Reconstruction Privacy: Enabling Statistical Learning

Ke Wang
Simon Fraser University
Singapore Management University
wangk@cs.sfu.ca

Chao Han
Simon Fraser University
hanchaoh@cs.sfu.ca

Ada Waichee Fu
Chinese University of Hong Kong
adafu@cse.cuhk.edu.hk

Raymond Chi Wing Wong
Hong Kong University of Science and Technology
raywong@cse.ust.hk

Philip S. Yu
University of Illinois at Chicago
Institute for Data Science
Tsinghua University (Beijing)
psyu@cs.uic.edu

ABSTRACT

Non-independent reasoning (NIR) allows the information about one record in the data to be learnt from the information of other records in the data. Most posterior/prior based privacy criteria consider NIR as a privacy violation and require to smooth the distribution of published data to avoid sensitive NIR. 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. A question is whether it is possible to (1) allow learning statistical relationships, yet (2) prevent sensitive NIR about an individual. 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).

Categories and Subject Descriptors

H.2.7 [**Database Management**]: Database Administration—*Security, integrity, and protection*; H.2.8 [**Database Applications**]: Data Mining

General Terms

Algorithm, Data Privacy, Theory

Keywords

Data Privacy, Differential Privacy

1. INTRODUCTION

1.1 Motivation

Many privacy definitions/criteria have been proposed in the literature and many ways exist to categorize them, such as semantic methods vs syntactic methods, prior/posterior methods vs differential methods, etc. See surveys [1][2][3] for details. Another way to categorize privacy definitions is by whether *non-independent reasoning (NIR)* is considered as a privacy violation. In NIR, the information about one record in the data can be learnt from the information of other records in the data, under the assumption that these records follow the same underlying distribution. A classifier is a master example of NIR where the class information of a new instance is learnt from the distribution in a related training set.

Most posterior/prior based privacy definitions consider NIR as a privacy violation, such as l-diversity [4], t-closeness [5], ρ_1-ρ_2 privacy [6], β-likeness [7], *small sum privacy* [8] and Δ-growth [9]. These criteria quantify the risk to an individual by the information learnt from the subpopulation containing that individual. To avoid privacy violation, the information learnt is required to have a small change compared to a prior of an adversary, and this often requires to "smooth" the distribution in the published data. One drawback of this approach is that it is hard to model the prior of the attacker [10][11]. Another drawback is that it limits the desired utility of learning statistical relationships. For example, Δ-growth postulates that the distribution in each subpopulation should be close to the global distribution in the whole data set. This requirement makes it difficult to learn novel statistical relationships such as "smokers tend to have lung cancer" in the subpopulation of smokers.

At the other side of the aisle, differential privacy [10] considers NIR as a non-privacy violation, as stated in [11] (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". Instead of avoiding the occurrence of disclosures, the differential privacy criterion seeks to mask the impact of a single individual on such occurrences. A popularized claim is that, even if an attacker knows all but one records, the attacker will not learn much about the remaining tuple. As indicated above, this comes with the price of permitting disclosures through NIR. Indeed, the recent study in [12] suggests that disclosures could occur under differential privacy if records are correlated, and the study in [13] demonstrates that a Bayes classifier could be built using only differentially private answers to predict the sensitive attribute of an individual. In this paper we propose that a sensitive disclosure of NIR could occur in more general cases: no correlation among

Table 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

records is required and only two differentially private query answers are needed to infer the sensitive attribute value. The example below demonstrates such a disclosure.

EXAMPLE 1. *Consider the* ADULT *data set [14] that contains 45,222 records (without missing values) from the 1994 Census database. We did not observe any record correlation in this data set. 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 $Conf = \frac{ans_2}{ans_1} = 0.8383$.

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

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 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. $Conf'$ 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 [15] 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 $\frac{ans'_2}{ans'_1}$. 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 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.

1.2 Our Approach

The question we study in this paper 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, posterior/prior based privacy criteria provide (B) but not (A), whereas the differential privacy criterion provides (A) but not (B). The difficulty of providing (A) and (B) is that they both make use of NIR, one for utility and one for privacy violation. The key lies at distinguishing these two types of learning. The next example illustrates the ideas of our approach.

EXAMPLE 2. *Consider a table $D(Gender, Job, Disease)$, 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 D, uniform perturbation [16] 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 D^* denote the perturbed data.*

D^* can be utilized to reconstruct the distribution of Disease in a given subset of records. Consider any subset S of D, the counterpart S^* for 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 following equation follows from the perturbation operation applied to the data:

$$E[F^*_d] = (0.2 + 0.8/10) f_d + (0.8/10)(1 - f_d) \qquad (1)$$

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) *[16] computed using the perturbed S^*.*

Given the published D^*, suppose that an adversary tries to learn the likelihood that Bob, a male engineer with a record in D, has breast cancer or BC for short. One way is considering the subset S_{me} for all male engineers in D, and another is considering the subset S_e for all engineers in D. Let M^{me}_d and M^e_d be the MLE for a disease d in S_{me} and S_e, respectively. Two questions can be asked.

