Sensitive Disclosures under Differential Privacy Guarantees

Chao Han
Simon Fraser University
Burnaby, Canada
hanchaoh@cs.sfu.ca

Ke Wang
Simon Fraser University
Burnaby, Canada
wangk@cs.sfu.ca

***Abstract*—*Non-independent reasoning* (NIR) refers to learning the information of one record from other records, under the assumption that these records share the same underlying distribution. Accurate NIR could disclose private information of an individual. An important assumption made by differential privacy is that NIR is considered to be non-violation of privacy. In this work, 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 means under NIR by randomized query answers. We then present a formal analysis on such disclosures by differentially private query answers. Our analysis on real life datasets 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 cannot be easily solved 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.**

***Keywords*-Differential Privacy; Data Privacy;**

I. Introduction

A. Motivation

The burst of networks has brought us to an age of big data. The growing number of people, devices, and sensors connected by the network are main data sources. While big data provides enormous value and benefits for the global economy and growth, it also brings significant privacy concerns. Privacy and security have been identified as a main challenge of big data [2]. 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 before her father knew [1]. Anonymizing data is not enough to maintain user privacy. AOL released anonymized search logs for academic purposes, but searchers were easily re-identified by their queries. There is a similar problem with Netflix's anonymized movie rating data.

Numerous methods have been proposed to protect data privacy, see [11][3][4] for surveys. In the past several years, differential privacy [10] emerges as the "gold standard" for data privacy definitions. Differential privacy provides statistically indistinguishable results with respect to the participation of a single individual in the dataset. This is achieved by adding noises to the results of statistical queries. A popularized claim is that differential privacy provides a good utility for data analytics while providing strong privacy guarantees regardless of the adversary's background knowledge. In particular, it is claimed that even if the adversary knows all but one record in the data set, he is not able to learn the sensitive information about the unknown record.

In this paper, 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 dataset, 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.

Sensitive disclosures under differentially private query answers have been examined by recent works [13][9][14]. The author of [13] pointed out that in the presence of record correlation sensitive information of an individual could be learnt from query answers even though a notion of differential privacy is satisfied. Intuitively, in the presence of record correlation, the absence of a record is no longer sufficient for hiding the sensitive information of a record because such information could be learnt from correlated records. The disclosures identified in our work do not assume correlation of records. The reference [9] demonstrated that a Bayes classifier could be built using noisy query answers published by a differential privacy mechanism. We identify disclosures that are based on the results of two queries, instead of multiple queries as required by a Bayes classifier. Involving fewer queries certainly increases the chance of disclosures. The reference [14] demonstrated some disclosures under differential privacy. Our work provides a theoretical explanation for such disclosures and shows an arbitrary accuracy provided that true query answers are sufficiently large (i.e., Theorem 1).

B. *Contributions*

The contributions of this work are as follows.

Firstly, we formalize a notion of disclosures in terms of the probability of a small error in learning sensitive information through *non-independent reasoning* (NIR). NIR refers to learning the information of one record from other records under the assumption that these records share the same underlying distribution.

Secondly, we present a formal analysis on disclosures through query answers that satisfy typical settings of differential privacy. Specifically, we model the probability of the error of learning sensitive information through NIR by a ratio distribution of two Laplace variables. To our knowledge, this is the first study on the probability of ratio distribution for two Laplace variables. This modeling yields an efficient way of determining the disclosures of query answers produced by the differential privacy mechanism.

Thirdly, we study the above type of disclosures on several real life datasets while a notion of differential privacy is satisfied, and the impact of eliminating such disclosures on data utility. 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 vioation, differential privacy does not provide both privacy and utility.

C. *Organization*

The rest of the paper is organized as follows. Section II formalizes the notion of disclosures of noisy query answers through NIR. Section III presents a formal analysis on such disclosures under the differential privacy mechanism with the Laplace noise distribution. Section IV studies the interplay between information disclosure and data utility for query answers published by the differential privacy mechanism on real life data sets. Finally, Section V concludes the paper.

