. IEEE, 2005.

[38] Benjamin C. M. Fung, Ke Wang, and Philip S. Yu. Anonymizing classification data for privacy preservation. *Knowledge and Data Engineering, IEEE Transactions on*, 19(5):711–725, 2007.

[39] Tamas S Gal, Zhiyuan Chen, and Aryya Gangopadhyay. A privacy protection model for patient data with multiple sensitive attributes. *IGI Global*, pages 28–44, 2008.

[40] Michaela Götz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao, and Johannes Gehrke. Publishing search logs — a comparative study of privacy guarantees. *Knowledge and Data Engineering, IEEE Transactions on*, 24(3):520–532, 2012.

[41] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems*, pages 2339–2347, 2012.

[42] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *STOC*, pages 705–714. ACM, 2010.

[43] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1-2):1021–1032, 2010.

[44] John Hopcroft and Robert Tarjan. Algorithm 447: Efficient algorithms for graph manipulation. *Communications of the ACM*, 16(6):372–378, 1973.

[45] Vijay S Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 279–288. ACM, 2002.

[46] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[47] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204. ACM, 2011.

[48] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. In *Proceedings of the 18th international conference on World wide web*, pages 171–180. ACM, 2009.

[49] Samuel Kotz, Yan Lumelskii, and Marianna Pensky. The stress-strength model and its generalizations. volume 43, page 44. World Scientific, 2003.

[50] Jaewoo Lee and Chris Clifton. Differential identifiability. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1041–1049. ACM, 2012.

[51] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49–60. ACM, 2005.

[52] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pages 25–25. IEEE, 2006.

[53] C. Li. *Optimizing liner queries under differential privacy*. PhD thesis, Computer Science, University of Massachusetts Amherst, 2013.

[54] Chao Li, Michael Hay, Gerome Miklau, and Yue Wang. A data-and workload-aware algorithm for range queries under differential privacy. *Proceedings of the VLDB Endowment*, 7(5):341–352, 2014.

[55] Chao Li, Michael Hay, Vibhor Rastogi, Gerome Miklau, and Andrew McGregor. Optimizing linear counting queries under differential privacy. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 123–134. ACM, 2010.

[56] Chao Li and Gerome Miklau. An adaptive mechanism for accurate query answering under differential privacy. volume 5, pages 514–525. VLDB Endowment, 2012.

[57] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[58] Tiancheng Li and Ninghui Li. Injector: mining background knowledge for data anonymization. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 446–455. IEEE, 2008.

[59] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 277–286. IEEE, 2008.

[60] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1), March 2007.

[61] Ashwin Machanavajjhala, Aleksandra Korolova, and Atish Das Sarma. Personalized social recommendations: accurate or private. *Proceedings of the VLDB Endowment*, 4(7):440–450, 2011.

[62] David McClure and Jerome P Reiter. Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Transactions on Data Privacy*, 5(3):535–552, 2012.

[63] Frank McSherry and Ilya Mironov. Differentially private recommender systems: building privacy into the net. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–636. ACM, 2009.

[64] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM, 2004.

[65] Saralees Nadarajah and Samuel Kotz. A note on the ratio of normal and laplace random variables. volume 15, pages 151–158. Springer, 2006.

[66] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[67] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84. ACM, 2007.

[68] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[69] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 1988.

[70] Vibhor Rastogi and Suman Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 735–746. ACM, 2010.

[71] Vibhor Rastogi, Dan Suciu, and Sungho Hong. The boundary between privacy and utility in data publishing. In *Proceedings of the 33rd international conference on Very large data bases*, pages 531–542. VLDB Endowment, 2007.

[72] Pierangela Samarati. Protecting respondents identities in microdata release. *Knowledge and Data Engineering, IEEE Transactions on*, 13(6):1010–1027, 2001.

[73] Melvin Dale Springer. *The algebra of random variables*. Wiley New York, 1979.

