EAI Endorsed Transactions
on Security and Safety

Research Article EAI.EU

Exploring the Privacy Bound for Differential Privacy: From Theory to Practice

Xianmang He[1], Yuan Hong[2,*], Yindong Chen[3]

[1]School of Computer Network and Security, Dongguan University of Technology
[2]Department of Computer Science, Illinois Institute of Technology
[3]College of Engineering, Shantou University

Abstract

Data privacy has attracted significant interests in both database theory and security communities in the past few decades. Differential privacy has emerged as a new paradigm for rigorous privacy protection regardless of adversaries prior knowledge. However, the meaning of privacy bound ϵ and how to select an appropriate ϵ may still be unclear to the general data owners. More recently, some approaches have been proposed to derive the upper bounds of ϵ for specified privacy risks. Unfortunately, these upper bounds suffer from some deficiencies (e.g., the bound relies on the data size, or might be too large), which greatly limits their applicability. To remedy this problem, we propose a novel approach that converts the privacy bound in differential privacy ϵ to privacy risks understandable to generic users, and present an in-depth theoretical analysis for it. Finally, we have conducted experiments to demonstrate the effectiveness of our model.

Received on 19 December 2018; accepted on 21 January 2019; published on 25 January 2019

Keywords: Differential Privacy, Inference, Privacy Bound

doi:10.4108/eai.8-4-2019.157414

1. Introduction

Large volumes of sensitive personal data (e.g., medical records, transaction history) are being ubiquitously collected by assorted organizations (e.g., hospitals, retailers). To further unlock the utility of a wide variety of datasets, these organizations often need to provide public access to such datasets or share the data with a third party (e.g., researchers or data analysts). This may pose great threats to the privacy of individuals.

To this end, the data privacy research has attracted significant interests in both database and security in the past few decades. As a major branch of the existing works, privacy preserving data publishing techniques [9, 27] have been proposed to sanitize specific datasets such that the output data can be *publishable* by satisfying a pre-defined privacy notion. More specifically, such privacy preserving techniques can be categorized into two types: (1) anonymization [18, 19, 27], and (2) differential privacy [1, 4].

Anonymization. In this case, a trusted curator computes and publishes an anonymized output data. Figure 1(a) shows such an example (table T_1). Sufficient degree of anonymization is achieved by hiding a record in a crowd with some records having the same values. In such non-interactive setting [16, 18, 19, 27], data recipients passively receive the anonymized table, and then issues some queries over the anonymized data.

Name	Age	Zipcode	Diesease
Alex	20	15k	Bronchitis
Bob	25	42k	Pneumonia
Jane	33	71k	Flu
Cathy	38	25k	Gastritis
Eva	44	56k	Emphysema
Frank	47	18k	Dyspepsia
David	53	31k	Bronchitis
Helen	61	35k	Flu

(a) Input Data T_1

Name	Age	Zipcode	Diesease
Alex	[20-38]	[15k-71k]	Bronchitis
Bob	[20-38]	[15k-71k]	Pneumonia
Jane	[20-38]	[15k-71k]	Flu
Cathy	[20-38]	[15k-71k]	Gastritis
Eva	[44-61]	[18k-56k]	Emphysema
Frank	[44-61]	[18k-56k]	Dyspepsia
David	[44-61]	[18k-56k]	Bronchitis
Helen	[44-61]	[18k-56k]	Flu

(b) Generalized Data

Figure 1. Example of Anonymization

Differential Privacy. In this case, a curator sits between the data recipients and the database. Data recipients issue statistical queries over the database (as the utility of the data). Data recipients can infer information from the database via their queries. Thus, in order to protect

*Corresponding author. Email: Yhong26@iit.edu

the privacy of the respondents, the statistical queries issued by the data recipients and/or the responses to these queries, can be modified by the curator.

More specifically, without loss of generality, we adopt *count* queries as our example in the following, as illustrated in Example 1 (e.g., in Table T_1). Inferences from other types of queries such as *min*, *max*, *mean* or *sum* can be handled in a similar way.

Example 1. Consider an adversary who have some background knowledge of the individuals (e.g., Alex with age 20 and Zipcode 15k), and would like to infer the fact whether Alex is involved in Table T_1 (which is used for analysis) in Figure 1(a) or not.