II. 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 III.

A. *Attacks*

We consider a micro dataset $\mathcal{D}$ containing a 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. 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 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. 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, use the combined set of the records from multiple personal groups that match partially 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 $|g_{sa}|$ and $|g|$ through the following two queries:

$$\begin{aligned}
&Q_1: SELECT\ COUNT(*)\ FROM\ \mathcal{D}\ WHERE \\
&NA = t[NA] \\
&Q_2: SELECT\ COUNT(*)\ FROM\ \mathcal{D}\ WHERE \\
&NA = t[NA]\ AND\ SA = sa.
\end{aligned} \tag{1}$$

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 and Y denote the noisy answers for Q_1 and Q_2. 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.

$\frac{Y}{X}$ 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. In this work, we consider the true confidence of sa in the entire dataset $\mathcal{D}$ 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

Table I: Notations

Notation	Explanation
t	a target individual
sa	a domain value of *SA*
g	a personal group
g_{sa}	the subset in g with sa on *SA*
$\phi = \|g\|, \theta = \|g_{sa}\|$	true answers for query Q_1 and Q_2 in (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

of individuals, one of them being t, that are more frequently associated with sa than in general. In this sense, we say that a disclosure occurs to t.

B. 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 1. (CP_τ, Closeness Probability) *For $\tau > 0$, the closeness probability for a* SA *value* sa *in g is*

$$CP_\tau = \Pr\left[\left|\frac{\frac{\theta}{\phi} - \frac{Y}{X}}{\frac{\theta}{\phi}}\right| \leq \tau\right]. \qquad (2)$$

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. A larger CP_τ and a smaller τ imply that the noisy confidence is a more accurate estimate of the true confidence.

For (B), we introduce the variable $\mathcal{J} = \frac{\theta/\phi}{f}$ to calibrate the distance from the true confidence to the prior knowledge (i.e., $f = \frac{|\mathcal{D}_{sa}|}{|\mathcal{D}|}$).

Definition 2. (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 [6], other than considering the difference between the noisy confidence within a personal group and the global frequency (as in β-likeness), i.e., condition (2), we also consider the difference of confidence within a personal group before and after adding the noise, i.e., condition (1). Table I outlines the notation used in this paper.

III. Disclosure of Differential Privacy

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

A. Differential Privacy

We say that two datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ are *neighboring* if they differ in at most one tuple.

Definition 3. (ϵ-Differential Privacy) [10]. *A randomization mechanism $\mathcal{A}$ satisfies ϵ-differential privacy if for any output O of $\mathcal{A}$ and any neighboring datasets $\mathcal{D}_1$ and $\mathcal{D}_2$,*

$$\Pr[\mathcal{A}(\mathcal{D}_1) = O] \leq exp(\epsilon) \times \Pr[\mathcal{A}(\mathcal{D}_2) = O]. \qquad (3)$$

Typically ϵ is a small value close to zero and $exp(\epsilon)$ is a value close to 1. The inequality in (3) implies that any neighboring datasets $\mathcal{D}_1$ 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 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 (3) are closer. In the literature [9] [17], ϵ is typically chosen from the range of $[0.01, 2]$. In this paper, we use this range.

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 [5]: "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 Alice's HIV status is *accurately* inferred from differentially private query answers and if Alice cares about keeping this status private, Alice's privacy is breached, even if Alice's record is not in the data. The focus of this work is on this type of disclosures enabled by differentially private query answers; we want to know how likely and how accurately noisy answers can be used to learn the sensitive information about an individual, even if these noisy answers satisfy a differential privacy guarantee.

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, defined below. ϵ-differential privacy can be achieved by adding the Laplace noise with the scale factor $b = \Delta/\epsilon$, denoted $Lap(b)$, to the output of each query [10]. $Lap(b)$ has the zero mean and the probability density function $f(x) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$. While there are other variations of differential privacy and randomization mechanism [10], we adopt $Lap(b)$ as the randomization mechanism for count queries because they are most studied in the literature.