[74] Alan Stuart and Keith Ord. *Kendall's advanced theory of statistics*, volume 1. Arnold, London, 6 edition, 1998.

[75] Latanya Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002.

[76] Yufei Tao, Xiaokui Xiao, Jiexing Li, and Donghui Zhang. On anti-corruption privacy preserving publication. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 725–734. IEEE, 2008.

[77] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. volume 60, pages 63–69. Taylor & Francis Group, 1965.

[78] Xiaokui Xiao, Gabriel Bender, Michael Hay, and Johannes Gehrke. ireduct: Differential privacy with reduced relative errors. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 229–240. ACM, 2011.

[79] Xiaokui Xiao and Yufei Tao. Anatomy: Simple and effective privacy preservation. In *Proceedings of the 32nd international conference on Very large data bases*, pages 139–150. VLDB Endowment, 2006.

[80] Xiaokui Xiao, Yufei Tao, and Minghua Chen. Optimal random perturbation at multiple privacy levels. *Proceedings of the VLDB Endowment*, 2(1):814–825, 2009.

[81] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Differential privacy via wavelet transforms. *IEEE Transactions on Knowledge and Data Engineering*, 23(8):1200–1214, 2011.

[82] Yonghui Xiao, James Gardner, and Li Xiong. Dpcube: Releasing differentially private data cubes for health information. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1305–1308. IEEE, 2012.

[83] Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, Ge Yu, and Marianne Winslett. Differentially private histogram publication. *The VLDB Journal*, 22(6):797–822, December 2013.

[84] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 785–790. ACM, 2006.

[85] Bin Yang, Issei Sato, and Hiroshi Nakagawa. Bayesian differential privacy on correlated data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 747–762. ACM, 2015.

[86] Yin Yang, Zhenjie Zhang, Gerome Miklau, Marianne Winslett, and Xiaokui Xiao. Differential privacy in data publication and analysis. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 601–606. ACM, 2012.

Appendix A

Computing CDF of Y/X

In this section, we prove Lemma 2 in Section 3.3.1. Recall $F_Z(z) = Pr\left[\frac{Y}{X} \leq z\right]$, where $Z = \frac{Y}{X}$, both X and Y are random variables for noisy versions of the query answers ϕ and θ after adding a Laplace noise. So the noises $x - \phi$ and $y - \theta$ follow the Laplace distribution below.

$$f_X(x) = \frac{1}{2b} \exp\left(-\frac{|x-\phi|}{b}\right) \tag{A.1}$$

$$f_Y(y) = \frac{1}{2b} \exp\left(-\frac{|y-\theta|}{b}\right) \tag{A.2}$$

Note that $0 < \theta \leq \phi$, $\phi > 0$, and $b > 0$.

Lemma 2. Assume $z \neq 0$ and $z \neq \pm 1$.

- For $z < 0$,

$$\begin{aligned} F_Z^1(z) = &\left[\frac{z^2}{2(1-z^2)}\right]\left[\exp\left(\frac{\theta - z\phi}{zb}\right)\right] - \frac{1}{2(z+1)}\left[\exp\left(\frac{-(\theta+\phi)}{b}\right)\right] \\ &+ \frac{1}{2}\exp\left(-\frac{\phi}{b}\right) - \frac{1}{2(z^2-1)}\left[\exp\left(\frac{z\phi-\theta}{b}\right)\right] \end{aligned} \tag{A.3}$$

- For $0 < z \leq \frac{\theta}{\phi}$,

$$\begin{aligned} F_Z^2(z) = &\left[\frac{z^2}{2(z^2-1)}\right]\left[exp\left(\frac{z\phi - \theta}{zb}\right)\right] - \frac{1}{2(z+1)}\left[exp\left(\frac{-(\theta+\phi)}{b}\right)\right] \\ &+ \frac{1}{2}exp\left(-\frac{\phi}{b}\right) - \frac{1}{2(z^2-1)}\left[exp\left(\frac{z\phi-\theta}{b}\right)\right] \end{aligned} \tag{A.4}$$