Question 1: Which of M^{me}_{BC} and M^e_{BC} should be used to quantify the risk to Bob? *S_{me} contains exactly the records that match all Bob's public information, whereas S_e contains additional records that do not belong to Bob. Without further information, S_{me} is more relevant to Bob than S_e, so M^{me}_{BC} should be used as the risk to Bob. If the additional records for female engineers follow a different distribution on BC from those for male engineers, M^e_{BC} most likely is not useful for inferring whether Bob has breast*

cancer. We will discuss the case where the additional records have the same distribution as Bob in Section 3.4. On the other hand, the frequency M_d^e for some disease d (e.g., cervical spondylosis) may be useful for data analysis, such as learning the statistical relationship that career engineers tend to have d. This leads to the next question.

Question 2: How to limit the accuracy of M_{BC}^{me} while preserving the accuracy of M_d^e for data analysis? *The errors of M_d^{me} and M_d^e were caused by approximating the unknown $E[F_d^*]$ with the observed f_d^* in Equation (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_e^* contains more records than S_{me}^*, M_d^e is more accurate for estimating the frequency of d in S_e than M_d^{me} for estimating the frequency of d in S_e. We can leverage this gap to limit the accuracy of M_{BC}^{me} while preserving the accuracy of M_d^e.* □

This example illustrates two types of reconstruction for MLEs. The reconstruction of M_{BC}^{me} based on S_{me} is called *personal reconstruction* because it aims at a particular individual by matching all public attributes of Bob; the reconstruction of M_d^e based on S_e is called *aggregate reconstruction* because it aims at a large population without specifically targeting any individual. We argue (in Section 3.2) 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.

The small count privacy and large count utility in [8] use the number of records involved to distinguish the reconstruction for privacy concern and the reconstruction for utility. It is not clear how to set appropriate thresholds for such sizes. Indeed, it could be the case that two reconstructions are performed on two subsets of data with the same size but one aims at finding an individual's sensitive information while the other aims at finding general patterns.

1.3 Contributions

Here are the main contributions in this work:

Contribution 1 (Section 2): We present a condition to characterize the occurrence of disclosures of differentially private answers through NIR. For the Laplace noise distribution, this condition is simple and neat as it is expressed in terms of the ratio of the scale factor to the query answer.

Contribution 2 (Section 3): We propose an *inaccuracy requirement* on personal reconstruction as a new privacy criterion called *reconstruction privacy*. This criterion imposes a minimum value δ for the *best* upper bound on $\Pr\left[\frac{F'-f}{f} > \lambda\right]$ for the actual and estimated frequency f and F' of a sensitive value in a personal reconstruction, where δ and λ are privacy parameters. Note that $\frac{F'-f}{f}$ is the error of the reconstruction for f, which should *not* be confused with the relative increase of the attacker's belief such as the β-likeness [7], (n,t)-closeness [17] and $(c,2)$-diversity [4]. This criterion 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, this criterion avoids modeling the prior of an adversary, which can be tricky as shown in [10][11].

Contribution 3 (Sections 4): We present an efficient test of reconstruction privacy. First, we show a conversion between an upper bound for the tail probability of Poisson trials into an upper bound on $\Pr\left[\frac{F'-f}{f} > \lambda\right]$. 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.

Contribution 4 (Section 5): We present an efficient algorithm for producing a perturbed version D^* that satisfies a given specification of reconstruction privacy. The algorithm is highly efficient because it only needs to sort the records once and make another scan on the sorted data.

Contribution 5 (Section 6): We evaluate two claims. The first claim is that reconstruction privacy could 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.

2. OBSERVATIONS ON DIFFERENTIAL PRIVACY

In this section, we answer the question under what conditions would differentially private answers disclose sensitive information through NIR? The standard ϵ-differential privacy mechanism [10] ensures that, for any two data sets D_1 and D_2 differing on at most one record, for all queries Q of interest, and for any value α in the range for noisy answers, $\Pr[K(D_1,Q) = \alpha] \leq exp(\epsilon)\Pr[K(D_2,Q) = \alpha]$, where $K(D_i,Q)$ is a noisy answer $a+\xi$ for the actual answer a and a random noise ξ following some distribution. The *scale* $E[|\xi|]$ of noises depends on the query class and the privacy parameter ϵ. The purpose of the noise is to mask the impact of a single record on query answers.