Definition 4. (Sensitivity) [10]. *The sensitivity of a sequence Q of queries, denoted as Δ, is defined as:*

$$\Delta = \max_{\mathcal{D}_1, \mathcal{D}_2} ||Q(\mathcal{D}_1) - Q(\mathcal{D}_2)||_1, \tag{4}$$

for neighboring datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, where $||.||_1$ is the 1-norm of a vector.

B. Computing CP_τ

With the Laplace noise distribution $Lap(b)$, we can now derive for CP_τ in Definition 1. 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:

$$f_X(x) = \frac{1}{2b} \exp\left(-\frac{|x-\phi|}{b}\right) \tag{5}$$

$$f_Y(y) = \frac{1}{2b} \exp\left(-\frac{|y-\theta|}{b}\right). \tag{6}$$

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

It is easy to see that CP_τ is equal to

$$\Pr\left[\left|\frac{\theta}{\phi} - \frac{Y}{X}\right| \leq \tau'\right], \tag{7}$$

where $\tau' = \tau \times \frac{\theta}{\phi} > 0$. To compute the probability in (7), we define the following cumulative function for the ratio of two Laplace variables, $Z = \frac{Y}{X}$,

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

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 1. *Assume $z \neq 0$ and $z \neq \pm 1$.*
For $z < 0$,

$$\begin{aligned} F_Z^1(z) = &\left[\frac{z^2}{2(1-z^2)}\right]\left[\exp\left(\frac{\theta - z\phi}{zb}\right)\right] \\ &- \frac{1}{2(z+1)}\left[\exp\left(\frac{-(\theta+\phi)}{b}\right)\right] \\ &+ \frac{1}{2}\exp\left(-\frac{\phi}{b}\right) - \frac{1}{2(z^2-1)}\left[\exp\left(\frac{z\phi-\theta}{b}\right)\right]; \end{aligned} \tag{8}$$

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

$$\begin{aligned} F_Z^2(z) = &\left[\frac{z^2}{2(z^2-1)}\right]\left[exp\left(\frac{z\phi - \theta}{zb}\right)\right] \\ &- \frac{1}{2(z+1)}\left[exp\left(\frac{-(\theta+\phi)}{b}\right)\right] \\ &+ \frac{1}{2}exp\left(-\frac{\phi}{b}\right) - \frac{1}{2(z^2-1)}\left[exp\left(\frac{z\phi-\theta}{b}\right)\right]; \end{aligned} \tag{9}$$

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

$$\begin{aligned} F_Z^3(z) = &\frac{z^2}{2(1-z^2)}\left[exp\left(\frac{\theta - z\phi}{zb}\right)\right] \\ &- \frac{1}{2(z+1)}\left[exp\left(\frac{-(\theta+\phi)}{b}\right)\right] \\ &+ 1 + \frac{1}{2}exp\left(-\frac{\phi}{b}\right) - \frac{1}{2(1-z^2)}\left[exp\left(\frac{\theta-z\phi}{b}\right)\right]. \end{aligned} \tag{10}$$

□

The proof of Lemma 1 can be found at [12].

Theorem 1. *Assume $\theta > 0$, $\phi > 0$, $\tau' > 0$, $\frac{\theta}{\phi} \pm \tau' \neq 0$ and $\frac{\theta}{\phi} \pm \tau' \neq \pm 1$.*