- q_1:

  Select $Count(*)$ From T_1
  Group By Age, Zipcode
  HAVING Age =20 AND Zipcode = 15k

The response of q_1 would be the count of individuals with age 20 and zipcode 15k. Then, the adversary can infer if Alex is involved in the analysis (query) based on the query result. An effective approach to address the above inferences (via the queries) is output perturbation [2, 5], which injects small amount of random noise into each query result. Numerous output perturbation techniques are available in the literature of statistics. However, such techniques are not based on a rigorous definition of privacy.

Recently, differential privacy [1, 4] has been the breakthrough in this field, and it provides strong privacy guarantee to prevent adversaries from inferring the status of any individual in the database (included or not) from their queries over the databases. Roughly speaking, differential privacy ensures that the removal or addition of a single tuple in the input data (e.g., T_1 in Figure 1(a)) does not substantially affect the outcome of any analysis. Therefore, the privacy risks are bounded regardless of adversaries' background knowledge – the query results are *indistinguishable* if queries are posed over any two neighboring databases (with or without any individual's data).

1.1. Motivation: (1) Interpreting the Privacy Bound for Differential Privacy to Practical Privacy Risks.

Although the notion of differential privacy has successfully achieved the objectives of privacy protection in many applications [10, 12, 21], the meaning of the privacy budget ϵ is still unclear to real users in practice. In general, the degree of privacy preservation gained by ϵ-differentially private algorithms is anti-monotonic on the privacy bound ϵ. That is, the smaller ϵ is, the better privacy protection can be achieved.

From practitioners' point of view, while ϵ was originally derived from the mathematical or probabilistic domain, it is difficult to quantitatively measure the strength of the privacy protection provided by differential privacy with any specific ϵ. This makes it hard for ordinary users to select appropriate values of ϵ for privacy protection, while maximizing the utility of the analytical results produced by different privacy mechanisms. Cynthia Dwork and Adam Smith put forward related questions in [6]: "What does it mean not to provide differential privacy? Failure to provide ϵ-differential privacy might result in $2 \cdot \epsilon$-differential privacy. How bad is this? Can this suggest a useful weakening? How much residual uncertainty is enough?". The main idea underlying these questions is how to interpret the degree of protection provided by ϵ-differential privacy to practical privacy risks (e.g., the upper bound of the probability that any individual's data is included in the analysis).

Interpretive Inference Model. A practical approach to interpret the privacy bound ϵ would be propose an interpretive inference model on the ϵ-differentially private query results – the probability that any individual tuple (viz. individual's data) is inferred to be included in the input database (denoted as "inference probability" for simplicity). As a consequence, *the inference probabilities can be explicitly understandable to generic data owners*). Note that the interpretive inference model is proposed for interpreting privacy risks in ϵ to practical privacy risks of individual tuples.

1.2. Motivation: (2) How to Choose An Appropriate ϵ

Then, after realizing the degree of privacy guarantee with a given interpretive inference model, the next open question would be *how to choose appropriate values for* ϵ based on data owners' required privacy risks (e.g., probabilities that any individual's data is included). Referring to the interpretive inference model used for converting ϵ to probabilities (of inferences), this question can also be interpreted as follows. Given any maximum inference probability, what is the maximum ϵ (upper bound) tolerable in differential privacy to satisfy the practical privacy guarantee (represented as the maximum inference probabilities).

To this end, Jaewoo Lee et al. [15] proposed an interpretive inference model which can be used to derive the upper bounds of ϵ given any maximum inference probability. We illustrate such upper bounds in the following example.

Example 2. We utilize a real database CENSUS that is frequently studied in privacy research [17, 28, 32, 34]. It contains the demographic information of 600k American adults. Each tuple has eight attributes: Age, Gender, Education, Marital, Race, Work-class, Country, and Income.

For instance, we apply *count* queries on the database. Suppose that we want to enforce the adversary's

probability of successfully identifying the individuals from the counting queries to be no greater than $\rho \leq \frac{1}{10}$ (maximum inference probability). In order to achieve this protection goal with the existing theoretical result [15] (more details will be given in Section 3), we thus have: the upper bound of differential privacy budget ϵ should satisfy $\epsilon \leq \frac{\Delta f}{\Delta v} \ln \frac{(n-1)\rho}{1-\rho}$, where n represents the number of records, Δf is the sensitivity of query, Δv is the maximum distance between function values of every possible world (*the same information needed to calculate global sensitivity*) [15], and ρ is the maximum inference probability. Then, if given $n = 600,000$ records, the bound yields

$$\epsilon \leq \frac{1}{1} \ln \frac{(600,000-1)(\frac{1}{10})}{1-\frac{1}{10}} \approx 11.1 \quad (1)$$

where Δv is no greater than 1 for *count* queries.