Let us construct a disclosure by differentially private answers. Let SA denote the sensitive attribute (e.g., diseases) and NA denote the public attributes. Suppose that an adversary tries to determine whether a participating individual t has a particular value sa on SA. Let $t.NA$ denote the values for t on NA, which is known to the adversary. The adversary issues two count queries:

$$\begin{aligned} Q_1 &: NA = t.NA \\ Q_2 &: NA = t.NA \wedge SA = sa \end{aligned} \tag{2}$$

Let $X = x + \xi_1$ and $Y = y + \xi_2$ be the noisy answers for Q_1 and Q_2 returned by the ϵ-differential privacy mechanism, where x and y are actual answers and ξ_i's are the noises. Note $\frac{y}{x} \leq 1$ and $\frac{y}{x}$ represents the chance that t has the sa value on SA. Note $\frac{Y}{X} = \frac{y+\xi_2}{x+\xi_1} = \frac{y/x+\xi_2/x}{1+\xi_1/x}$.

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 x increases, ξ_2/x and ξ_1/x decrease and $\frac{Y}{X}$ approaches $\frac{y}{x}$. If $\frac{y}{x}$ 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 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 x and y be the true answers to Q_1 and Q_2, $x \neq 0$. Let $X = x + \xi_1$ and $Y = y + \xi_2$ be the noisy answers for Q_1 and Q_2 with the noises ξ_i having the zero mean and the variance V. Then*

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

PROOF. Note that $E[\frac{Y}{X}]$ is *not* equal to $\frac{E[Y]}{E[X]}$. Using the Taylor expansion technique [18, 19], $E[\frac{Y}{X}]$ 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. $E[X] = x$ and $E[Y] = y$ (because noises have the zero mean), $Var[X] = Var[Y] = V$, and the covariance $cov[X,Y] = cov[x+\xi_1, y+\xi_2] = cov[\xi_1, \xi_2]$. 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 x increases, $\frac{V}{x^2}$ decreases, $E[\frac{Y}{X}]$ approaches $\frac{y}{x}$ and $Var[\frac{Y}{X}]$ approaches zero. In general, $E[\frac{Y}{X}]$ approaching $\frac{y}{x}$ does not entail $\frac{Y}{X}$ approaching $\frac{y}{x}$, 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 $\frac{y}{x}$. This is summarized in the next corollary.

COROLLARY 1. *For any noise distribution with the zero mean and a fixed variance V, as the query answer x increases, $\frac{Y}{X}$ approaches $\frac{y}{x}$.*

To our knowledge, Corollary 1 covers all noise distributions employed by the differential privacy mechanism, including Laplace mechanism [10], Gaussian mechanism [20], and Matrix mechanism [21], because these distributions have a zero mean and a fixed variance. To see how large x is needed for $\frac{Y}{X}$ to be accurate enough for $\frac{y}{x}$, let us consider the Laplace mechanism $Lap(b) = \frac{1}{2b}exp(-|\xi|/b)$, but a similar analysis can be performed for other mechanisms. b is the *scale factor*. $Lap(b)$ 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 $\frac{y}{x} \leq 1$ and $\frac{V}{x^2} = \frac{2b^2}{x^2} = 2\left(\frac{b}{x}\right)^2$ into Lemma 1 and simplifying, we get a simple bound on $|E[\frac{Y}{X}] - \frac{y}{x}|$ and $Var[\frac{Y}{X}]$ in terms of the scale factor b and the query answer x (but not the privacy parameter ϵ or the sensitivity Δ of queries).

COROLLARY 2. *Let X and Y be the noisy answers for actual answers x and y, where the noises follow the Laplace distribution $Lap(b)$. (i) $|E[\frac{Y}{X}] - \frac{y}{x}| \leq 2\left(\frac{b}{x}\right)^2$. (ii) $Var[\frac{Y}{X}] \leq 4\left(\frac{b}{x}\right)^2$.*

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

b \ x	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

Thus, the value of $2\left(\frac{b}{x}\right)^2$ is an indicator of how close $\frac{Y}{X}$ is to $\frac{y}{x}$. Table 2 shows the values of $2\left(\frac{b}{x}\right)^2$ for various query answers x and settings of b (within the brackets is the corresponding privacy parameter ϵ for the setting of $\Delta = 2$, which accounts for answering the two queries Q_1 and Q_2 in a row). The boldface highlights where $2\left(\frac{b}{x}\right)^2$ is small enough so that $\frac{Y}{X}$ is a good indicator of $\frac{y}{x}$. Take $(b = 20, x = 500)$ as an example where $2\left(\frac{b}{x}\right)^2 = 0.0032$. $|E[\frac{Y}{X}] - \frac{y}{x}| \leq 0.0032$ 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 $\frac{b}{x}$: as a rule of thumb, a ratio $\frac{b}{x} \leq \frac{1}{20}$ would ensure that $\frac{Y}{X}$ is a good indicator of $\frac{y}{x}$ because $2\left(\frac{b}{x}\right)^2 \leq \frac{2}{400}$. In this case, if $\frac{y}{x}$ 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 cannot be avoided by choosing a large scale factor b if the actual answer x can be arbitrarily large.