- For $z > \frac{\theta}{\phi}$,

$$F_Z^3(z) = \frac{z^2}{2(1-z^2)}\left[exp\left(\frac{\theta - z\phi}{zb}\right)\right] - \frac{1}{2(z+1)}\left[exp\left(\frac{-(\theta+\phi)}{b}\right)\right] + 1 + \frac{1}{2}exp\left(-\frac{\phi}{b}\right) - \frac{1}{2(1-z^2)}\left[exp\left(\frac{\theta - z\phi}{b}\right)\right] \tag{A.5}$$

□

Proof. Rewrite $F_Z(z)$ as follows.

$$\begin{aligned} F_Z(z) &= Pr\left[\frac{Y}{X} \leq z\right] \\ &= Pr\left[Y \geq zX, X < 0\right] + Pr\left[Y \leq zX, X > 0\right] \\ &= A + B \end{aligned} \tag{A.6}$$

where $A = \int_{-\infty}^{0^-}\left[\int_{zx}^{\infty} f_Y(y)dy\right] f_X(x)\,dx$, $B = \int_{0^+}^{\infty}\left[\int_{-\infty}^{zx} f_Y(y)dy\right] f_X(x)\,dx$ and $\int_{-\infty}^{0^-}$ represents the integration range of $(-\infty, 0)$ and $\int_{0^+}^{\infty}$ indicates the integration range of $(0, \infty)$.

z could be either positive or negative. We discuss the situation when $z > 0$ in detail. And the case of $z < 0$ could be computed in similar way.

A.1 The Case of $z > 0$

A.1.1 Computing A

Instantiating Equation (A.2) for $f_Y(y)$ into A,

$$\begin{aligned} A &= \int_{-\infty}^{0^-}\left[\int_{zx}^{\infty} f_Y(y)dy\right] f_X(x)\,dx \\ &= \int_{-\infty}^{0^-}\left[\int_{zx}^{\infty} \frac{1}{2b}exp\left(-\frac{|y-\theta|}{b}\right)dy\right] f_X(x)\,dx \end{aligned} \tag{A.7}$$

To remove the absolute sign, we consider two cases:

- CASE 1: $zx \geq \theta$, i.e., $x \geq \frac{\theta}{z}$ (because $z > 0$). Since $\theta \geq 0$, we get $x \geq 0$, which is contradicting the integration range for x, i.e., $(-\infty, 0)$. Therefore, this case is not possible.
- CASE 2: $zx < \theta$, i.e., $x < \frac{\theta}{z}$ (because $z > 0$). We divide the range of y, $[zx, \infty)$, into $[zx, \theta]$ and (θ, ∞), so $A = A_1 + A_2$, where
  $A_1 = \int_{-\infty}^{0^-}\left[\int_{zx}^{\theta} \frac{1}{2b}exp\left(\frac{y-\theta}{b}\right)dy\right] f_X(x)\,dx$ and
  $A_2 = \int_{-\infty}^{0^-}\left[\int_{\theta^+}^{\infty} \frac{1}{2b}exp\left(\frac{\theta-y}{b}\right)dy\right] f_X(x)\,dx$.

To compute A_1, let u be $\frac{y-\theta}{b}$: $du = \frac{1}{b}dy$; at $y = zx$, $u = \frac{zx-\theta}{b}$; at $y = \theta$, $u = 0$. We get

$$\begin{aligned} A_1 &= \int_{-\infty}^{0} \left[\int_{\frac{zx-\theta}{b}}^{0} \frac{1}{2} exp(u) du \right] f_X(x)\, dx \\ &= \int_{-\infty}^{0} \left[\frac{1}{2} exp(u) \Big|_{\frac{zx-\theta}{b}}^{0} \right] f_X(x)\, dx \\ &= \frac{1}{2} \int_{-\infty}^{0} \left[1 - exp\left(\frac{zx-\theta}{b} \right) \right] f_X(x)\, dx \end{aligned} \tag{A.8}$$