In the above example, the upper bound of ϵ would be 11.1, which might exceed our expectation. In other words, such large ϵ can satisfy $\rho \leq \frac{1}{10}$ in their interpretive inference model but can be vulnerable in other cases (for instance, in our proposed interpretive inference model, ϵ=11.1 would result in $\rho > \frac{1}{10}$) (*higher privacy risks than the data owners' demand*).

Furthermore, in the interpretive inference model proposed in [15], ϵ is proportional to $\ln(n)$ where n is data size. As n increases, the upper bound of ϵ also increases. In case of a large or small n, the derived bound would be meaningless (unbounded or negative). From the above examples, we can see that existing solutions have their inherent drawbacks. Motivated by such observations, we propose a novel interpretive inference model, which can be used to evaluate the probability or confidence that the adversary will be able to identify any individual from the *noise-injected* queries over the dataset. This enables us to understand the privacy implications of differentially private techniques in a much clearer way.

1.3. Our Contributions

The major contributions of this paper are summarized as follows.

- This paper presents a novel interpretive inference model to convert the privacy bound ϵ in differential privacy to inference probabilities that any individual is included in the input data (for queries). The proposed interpretive inference model and converted inference probabilities have addressed the drawbacks of the existing models [15, 24].
- Based on the proposed interpretive inference model, we present an instantiation for choosing appropriate ϵ (maximum privacy bound in differential privacy), which should effectively bound the risks of inferring the presence/absence of individuals (given the maximum inference probability) in generic differentially private algorithms.
- An in-depth theoretical analysis of our approach is provided, and a set of experiments are conducted to confirm the effectiveness of our approach.

The rest of the paper is organized as follows. In Section 2, we describe the preliminaries for differential privacy. In Section 3, we present the analysis for two representative existing works. Then, in Section 4, we propose our interpretive inference model and the upper bound for ϵ in differential privacy (given the maximum inference probability). Section 5 demonstrates the experimental results, and Section 6 reviews related work. Finally, Section 7 gives the concluding remarks.

2. Preliminaries

In this section, we will first describe the basic mechanism of differential privacy, and then present the Laplace distribution which contributes to a generic differentially private approach.

2.1. Differential Privacy

The most commonly-used definition of differential privacy is ϵ-differential privacy, which guarantees that any individual tuple has negligible influence on the published statistical results, in a probabilistic sense. Specifically, a randomized algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if and only if for any two databases D_1, D_2 that differ in exactly one record, and any possible output O of A, the ratio between the probability that A outputs O on D_1 and the probability that A outputs O on D_2 is bounded by a constant. Formally, we have

$$\frac{|Prob(\mathcal{A}(D_1) = O)|}{|Prob(\mathcal{A}(D_2) = O)|} \leq e^{\epsilon} \quad (2)$$

where ϵ is a constant specified by the user, D_1, D_2 differ in at most one element, and e is the base of the natural logarithms. Intuitively, given the output O of $\mathcal{A}$, it is hard for the adversary to infer whether the original data is D_1 or D_2, if the parameter ϵ is sufficiently small. Similarly, ϵ-differential privacy also provides any individual with plausible deniability that her/his record was in the databases.

The earliest and most widely-adopted approach for enforcing ϵ-differential privacy is the Laplace mechanism [4], which works by injecting random noise $x \propto lap(\lambda)$ that follows a Laplace distribution into the output of the original O, and the deterministic algorithm A obtains its randomized version $O + x$, that is, $A(D) = O + x$, where $\lambda = \frac{\triangle f}{\epsilon}$.