We end this section with an explicit acknowledgement of disclosures by differential privacy from [10]: "Note that a bad disclosure can still occur, but our guarantee assures the individual that it will not be the presence of her data that causes it, nor could the disclosure be avoided through any action or inaction on the part of the user". 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.

3. PROBLEM STATEMENT

We define our model of data perturbation, privacy criterion, and the problems we will study.

3.1 Data Perturbation

As in [7, 9, 22], we consider a table 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 [23].

Prior knowledge on negative correlation [24] 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 [25][23]. 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, Dis-

ease=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.

We produce the perturbed version D^* of D by applying *uniform perturbation* [25][16][6] on SA as follows. For a given retention probability p, where $0 < p < 1$, for each record in D, we toss a coin with head probability p. If the coin lands on head, retain the SA value in the record; if the coin lands on tail, replace the SA value in the record 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 $\mathbb{P}_{m\times m}$:

$$\mathbb{P}_{ji} = \begin{cases} p + \frac{1-p}{m} & \text{if } j{=}i \text{ (retain } sa_i\text{)} \\ \frac{1-p}{m} & \text{if } j{\neq}i \text{ (perturb } sa_i \text{ to } sa_j\text{)} \end{cases} \quad (3)$$

A proper choice of the retention probability p can ensure some privacy requirements, such as ρ_1-ρ_2 privacy [6][25]. We end this section with a comparison between *output perturbation* and *data perturbation* in the current work. In output perturbation, such as 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 Sections 1.1 and 2, 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.

3.2 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. $D(x_1, \cdots, x_n)$ denotes the subset of records in D that match x_i on every A_i, and $D^*(x_1, \cdots, x_n)$ denotes the corresponding subset for D^*. If, for $1 \leq i \leq n$, x_i is a non-wildcard, $D(x_1, \cdots, x_n)$ is a *personal group*. If at least one x_i is a wildcard, $D(x_1, \cdots, x_n)$ is an *aggregate group*. For example, for $NA = \{Gender, Job\}$, $D(male, eng)$ is a personal group and $D(-, eng)$ 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 $D(male, eng)$ because all records in the personal group are exactly identical on NA. Without confusion, we call both $D(x_1, \cdots, x_n)$ and $D(x_1, \cdots, x_n)^*$ a personal or aggregate group as there is an one-to-one correspondence between the two.

In Example 2, we argued that the personal group $D^*(male, eng)$ should be used to quantify the risk of inferring the disease breast cancer for the male engineer Bob, instead of the aggregate groups $D^*(-, eng)$, $D^*(male, -)$, or $D^*(-, -)$. 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. In Section 3.4 we will consider the case where further information is available to the adversary and using additional records not belonging to Bob may help the adversary. An analogy is short-listing the suspect of a robbery: if the eyewitness has reported that the suspect was a male blonde caucasians (i.e., the public attributes), it makes sense to focus on the subset of male blonde caucasians in the police database, instead of examining all male caucasians records. The above observation motivates the following two types of reconstruction.

DEFINITION 1. *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. These different roles of reconstruction are stated in the next principle.

DEFINITION 2 (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. In fact, any statistical database with any non-trivial utility incurs some amount of disclosure [10]. Our principle assures that only a limited amount of disclosure is incurred by enabling non-trivial utility.

3.3 Reconstruction Privacy

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^*. Note that F' is a random variable because 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 3 (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^*. 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.)* □

Note that reconstruction privacy is a property of the perturbation matrix $\mathbb{P}$, not a property of a particular instance of 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. $F' - f$ 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 [10][11]. 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 $\Pr\left[\frac{F'-f}{f} > \lambda\right]$.

DEFINITION 4 (ENFORCING PRIVACY). *Given a database 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 D^* while preserving aggregate reconstruction as much as possible.* □

By leaving the retention probability p as an input parameter to our problem, other privacy criteria, such as ρ_1-ρ_2 privacy, can be enforced through a proper choice of p. In this sense, reconstruction privacy can be considered as an *additional* protection on top of other privacy criteria.

3.4 Generalized Personal Groups

Consider two personal groups $g^* = D(male, eng)$ and $g'^* = D(female, eng)$. Our reconstruction privacy limits the reconstruction for each personal group, but does not limit the reconstruction for the combined $g^* \cup g'^*$, i.e., the aggregate group $D^*(-, eng)$, 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 SA values. For example, suppose that $FavoriteColor$ is another public attribute and that the favorite color 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 two domain values x_i and x_i' (e.g., $male$ and $female$) of A_i have the same impact on SA, we will merge x_i and x_i' 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 combine several sub-populations that follow a different distribution on SA.