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

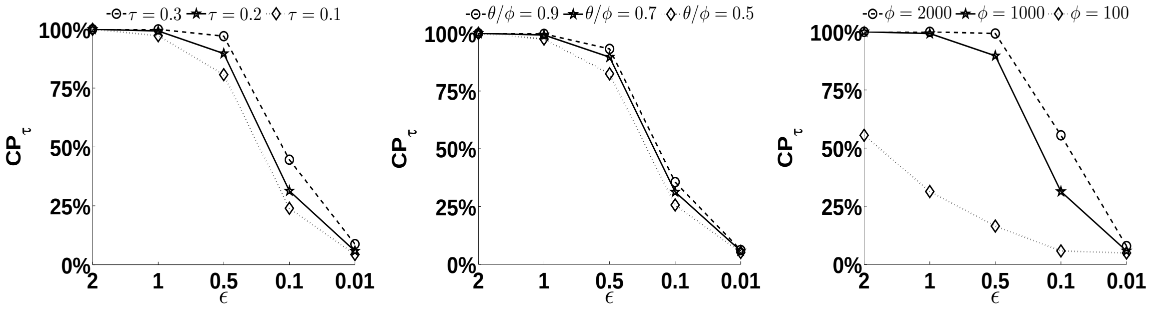
$$\begin{cases} F_Z^3\left(\frac{\theta}{\phi} + \tau'\right) - F_Z^1\left(\frac{\theta}{\phi} - \tau'\right) & \textit{if } \frac{\theta}{\phi} - \tau' < 0 \\ F_Z^3\left(\frac{\theta}{\phi} + \tau'\right) - F_Z^2\left(\frac{\theta}{\phi} - \tau'\right) & \textit{if } \frac{\theta}{\phi} - \tau' > 0 \end{cases} \tag{11}$$

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

Proof: 1) Let $Z = \frac{Y}{X}$. Rewrite $\left|\frac{\theta}{\phi} - Z\right| \leq \tau'$ into $\frac{\theta}{\phi} - \tau' \leq Z \leq \frac{\theta}{\phi} + \tau'$. $Pr\left[\left|\frac{\theta}{\phi} - Z\right| \leq \tau'\right] = F_Z\left(\frac{\theta}{\phi} + \tau'\right) - F_Z\left(\frac{\theta}{\phi} - \tau'\right)$. Because $\tau' > 0$, $F_Z\left(\frac{\theta}{\phi} + \tau'\right)$ is $F_Z^3\left(\frac{\theta}{\phi} + \tau'\right)$. $F_Z\left(\frac{\theta}{\phi} - \tau'\right)$ is $F_Z^1\left(\frac{\theta}{\phi} - \tau'\right)$ when $\frac{\theta}{\phi} - \tau' < 0$, and $F_Z\left(\frac{\theta}{\phi} - \tau'\right)$ is $F_Z^2\left(\frac{\theta}{\phi} - \tau'\right)$ when $\frac{\theta}{\phi} - \tau' > 0$. Note that $\frac{\theta}{\phi} + \tau' \neq 0$ and $\frac{\theta}{\phi} - \tau' \neq \pm 1$ are required because $z \neq 0$ and $z \neq \pm 1$ are required in Lemma 1. This shows Theorem 1, 1).

2) As ϕ approaches to ∞, all exp terms in F_Z^1, F_Z^2 and F_Z^3 in (11) 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 1 provides the theoretical basis for why the noisy confidence is a good estimate of the true confidence: when ϕ is large the adversary is able to use the noisy confidence to estimate the true confidence more accurately. It also provides a way to compute the probability in (7), thus, CP_τ. By replacing τ' with $\tau \times \frac{\theta}{\phi}$ in (11), we have the following computation of CP_τ.



(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 1: CP_τ ($\Delta = 25$)

Table II: 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**

Corollary 1.

$$CP_\tau = \begin{cases} F_Z^3\left((1+\tau)\frac{\theta}{\phi}\right) - F_Z^1\left((1-\tau)\frac{\theta}{\phi}\right) & \textit{if } \tau > 1 \\ F_Z^3\left((1+\tau)\frac{\theta}{\phi}\right) - F_Z^2\left((1-\tau)\frac{\theta}{\phi}\right) & \textit{if } 0 < \tau < 1 \end{cases} \tag{12}$$

The value of CP_τ depends on τ, b, ϕ, and θ/ϕ. Let us evaluate CP_τ for typical values of these parameters. 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 datasets in our experiments Section IV. The settings for the other parameters can be found in Table II with the default settings in bold face. The range of ϵ has been discussed in Section III.