Definition 1 (Sensitivity). The sensitivity $\triangle f$ of the query function f is defined as the maximal L_1-norm distance between the exact answers of the query q (i.e., $q(D_1)$ and $q(D_2)$) on any neighboring databases:

$$\triangle f = max_{D_1,D_2} \|q(D_1) - q(D_2)\|$$

2.2. The Laplace Distribution

A random variable has a Laplace(μ, b) distribution if its probability density function is

$$f(x|\mu, b) = \frac{1}{2b} e^{-\frac{|x-u|}{b}} = \begin{cases} \frac{1}{2b} e^{\frac{x-\mu}{b}} & x < \mu \\ \frac{1}{2b} e^{\frac{\mu-x}{b}} & x \geq \mu \end{cases} \quad (3)$$

It is straightforward to integrate the Laplace distribution. Its cumulative distribution function is as follows:

$$F(x) = \begin{cases} \frac{1}{2} e^{\frac{x-\mu}{b}} & x < \mu \\ 1 - \frac{1}{2} e^{\frac{\mu-x}{b}} & x \geq \mu \end{cases} \quad (4)$$

3. Inferences on the Privacy Bound ϵ

In this section, we first introduce a key concept for describing the adversary's inference ability: *Potential Input Set*. Then, we provide an in-depth analysis of two existing work [15, 24] on converting the guarantee of privacy budget ϵ to inference probabilities.

3.1. Potential Input Set

Definition 2 (Potential Input Set). Given any output S of a differentially private algorithm $\mathcal{A}$, the potential input set Ψ is a set of corresponding possible inputs $\forall D_i$. Then, $\Psi = \{D_i | \mathcal{A}(D_i) = S\}$

Note that our interpretive inference model can be applied to different kinds of queries, e.g., *count*, *sum*, *min* and *max*. We use *sum* queries in the following example.

Example 3. In Example 2, if *sum* queries are given, there will be $n = 60,000$ possibilities of the potential input set Ψ: any subset $X \subseteq$ CENSUS with $n-1$ tuples sampled from CENSUS can be a possible input D_i. While in Example 1, there are only 2 possibilities for query q_1: $\Psi = \{\emptyset, \{Alex\}\}$.

The cardinality of the potential input set can be very large or extremely small, which depends on external knowledge. Basically, differential privacy hides the presence of any individual in the database from data users by making any two output distributions in Ψ (one is with individual and the other is without) computationally indistinguishable. The adversary's goal is to figure out whether $D_i \in \Psi$ is true or not.

3.2. Lee and Clifton's Interpretive Inference Model [15]

In order to bound ϵ, a few assumptions are made by Lee and Clifton [15], e.g., assuming that the adversary has a database D consisting of n tuples, has an infinite computational power, and knows everything about the universe besides which individual is missing in the database. The adversary maintains a set of tuples $< w, \alpha, \beta >$ for each possible combination w of D', with $n-1$ records sampled from D (i.e., $D' \in D$ and $|D'| = |D| - 1$).

Consider the discussions in Section 2, we can infer that, either $D = D_1, D' = D_2$ or $D = D_2, D' = D_1$ holds. Let Ψ denote the set of all possible combinations of D' ($|\Psi| = n$). α and β are the adversary's prior belief and posterior belief on $w = D'$, respectively, after given a query response.

For simplicity, they assume that α is a uniform prior, i.e., $\forall w \in \Psi$, $\alpha(w) = \frac{1}{n}$. They refer to each possible combination w in Ψ as a possible world. The posterior belief β is defined as

$$\beta(w) = P(w = D'|\gamma) = \frac{Prob(A_f(w) = \gamma)}{\sum_{\varphi \in \Psi} Prob(A_f(\varphi) = \gamma)} \quad (5)$$

where A_f is an ϵ-differentially private mechanism for query function f.

Given the best guess w', the confidence of the adversary's guess is calculated as $\beta(w') - \alpha(w')$. Authors treated the adversary's posterior belief on each possible world as the risk of disclosure. Starting from Equation (5), after several steps of deduction, the upper bound of ϵ can be derived as follows:

$$\epsilon \leq \frac{\Delta f}{\Delta v} \ln \frac{(n-1)\rho}{1-\rho} \quad (6)$$

Although such study is the state-of-the-art, the upper bound provided in Equation (6) has two drawbacks that greatly limit its practical applicability.

- First, it is somewhat surprising that the upper bound is directly proportional to $\ln(n)$, and n is the size of the potential input set. As illustrated in Example 2, the size n is a crucial component in deciding ϵ. Therefore, the solution may be not suitable when the size is too small or large. Moreover, the upper bound would be infinite if the potential input set contains only a single tuple. Therefore, we can safely make a conclusion that the upper bound described above is not always applicable and has its inherent disadvantage.

- Second, this solution still cannot estimate the *probability that a certain tuple is included or*

not. Specifically, when the differentially private algorithm $\mathcal{A}$ returns a query, its response means nothing for interpreting the privacy risks for individual tuples regardless of whether the answer is large or small.