So the question is how to identify the values of A_i that have the same impact on SA. To this end, the well studied χ^2-squared test that tells if two data sets are from different distributions can be used. For two domain values x_i and x_i' of A_i, let o_{ij} (resp. o_{ij}') be the number of records in D satisfying $A_i = x_i$ (resp. $A_i = x_i'$) and $SA = sa_j$, $1 \leq j \leq m$. Let $O_i = [o_{i1}, \cdots, o_{im}]$ and $O_i' = [o_{i1}', \cdots, o_{im}']$, which represents 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 proves that the data sets are from different distributions.

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 [26]:

$$\chi^2 = \sum_{j=1}^{m} \frac{\left(\sqrt{|O_i'|/|O_i|}o_{ij} - \sqrt{|O_i|/|O_i'|}o_{ij}'\right)^2}{o_{ij} + o_{ij}'} \quad (4)$$

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) 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) 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 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 above null hypothesis. Finally, for each connected component of the graph, we merge all the values in the component into a single generalized value. This method ensures that any two values x_i and x_i' from different components have a different impact on SA.

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 3.2 are based on such generalized domain values.

4. TESTING PRIVACY

An immediate question is how to test (λ, δ)-reconstruction-privacy. From Definition 3, 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 follow-

Table 3: Notations

Symbols	Meaning
D, 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 (3)
p	the retention probability
(λ, δ)	privacy parameters

ing discussion refers to a subset S of D and the corresponding subset S^* of 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 3 summarizes the notations used in this paper.

4.1 Computing F'

First of all, let us examine the computation of F'. Example 2 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 (3). 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 1 (THEOREM 2, [16]). *$\mathbb{P}^{-1} \cdot \frac{\overleftarrow{O^*}}{|S|}$ 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 2. *For any subset S of 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 1, $\overleftarrow{F'} = \mathbb{P}^{-1} \cdot \frac{\overleftarrow{O^*}}{|S|}$. Let $\overleftarrow{\frac{1-p}{m}}$ denote the column-vector of the constant $\frac{1-p}{m}$ of length m. We have

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

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 2(iii) implies that F' is an unbiased estimator of f. Lemma 2(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 3).

4.2 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 2(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 2 (BOUND CONVERSION). *Consider any subset S of 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 2(ii)), $O^* = |S|(F'p + (1-p)/m)$, and from Lemma 2(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 2, 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. Markov's inequality and Chebyshev's inequality are some early upper bounds, for example. The Chernoff bound, due to [27], 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 3 (CHERNOFF BOUNDS, [27]). *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 $\omega \in (0, \infty)$,*

$$\Pr\left[\frac{X-\mu}{\mu} > \omega\right] < U(\omega, \mu) = exp(-\frac{\omega^2\mu}{2+\omega}) \quad (5)$$

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

$$\Pr\left[\frac{X-\mu}{\mu} < -\omega\right] < L(\omega, \mu) = exp(-\frac{\omega^2\mu}{2}).\square \quad (6)$$

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 2). To obtain the upper bounds for F', we instantiate the upper bounds U and L for O^* in Equations (5) and (6) into Theorem 2. This gives the next corollary.

COROLLARY 3 (UPPER BOUNDS FOR F'). *Let* $\omega = \frac{\lambda|S|pf}{\mu}$ *and* $\mu = |S|(fp + (1-p)/m)$. *For* $\omega \in (0, \infty)$,

$$\Pr\left[\frac{F'-f}{f} > \lambda\right] < U(\omega, \mu) = exp(-\frac{\omega^2\mu}{2+\omega}) \tag{7}$$

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

$$\Pr\left[\frac{F'-f}{f} < -\lambda\right] < L(\omega, \mu) = exp(-\frac{\omega^2\mu}{2}).\square \tag{8}$$

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 3 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 not. 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 2 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 3 are the best upper bounds for F'. This assumption is not a real restriction because Theorem 2 allows us to "plug in" any better bound for O^* for a better bound for F'. If 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 3.1, 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.

4.3 Testing

With the upper bounds L and U in Corollary 3, 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 expressions for ω and μ in Corollary 3 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 4. *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*

$$|g| \leq \frac{-2(fp + (1-p)/m)\ln\delta}{(\lambda pf)^2}\square \tag{9}$$

Given D, the personal groups g and the frequencies f for all SA values in g can be found by sorting the records in D in the order of all attributes in NA followed by SA. Therefore, all the quantities in Equation (9) 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.

5. ENFORCING PRIVACY

If reconstruction privacy is not satisfied, we can restore reconstruction privacy by satisfying the condition in Equation (9) for every SA value and every personal group. Observe that the right-hand side of Equation (9) decreases as f increases. Therefore, a personal group g^* satisfies reconstruction privacy if and only if $|g| \leq s_g$, where

$$s_g = \frac{-2(fp + (1-p)/m)\ln\delta}{(\lambda pf)^2} \tag{10}$$

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 (10) unchanged after sampling). Preserving frequencies also helps minimize the distortion to data distribution. After perturbing 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.