Fig. 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., 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 1.

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 dataset, 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 for data analysis because both noisy confidence and utility are built 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.

IV. Reality Check

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. We also study the negative impact on utility by simply using larger noises to remove such disclosures.

A. Experimental Setup

All experiments are implemented in Python and ran on an Intel Xeon(R) E5630 CPU 2.53GHZ PC with 12GB of RAM. All three dataset *SALARY*, *EDU* and *OCC* we used are extracted from the US census data [1] about personal information of American adults. *SALARY* was used in [16][7], and both *EDU* and *OCC* were used in [16][8]. All datasets have the same number of records (i.e., 500k). Table III shows the information of attributes with domain size in brackets, and the *SA* is marked in bold.

B. Publishing scenarios

Many typical data analysis tasks make use of low dimensional marginals, where each marginal corresponds to some subset of attributes with each row being the count for one combination of the values on those attributes. Therefore, a marginal consists of the answers for all the count queries over those attributes. For a given dataset, we considered publishing all $2D$ (two dimensional) marginals on the attributes in *NA*, and for each, publishing the corresponding marginal expanded with the extra attribute *SA*, denoted by *2D+SA*.

[1] Downloadable from http://www.ipums.org.

Table III: Attributes in Datasets

Dataset	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)**

Table IV: Dataset and Sample Marginals in Example 1

(a) A dataset $\mathcal{D}$

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 $2D$ marginal

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

(c) A *2D+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

There are $\binom{|NA|}{2}$ $2D$ marginals and the same number of *2D+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 1. *Assume that a dataset $\mathcal{D}$ has three* NA *attributes:* Age, Gender *and* Occupation, *and* SA, Disease. *Tables IVa, IVb, and IVc illustrate $\mathcal{D}$, a $2D$ marginal on $\{Gender, Occupation\}$, and the corresponding* 2D+SA *marginal on $\{Gender, Occupation, Disease\}$, respectively. $|NA|$ in this dataset is 3, thus, there are $\binom{3}{2}$ (i.e., three) $2D$ marginals and three* 2D+SA *marginals in total. The three $2D$ marginals are $\{Age, Gender\}$, $\{Age, Occupation\}$ and $\{Gender, Occupation\}$, and the three* 2D+SA *marginals are $\{Age, Gender, Disease\}$, $\{Age, Occupation, Disease\}$ and $\{Gender, Occupation, Disease\}$. Here the counts in the marginals are based on the raw dataset 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 $2D$ marginals would not be the aggregated results of corresponding* 2D+SA *marginals.*

In the classic mechanism in [10], 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

$$\Delta = 2 \cdot |\mathcal{M}| = 4 \cdot \binom{|NA|}{2}. \tag{13}$$

This is because changing one record affects at most 2 counts in a marginal and there are $|\mathcal{M}|$ marginals. Note that if only deleting and adding a record are allowed in the dataset then changing one record affects at most 1 count in a marginal, in which case our result could be simply adjusted by 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 [17][16][15] 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.