Let us illustrate these with Example 4.

Example 4. With the purpose of identifying the victim Alex, the adversary issues a query q_1 as shown in Example 1. As a running example, the ϵ-differential private algorithm $\mathcal{A}$ answers its query with noise -1.1, then returns $\mathcal{A}(q_1) = -0.1$ to the adversary. No matter what the answer is, the adversary cannot make an assertion regarding whether Alex has contributed to the result or not.

3.3. Naldi et al.'s Interpretive Inference Model [24]

In [24], Naldi et al. provided a different approach for choosing ϵ. With respect to the above notation, the new equation for choosing ϵ can be given as follows:

$$\epsilon = -\frac{\ln(1-\rho)}{q(D) \times w} \tag{7}$$

The parameter $w > 0$ indicates that the true value of $q(D)$ is within interval $[q(D)(1-w),\ q(D)(1+w)]$ with probability ρ.

Equation (7) is derived from below:

$$\rho = \mathbb{P}[q(D) \times (1-w) < X < q(D) \times (1+w)] \tag{8}$$

where X is the guess value.

Example 5. Suppose the query result is $q(D) = 100$ and we wish the query output to be within ± 20% of the true value (i.e., 100) with 80% probability, then $\epsilon \leq \ln(1-0.8)/(0.2 \cdot 100) = 0.08$.

On the other hand, while $q(D)$ is negative, for instance, the *min* query equals −100, Equation (7) might be totally wrong.

From Examples 4 and 5, we can see that although choosing proper ϵ for differential privacy has attracted many attentions, and has successfully achieved some upper bounds, state-of-the-art upper bounds derived via Equations (6) or (7) do not function properly in many cases. Based on such observation, we claim that the problem of choosing proper values of ϵ remains open.

4. Novel Interpretive Inference Model on the Privacy Bound ϵ

In this section, we present our interpretive inference model and how to choose ϵ.

4.1. Inferring Query Results

First, for each noise variable x added by algorithm $\mathcal{A}$, which follows Laplace distribution $f(\mu, b)$, it is impossible for the adversary to guess it accurately. However, the adversary can guess x if it falls in a short interval. That is, for each noise variable x, the adversary generates x', which is also governed by the Laplace distribution $f(\mu, b)$. If $|x - x'| < L$, we consider that the inference is successful. The length L is related to query types. For example, in the case of *count* queries, L equals 0.5.

Consider that the guess x' is also generated according to Laplace distribution. Then, the confidence of the adversary's guess can be calculated using Definition 3.

Definition 3. Given a guess x', the adversary's confidence in guessing the exact x that falls into interval L is defined as

$$Prob(|x - x'| < L) = F(x+L) - F(x-L) \tag{9}$$

The definition of F is presented in Equation (4).

We note that probability $Prob(|x - x'| < L)$ is not a fixed value, but varies with x, which is sampled from a Laplace distribution. The nearer that x appears in location u, the larger the probability $Prob(|x - x'| < L)$ is. On the contrary, while x is picked far away from the the location u, probability $Prob(|x - x'| < L)$ becomes much smaller.

Therefore, it makes no sense to calculate the probability for each noise variable x, since such noise can be sampled as much as you want, which is inexhaustible. Moreover, the adversary cannot infer the exact x after deriving response $q(D) + x$. Luckily, we can compute its mathematical expectation that reflects the average level.

Theorem 1. Given response $q(D) + x$, interval L, the mathematical expectation of the probability that the adversary can guess the exact value falling into interval $[q(D) - L, q(D) + L]$ can be given as follows: $E(Prob(|x - x'| < L)) = 1 - (1 + \frac{L \cdot \epsilon}{2 \cdot \Delta}) \cdot e^{-\frac{\epsilon \cdot L}{\Delta}}$.