Sampling-Perturbing-Scaling (SPS) algorithm. The input is a database 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 D^* that satisfies (λ, δ)-reconstruction-privacy. For each personal group g in D, this algorithm computes a modified version g_2^* of g^*, then outputs $D_2^* = \bigcup g_2^*$. In a preprocessing step, we sort the records in D by the attributes in NA and followed by SA. The result is a collection of personal groups g together with the frequencies f of every SA value in g.

For each personal group g in D, compute s_g as in Equation (10), if $|g| \leq s_g$, g already satisfies the maximum group size constraint, let $g_2^* = g^*$. We assume $|g| > s_g$. In the following, g_2^* is produced in three steps: *Sampling*, *Perturbing*, and *Scaling*, described below. Let $\tau = s_g/|g|$, called the *sampling rate*.

1. $Sampling(g, s_g)$ takes a sample of the records in g while preserving the frequency of each SA value. For each SA value sa occurring in g, let g_{sa} denote the subset of the records

in g that have sa. Note that all records in g_{sa} are identical. We pick any $\lfloor |g_{sa}|\tau \rfloor$ records from g_{sa} and pick one additional record from g_{sa} with the probability $|g_{sa}|\tau - \lfloor |g_{sa}|\tau \rfloor$. Let g_1 be the set of the picked records. Return g_1.

2. $Perturbing(g_1, p, m)$ perturbs the SA values of the records in g_1 with the retention probability p, as in the Uniform Perturbation described in Section 3.1. Return g_1^*.

3. $Scaling(g_1^*, |g|)$ scales up g_1^* to the original size $|g|$ while preserving the frequency of each SA value. Let $\tau' = |g|/|g_1^*|$. For each record r^* in g_1^*, let g_2^* contain $\lfloor \tau' \rfloor$ duplicates of r^* and one additional duplicate of r^* with the probability $\tau' - \lfloor \tau' \rfloor$. Return g_2^*.

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 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.

Complexity analysis. Let $|D|$ denote the number of records in D. The sorting step takes $|D|log|D|$ 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 $(|D|log|D| + |D|)$ time.

5.1 Analysis

We prove two claims about the output $D_2^* = \cup g_2^*$. The first claim is on privacy guarantee: each g_2^* in 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 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 2(ii)) and Fact 2.

- Fact 4: $E[O_{g_2}^*] \simeq E[O_g^*]$. From Lemma 2(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 2(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 4 (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 4, 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 5 (UTILITY). *Let S be a set of records for one or more personal groups in D, S^* be the corresponding set for D^*, and S_2^* be the corresponding set for 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 2(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 2(iii), $E[F'] \simeq f$, thus, $E[F'_{S_2}] \simeq f$. □

Intuitively, Theorem 5 says that the estimate reconstructed using the corresponding records in D_2^* is an unbiased estimator of the actual frequency.

6. EXPERIMENTAL STUDIES

We evaluate two claims. 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.

6.1 Experimental Setup

We implemented the proposed SPS algorithm as described in Section 5 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 [14]. 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 *CENSUS* data previously used in [28][22]. 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 CENSUS 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 CENSUS 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.

As discussed in Section 3.4, the values for public attributes with the same impact on SA have to be aggregated before generating personal groups. The aggregation affects data sets to some extent. Tables 4 and 5 show the impacts on the domain size of each public attribute, the total number of personal groups (e.g., $|G|$), and the averaged personal groups size (e.g., $|D|/|G|$ with $|D|$ as the total number of records) of ADULT and CENSUS $300K$. In the rest of this section, we use the generalized values of public attributes.

Table 4: NA Aggregation Impact on ADULT

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

Table 5: NA Aggregation Impact on CENSUS $300K$

	Domain Size of NA					$\|G\|$	$\|D\|/\|G\|$
	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:

$$SELECT\ COUNT\ (*)\ FROM\ D$$
$$WHERE\ A_1 = a_1 \wedge \cdots \wedge A_d = a_d \wedge SA = sa_i \quad (11)$$

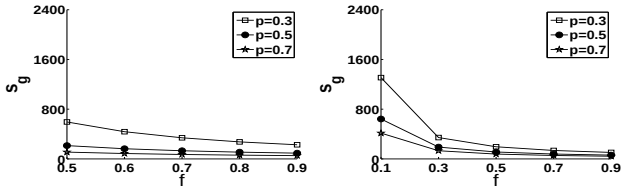
where $A_j \in NA$, $a_j \in dom(A_j)$, and $sa_i \in dom(SA)$. The answer to the query, ans, is the number of records in D satisfying the condition in the WHERE clause. Such answers can be used to learn statistical relationships between the attributes in NA and SA. Given the perturbed data D^*, ans is approximated by $est = |S^*| * F'$, where S^* is the set of records in 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 2(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 1D or 2D marginals with a size above a sanity bound [29]. We generated a pool of 5,000 count queries with the query dimensionality d in $\{1, 2, 3\}$ and with the selectivity $ans/|D| \geq 0.1\%$. 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 replaced the NA value with aggregated values and then added it to the pool. Recall that we aggregated NA values based on their impact on SA as in Section 3.3. The query pool simulates the set of possible queries generated from real life, therefore, the original NA value (before aggregation) is used to generate the query pool. Since we protect reconstruction privacy on aggregated personal groups we evaluate relative error on these aggregated personal groups as well. We report the average of relative error over all queries in this pool. In addition, since D^* is randomly generated in each run, we reported the average of 10 runs to avoid the bias of a particular run.