Similarly we could get $A_2 = \frac{1}{2} \int_{-\infty}^{0^-} f_X(x)\, dx$.

$$A = A_1 + A_2 = \frac{1}{2} \int_{-\infty}^{0^-} \left[2 - exp\left(\frac{zx-\theta}{b} \right) \right] f_X(x)\, dx \tag{A.9}$$

Instantiating $f_X(x)$ in Equation (A.9) using Equation (A.1), and since $x < 0$ and $\phi > 0$, $-\frac{|x-\phi|}{b}$ is equal to $\frac{x-\phi}{b}$.

$$\begin{aligned} A &= \frac{1}{2} \int_{-\infty}^{0^-} \left[2 - exp\left(\frac{zx-\theta}{b} \right) \right] \frac{1}{2b} exp\left(-\frac{|x-\phi|}{b} \right) dx \\ &= \frac{1}{2} \int_{-\infty}^{0^-} \frac{1}{b} exp\left(\frac{x-\phi}{b} \right) dx - \frac{1}{4} \int_{-\infty}^{0^-} \frac{1}{b} \left[exp\left(\frac{(z+1)x - (\theta+\phi)}{b} \right) \right] dx \\ &= \frac{1}{2} exp(-\frac{\phi}{b}) - \frac{1}{4(z+1)} \left[exp\left(\frac{-(\theta+\phi)}{b} \right) \right] \end{aligned} \tag{A.10}$$

A.1.2 Computing B

Similarly we can compute B. Replacing $f_Y(y)$ with Equation (A.2), we get

$$\begin{aligned} B &= \int_{0^+}^{\infty} \left[\int_{-\infty}^{zx} f_Y(y) dy \right] f_X(x)\, dx \\ &= \int_{0^+}^{\infty} \left[\int_{-\infty}^{zx} \frac{1}{2b} exp\left(-\frac{|y-\theta|}{b} \right) dy \right] f_X(x)\, dx \end{aligned} \tag{A.11}$$

To remove the absolute sign, we consider two cases:

- CASE 1: one is $zx \geq \theta$, i.e., $x \geq \frac{\theta}{z}$ (because $z > 0$), and the integration of y over $(-\infty, zx]$ is equal to the integration over $(-\infty, \theta]$ with $-\frac{|y-\theta|}{b}$ being equal to $\frac{y-\theta}{b}$, plus the integration over $(\theta, zx]$ with $-\frac{|y-\theta|}{b}$ being equal to $-\frac{y-\theta}{b}$.
- CASE 2: the second case is $zx < \theta$, i.e., $x < \frac{\theta}{z}$ (because $z > 0$), in this case $-\frac{|y-\theta|}{b}$ is equal to $\frac{y-\theta}{b}$.

Through elementary integration we get the following (Notice that the two cases have been reflected by the two integrations).

$$B = \frac{1}{2}\int_{\left(\frac{\theta}{z}\right)}^{\infty}\left[2 - exp\left(\frac{\theta - zx}{b}\right)\right] f_X(x)\,dx + \frac{1}{2}\int_{0+}^{\left(\frac{\theta}{z}\right)^-}\left[exp\left(\frac{zx-\theta}{b}\right)\right] f_X(x)\,dx \quad \text{(A.12)}$$

Substituting Equation (A.1) for $f_X(x)$, we get $B = B_L + B_R$ where

$$B_L = \frac{1}{2}\int_{\left(\frac{\theta}{z}\right)}^{\infty}\left[2 - exp\left(\frac{\theta - zx}{b}\right)\right]\left[\frac{1}{2b}exp\left(-\frac{|x-\phi|}{b}\right)\right] dx$$

$$B_R = \frac{1}{2}\int_{0+}^{\left(\frac{\theta}{z}\right)^-}\left[exp\left(\frac{zx-\theta}{b}\right)\right]\left[\frac{1}{2b}exp\left(-\frac{|x-\phi|}{b}\right)\right] dx$$