Proof. We can use $p(x)$ to denote $F(x+L) - F(x-L)$ for simplicity. In the light of the definition of mathematical expectation in the area of probability theory, we have the following equation:

$$E(Pr(|x - x'| < L)) = \int_{-\infty}^{+\infty} p(x) \cdot f(x) \mathrm{d}x \tag{10}$$

where the definition of $f(x)$ is given in Equation (3). In order to calculate the final result, the integration Equation (10) is divided into 4 parts:

$$\begin{aligned}
&\int_{-\infty}^{+\infty} p(x)\cdot f(x)\mathrm{d}x \\
&= \int_{-\infty}^{\mu-L} p(x)\cdot f(x)\mathrm{d}x + \int_{\mu-L}^{\mu} p(x)\cdot f(x)\mathrm{d}x \\
&+ \int_{\mu}^{\mu+L} p(x)\cdot f(x)\mathrm{d}x + \int_{\mu+L}^{+\infty} p(x)\cdot f(x)\mathrm{d}x
\end{aligned} \tag{11}$$

Then, after Equation (9) is substituted into Equation (11), thus we have:

$$\begin{aligned}
&= \int_{-\infty}^{\mu-L} (\frac{1}{2}e^{\frac{x+L-\mu}{b}} - \frac{1}{2}e^{\frac{x-L-\mu}{b}})\cdot\frac{1}{2b}e^{\frac{x-\mu}{b}}\mathrm{d}x \\
&+ \int_{\mu-L}^{\mu} \left(\left(1-\frac{1}{2}e^{\frac{\mu-(x+L)}{b}}\right) - \frac{1}{2}e^{\frac{x-L-\mu}{b}}\right)\cdot\frac{1}{2b}e^{\frac{x-\mu}{b}}\mathrm{d}x \\
&+ \int_{\mu}^{\mu+L} \left(\left(1-\frac{1}{2}e^{\frac{\mu-(x+L)}{b}}\right) - \frac{1}{2}e^{\frac{x-L-\mu}{b}}\right)\cdot\frac{1}{2b}e^{\frac{\mu-x}{b}}\mathrm{d}x \\
&+ \int_{\mu+L}^{+\infty} \left(\left(1-\frac{1}{2}e^{\frac{\mu-(x+L)}{b}}\right) - \left(1-\frac{1}{2}e^{\frac{\mu-(x-L)}{b}}\right)\right)\cdot\frac{1}{2b}e^{\frac{\mu-x}{b}}\mathrm{d}x
\end{aligned} \tag{12}$$

Now, we can integrate the four parts in Equation (12), respectively.

$$\begin{aligned}
&= \frac{1}{8}\left(e^{\frac{2x-2\mu+L}{b}} - e^{\frac{2x-2\mu-L}{b}}\right)\Big|_{-\infty}^{\mu-L} \\
&+ \left(-\frac{1}{4b}e^{\frac{-L}{b}}x + \frac{1}{2}e^{\frac{x-\mu}{b}} - \frac{1}{8}e^{\frac{2x-2\mu-L}{b}}\right)\Big|_{\mu-L}^{\mu} \\
&+ \left(-\frac{1}{4b}e^{\frac{-L}{b}}x - \frac{1}{2}e^{\frac{\mu-x}{b}} + \frac{1}{8}e^{\frac{2\mu-2x-L}{b}}\right)\Big|_{\mu}^{\mu+L} \\
&+ \frac{1}{8}\left(e^{\frac{2\mu-2x-L}{b}} - e^{\frac{2\mu-2x+L}{b}}\right)\Big|_{\mu+L}^{+\infty}
\end{aligned} \tag{13}$$

Finally, we arrive at:

$$\begin{aligned}
&= \left(\frac{1}{8}e^{\frac{-L}{b}} - \frac{1}{8}e^{\frac{-3L}{b}}\right) + \left(\frac{1}{2} - \left(\frac{5}{8}+\frac{L}{4b}\right)e^{\frac{-L}{b}} + \frac{1}{8}e^{\frac{-3L}{b}}\right) \\
&+ \left(\frac{1}{2} - \left(\frac{5}{8}+\frac{L}{4b}\right)e^{\frac{-L}{b}} + \frac{1}{8}e^{\frac{-3L}{b}}\right) + \left(\frac{1}{8}e^{\frac{-L}{b}} - \frac{1}{8}e^{\frac{-3L}{b}}\right) \\
&= 1 - \left(1+\frac{L}{2b}\right)e^{\frac{-L}{b}}
\end{aligned} \tag{14}$$

Recall that, to implement ϵ-differential privacy, the noise λ is governed by $Lap(\frac{\triangle}{\epsilon})$. Therefore, by replacing b with $\frac{\triangle}{\epsilon}$, the theorem has been proven. □

Example 6. Let $\epsilon = 1$, and the adversary submits *count* queries, then the probability of successful inference that the adversary can achieve is 24.18%. For example, if the adversary submits a workload of 10,000 queries, then there are about 2,418 queries, of which the value of $q(D)$ could be correctly inferred (Note that $\triangle = 1, L = 0.5$).