Table 6: 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

The uniform perturbation, denoted by UP, as described in Section 3.1 has been used as a privacy mechanism in [25][16][6]. But these privacy mechanisms do not address the disclosure of personal reconstruction. Our method addresses this disclosure by applying UP to sampled data. So our evaluation has two parts. First, we evaluate how often reconstruction privacy is violated by the perturbed data 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 D_2^* produced by SPS, compared to the relative error of queries answered using D^* produced by UP. The same retention probability p is used for both UP and SPS. Table 6 shows the settings of p, λ, and δ with the default settings in boldface.



(a) ADULT (b) CENSUS

Figure 1: Maximum group size s_g vs. maximum frequency f

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 5, where s_g is the maximum threshold on the group size defined as

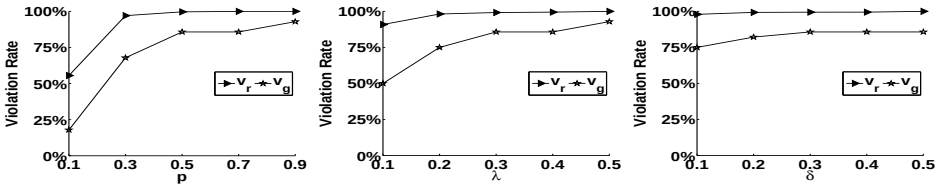
$$s_g = \frac{-2(fp + (1-p)/m)\ln\delta}{(\lambda pf)^2} \quad (12)$$

f is the maximum frequency of any SA value occurring in g. Figure 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 CENSUS. 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 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 1 and Equation (12) 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 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.

6.2 ADULT Data Set

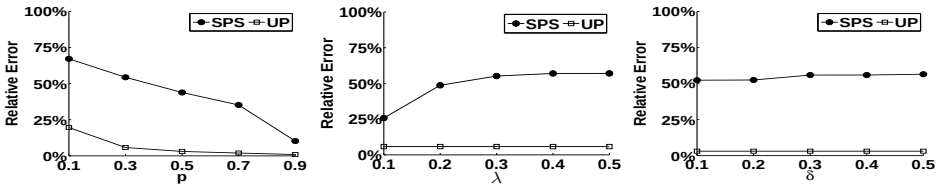
Violation. Figure 2 shows the extent to which reconstruction privacy is violated on the perturbed ADULT data set D^* produced by UP. This *extent* is measured at two levels. v_g represents the percentage of groups that violate reconstruction privacy. v_r represents the percentage of records contained in a violating personal group, 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



(a) vs. p (b) vs. λ (c) vs. δ

Figure 2: ADULT: Privacy Violation



(a) vs. p (b) vs. λ (c) vs. δ

Figure 3: ADULT: Relative Error

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 1).

Cost. Figure 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 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 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 CENSUS data set. Choosing a small p helps eliminate violation, but also quickly increases the relative error for both UP and SPS (Figures 2a and 3a). Indeed, a too small p makes the perturbed data become nearly pure noises. This study confirms our discussion at the beginning of Section 5 that the approach of reducing p does not preserve utility.

6.3 CENSUS Data Set

Violation. CENSUS 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 shows the extent to which reconstruction privacy is violated. The default data size is 300K when $|D|$ 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 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 CENSUS would violate reconstruction privacy, which explains the much smaller v_g and also confirms our claim on Figure 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.

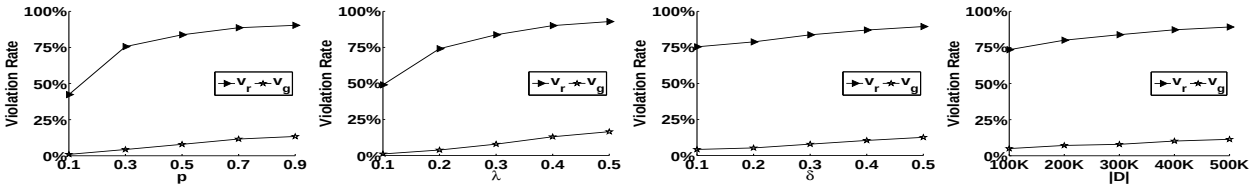
Cost. Figure 5 compares the relative error of UP and SPS. A big difference from the ADULT data set is that there is less increase in the relative error (e.g., less than 10% for most of settings) for SPS compared to the relative error for UP across all settings of parameters. This is a consequence of 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 $|D|$ causes more violations of reconstruction privacy (Figure 4d), it actually decreases the relative error for SPS (Figure 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.