To remove the absolute sign, we consider the following two cases:

- CASE i: $\frac{\theta}{z} \geq \phi$, i.e., $z \leq \frac{\theta}{\phi}$. For B_L, $-\frac{|x-\phi|}{b}$ is equal to $\frac{\phi-x}{b}$. For B_R, we can express the integration range of x over $(0, \frac{\theta}{z})$ as the sum of those over $[0, \phi]$ and $(\phi, \frac{\theta}{z}]$. Through elementary integration we get

$$\begin{aligned} B = &\left[\frac{z^2}{2(z^2-1)}\right]\left[exp\left(\frac{z\phi-\theta}{zb}\right)\right] - \frac{1}{2(z^2-1)}\left[exp\left(\frac{z\phi-\theta}{b}\right)\right] \\ &- \frac{1}{4(z+1)}\left[exp\left(\frac{-(\theta+\phi)}{b}\right)\right] \end{aligned} \quad \text{(A.13)}$$

- CASE ii: $\frac{\theta}{z} < \phi$, i.e., $z > \frac{\theta}{\phi}$. In this case, for B_L, we express the integration range of x over $[\frac{\theta}{z}, \infty)$ as the sum of those over $[\frac{\theta}{z}, \phi]$ and (ϕ, ∞). For B_R, $-\frac{|x-\phi|}{b}$ is equal to $\frac{x-\phi}{b}$. Through elementary integration we can get

$$\begin{aligned} B = &\frac{z^2}{2(1-z^2)}\left[exp\left(\frac{\theta-z\phi}{zb}\right)\right] - \frac{1}{2(1-z^2)}\left[exp\left(\frac{\theta-z\phi}{b}\right)\right] \\ &- \frac{1}{4(z+1)}\left[exp\left(\frac{-(\theta+\phi)}{b}\right)\right] + 1 \end{aligned} \quad \text{(A.14)}$$

A.2 The Case of $z < 0$

Following a similar procedure, we get A and B as following when $z < 0$.

$$A = \frac{z^2}{2(1-z^2)}\left[\exp\left(\frac{\theta-z\phi}{zb}\right)\right] - \frac{1}{4(z+1)}\left[\exp\left(\frac{-(\theta+\phi)}{b}\right)\right] + \frac{1}{2}\exp\left(-\frac{\phi}{b}\right) \quad \text{(A.15)}$$

$$B = -\frac{1}{2(z^2-1)}\left[\exp\left(\frac{z\phi-\theta}{b}\right)\right] - \frac{1}{4(z+1)}\left[\exp\left(\frac{-(\theta+\phi)}{b}\right)\right] \tag{A.16}$$

Notice that $z < 0$ implies that Case 2, i.e., $z > \frac{\theta}{\phi}$, never occurs (because $\frac{\theta}{\phi} > 0$) , thus, B has only Case 1, i.e., $z \leq \frac{\theta}{\phi}$.

Table A.1: $F_Z(z)$

$F_Z(z)$	= A + B
$F_Z^1(z) \quad (z < 0)$	A is given by Equation (A.15), B is given by Equation (A.16)
$F_Z^2(z) \quad (0 < z \leq \frac{\theta}{\phi})$	A is given by Equation (A.10), B is given by Equation (A.13)
$F_Z^3(z) \quad (z > \frac{\theta}{\phi})$	A is given by Equation (A.10), B is given by Equation (A.14)

A.3 Sum Things Up

Recall $F_Z(z) = A + B$, and $F_Z^1(z)$ is $F_Z(z)$ for $z < 0$, $F_Z^2(z)$ is $F_Z(z)$ for $0 < z \leq \frac{\theta}{\phi}$, and F_Z^3 is $F_Z(z)$ for $z > \frac{\theta}{\phi}$. We summarize the computation of $F_Z(z)$ for the various cases of z in Table A.1. This completes the proof of Lemma 2. □