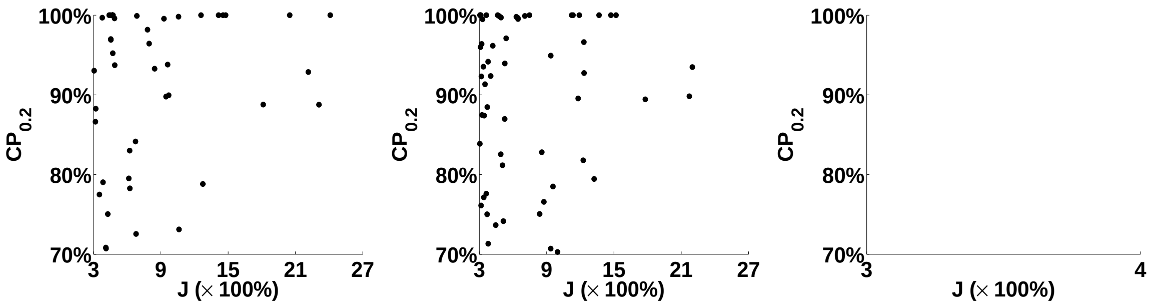
C. Disclosures

In a personal group, a *SA* value sa is *disclosed* if the following conditions are satisfied (Definition 2): $CP_\tau \geq \mathcal{K}_{CP}$, that is, the probability that the noisy confidence $\frac{Y}{X}$ has no more than the relative error of τ is at least $\mathcal{K}_{CP}$, and $\mathcal{J} \geq \mathcal{K}_{\mathcal{J}}$, that is, the true confidence in the personal group is at least $\mathcal{K}_{\mathcal{J}}$ times higher than the confidence in the entire dataset.

Fig. 2 shows the disclosures found in the three datasets for the privacy setting $\epsilon = 0.5$, $\tau = 0.2$, $\mathcal{K}_{CP} = 0.7$, and $\mathcal{K}_{\mathcal{J}} = 3$. Each point 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 participant belongs to a personal group, identified by her/his known *NA*, and in this group, it is 20 times more likely to have some *SA* value than in the entire dataset 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., $CP_{0.2} \approx 100\%$).

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. This is because *SALARY* has more *NA* attributes, thus, a larger $|\mathcal{M}|$ and a larger sensitivity Δ according to (13), which leads to a larger scale factor b for a given privacy setting ϵ.

Fig. 3 shows the number of disclosures for the three datasets 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) EDU

(b) OCC

(c) SALARY

Figure 2: Disclosures in terms of CP_τ and J ($\tau = 0.2, \mathcal{K}_{CP} = 0.7, \mathcal{K}_{\mathcal{J}} = 3, \epsilon = 0.5$)

a personal group g is counted as the number of disclosed *SA* values in g, and the total number of disclosures of the dataset is counted as the sum of the disclosures in all personal groups g. 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, which is consistent with the findings in Fig. 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.

D. Utility

The study in Section IV-C 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 $2D$ and $2D$+*SA* marginals defined in Section IV-B, and we measure the utility of a noisy query answer by the *relative error* of the answer. 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 [15], we set $\delta = 10^{-4} \times |\mathcal{D}|$, where $|\mathcal{D}|$ is the number of records in the dataset $\mathcal{D}$. 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.

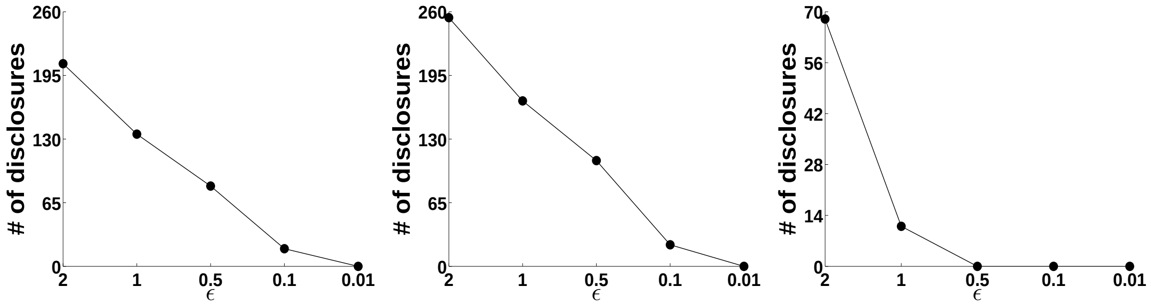
It is interesting to cross-examine the overall error in Fig. 4 and the number of disclosures in Fig. 3. While a smaller ϵ helps prevent disclosures, it degrades the utility by a larger overall error. For example, at $\epsilon = 0.01$, no disclosure is found in any dataset, but the overall error is awfully 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 Fig. 3.