In 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.

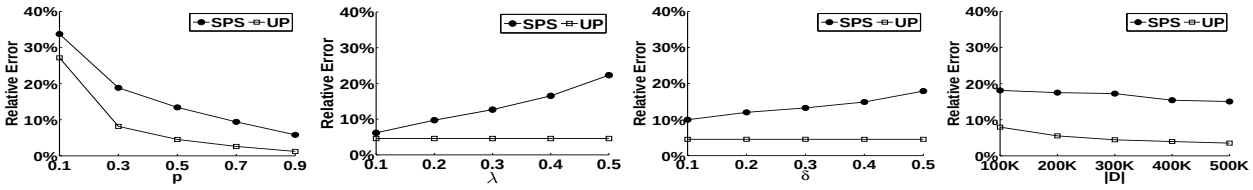
7. CONCLUSION

Differential privacy has become a popular privacy definition for sharing statistical information thanks to good utility. However, this good utility comes with the cost of disclosures through non-independent reasoning. In this work, we presented a data perturbation approach to prevent sensitive non-independent reasoning 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 use record sampling to reduce the random trials in data perturbation, which mostly affects non-independent reasoning specific to an individual while having only a limited effect on statistical learning.



(a) vs. p (b) vs. λ (c) vs. δ (d) vs. $|D|$

Figure 4: CENSUS: Privacy Violation



(a) vs. p (b) vs. λ (c) vs. δ (d) vs. $|D|$

Figure 5: CENSUS: Relative Error

8. ACKNOWLEDGEMENTS

Ke Wang's work is partially supported by a Discovery Grant of the Natural Sciences and Engineering Research Council, Canada. Raymond Chi-Wing Wong's is partially supported by the grant F-SGRF14EG34. Philip S. Yu's is partially supported by US NSF through grants CNS-1115234 and OISE-1129076.

9. REFERENCES

[1] R. Adam and J. Worthmann. Security-control methods for statistical databases: A comparative study. *ACM Comput. Surv.*, 21(4):515–556, December 1989.

[2] B. Fung, K. Wang, R. Chen, and P. Yu. Privacy-preserving data publishing: a survey of recent developments. *ACM Comput. Surv.*, 42(4):14:1–14:53, June 2010.

[3] B. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala. Privacy-preserving data publishing. *Found. Trends databases*, 2(1-2):1–167, January 2009.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: privacy beyound k-anonymity. In *ICDE*, 2006.

[5] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.

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

[7] J. Cao and P. Karras. Publishing microdata with a robust privacy guarantee. In *VLDB*, 2012.

[8] A. Fu, K. Wang, R. Wong, J. Wang, and M. Jiang. Small sum privacy and large sum utility in data publishing. *Journal of Biomedical Informatics*, 50:20–31, 2014.

[9] Y. Tao, X. Xiao, J. Li, and D. Zhang. On anti-corruption privacy preserving publication. In *ICDE*, 2008.

[10] C. Dwork. Differential privacy. In *ICALP*, 2006.

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

[12] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD*, 2011.

[13] G. Cormode. Personal privacy vs population privacy: learning to attack anonymization. In *SIGKDD*, 2011.

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

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

[16] R. Agrawal, R. Srikant, and D. Thomas. Privacy preserving olap. In *SIGMOD*, 2005.

[17] M. NarasimhaRao, J. VenuGopalkrisna, R. Murthy, and C. Ramesh. Closeness: Privacy measure for data publishing using multiple sensitive attributes. 2(2):278–284, 2012.

[18] R. C. Elandt-Johnson and N.L. Johnson. *Survival models and data analysis*. John Wiley & Sons NY, 1980.

[19] A. Stuart and K. Ord. *Kendall's advanced theory of statistics*, volume 1. Arnold, London, 6 edition, 1998.

[20] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: privacy via distributed noise generation. In *EUROCRYPT*, volume 4004, pages 486–503, 2006.

[21] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear counting queries under differential privacy. In *PODS*, 2010.

[22] R. Chaytor and K. Wang. Small domain randomization: same privacy, more utility. In *VLDB*, 2010.

[23] V. Rastogi, S. Hong, and D. Suciu. The boundary between privacy and utility in data publishing. In *VLDB*, 2007.

[24] T. Li and N. Li. Injector: mining background knowledge for data anonymization. In *ICDE*, 2008.

[25] S. Agrawal and J. Haritsa. A framework for high-accuracy privacy preserving mining. In *ICDE*, 2005.

[26] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 1988.

[27] H. 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.

[28] X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In *VLDB*, 2006.

[29] X. Xiao, G. Bender, M. Hay, and J. Gehrke. ireduct: Differential privacy with reduced relative errors. In *SIGMOD*, 2011.