E. Discussion

Both the theoretical analysis in Theorem 1 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.

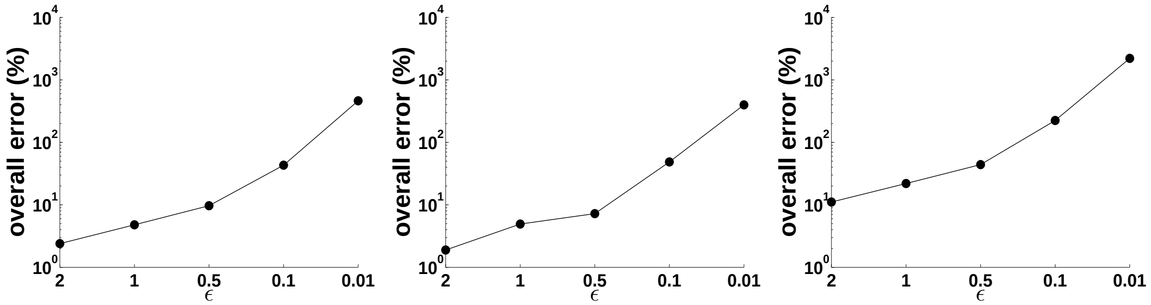
V. Conclusions

Differential privacy aims to hide the participation of an individual and considers learning sensitive information from other records, i.e., non-independent reasoning (NIR), as non-violation of privacy. The threat of NIR is that it does not require the participation of an individual but still can accurately infer the sensitive information of the individual. We formalized a notion of disclosures arising from accessing randomized query answers and NIR (Section II), presented a formal analysis and a theoretical explanation on such disclosures when published query answers are randomized by the differential privacy mechanism (Section III), and evaluated the occurrence of such disclosures on real life datasets (Section IV). Our study suggests that whenever the



(a) EDU (b) OCC (c) SALARY

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



(a) EDU (b) OCC (c) SALARY

Figure 4: Overall Error

differential privacy mechanism delivers a good utility for data analysis, this utility could also be used for learning sensitive information through NIR because both NIR and utility are based on accurate query answers. Therefore, if hiding sensitive information of an individual is the privacy goal, in contrast to hiding the participation of an individual, differential privacy is not a suitable choice.

VI. Acknowledgment

This work was partially supported by a Discovery Grant of Natural Sciences and Engineering Research Council of Canada.

References

[1] How companies learn your secrets. *The New York Times Magazine*, 2012.

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

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

[4] C. Bee-Chung, K. Daniel, L. Kristen, and M. Ashwin. Privacy-preserving data publishing. *Foundations and Trends in Databases*, 2(1-2):1–167, 2009.

[5] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: The sulq framework. In *PODS*, pages 128–138. ACM, 2005.

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

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

[8] R. Chaytor and K. Wang. Small domain randomization: Same privacy, more utility. *Proc. VLDB Endow.*, 3(1-2):608–618, Sept. 2010.

[9] G. Cormode. Personal privacy vs population privacy: Learning to attack anonymization. In *KDD*, pages 1253–1261. ACM, 2011.

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

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

[12] C. Han. Phd thesis. Simon Fraser University, Canada.

[13] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD*, pages 193–204. ACM, 2011.

[14] K. Wang, C. Han, A. W. Fu, R. C. W. Wong, and P. S. Yu. Reconstruction privacy: Enabling statistical learning. In *EDBT*, 2015.

[15] X. Xiao, G. Bender, M. Hay, and J. Gehrke. ireduct: Differential privacy with reduced relative errors. In *SIGMOD*, pages 229–240. ACM, 2011.

[16] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, pages 139–150. VLDB Endowment, 2006.

[17] X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. *IEEE Transactions on Knowledge and Data Engineering*, 23(8):1200–1214, 2011.