4.2. Inferring Record Status: Present or Absent

Recall that given a query answer $q(D)+x$, we can derive its exact answer $q(D)$ with a certain probability. However, it is not our ultimate goal, as the goal of the differential privacy is to determine whether a certain tuple has contributed to the query result or not. Hence, we need to go further into reduction of Theorem 1, which aligns with the objectives of differential privacy.

Example 7. Let's continue with Example 1, and assume that $\epsilon = 1$. Suppose the adversary tries to identify victim Alex, and issues a query q_1 as shown in Example 1. The random noise drawn from the Laplace distribution with mean 0 and scale factor $b = 1$ is 0.7, then the response is produced as $\mathcal{A}(q_1) = 1 + 0.7 = 1.7$ by the ϵ-differentially private algorithm $\mathcal{A}$.

As a result, the adversary generates an additional noise x'. If $0.7 - x' > 0.5$, that is, $x' < 0.2$, the inference is successful, and the adversary can infer that Alex's data is included. Otherwise, the inference fails.

This example inspires us that *a certain tuple is absent or not can be determined by extrapolating the query result with sufficient background*. Generally, let q be a *count* query such that $q(D) = c$, and the adversary knows that $q(D)$ falls into the potential input set $\Psi = \{c_1, c_2, \ldots, c_n\}$ where some integer constants $c_1 \le c_2 \le \cdots \le c_n$ (note that $\forall i \in [2, n], c_i - c_{i-1}$, the intervals between every two consecutive integer constants can be greater than 1). The main task of the adversary is to make an answer for $c_i \in \Psi$ or not, which is close to the value of $y = x + q(D) - x'$. In other words, the adversary pick the answer that minimizes $|c_i - y|$, formally

$$\arg\min_{\forall i\in\{1,\ldots,n\}} |c_i - y| \tag{15}$$

Observing this, we propose an inference algorithm for the *count* query as shown in Algorithm 1. The adversary issues a query $q(D)$, which restricts the query to a specific victim (e.g., query q_1 in Example 1). The differentially private mechanism $\mathcal{A}$ generates noise x, which is governed by Laplace distribution $f(\mu, b)$. Then, noise x plus the real value $q(D)$, that is, $x + q(D)$, is returned to the adversary. The adversary generates some different noise x', and computes the value $y = x + q(D) - x'$. Eventually, the adversary concludes that: if the answer of y equals c, then the inference is successful, otherwise will be failure.

Theorem 2. For *count* queries, Algorithm 1 can correctly infer the record status with probability $1 - \frac{(1+\frac{\epsilon}{4})\cdot e^{-\frac{\epsilon}{2}}}{2}$ where $c = c_1$ or c_n, otherwise the probability is $1 - (1 + \frac{\epsilon}{4})e^{-\frac{\epsilon}{2}}$.

Algorithm 1 Inference For *Count* Query

Input: $\mathcal{A}(q(D)) = x + q(D), \Psi$
Output: Success or Failure
/* Lap(μ, b): the scale factor b, and the location u and $q(D) \in \{0, 1\}$*/
1: The adversary generates a noise x' variable, which is governed by Laplace distribution $f(\mu, b)$
2: $y \leftarrow x + q(D) - x'$
3: derive $\arg\min_{\forall c_i \in \Psi} |c_i - y|$ where $i = 1, 2, \ldots, n$
4: answer=c_i
5: **if** answer==c **then**
6: **return** Success
7: **else**
8: **return** Failure
9: **end if**

Proof. Consider the symmetry that the probability that y falls into $[c_2 + 0.5, +\infty)$ equals to that of y falls into $(-\infty, c_2 - 0.5]$. Combined with Theorem 1, we can complete the proof of this theorem. Notice that for *count* queries, L equals 0.5. □

Example 8. As for the *min* (*max* or *sum*) queries, the adversary may know the fact that the victim Alex is involved and his minimal income. Then, the adversary issues query q_3:

Select Min(Income) From CENSUS

This is the case with the potential input set $|\Psi| = 1$, which is different from the *count* query.

Indeed, it is sufficient for the adversary to determine whether the victim is present or not for the *count* query, while for the *min* query, the case would change. The adversary can only determine that the victim's income falls into a small interval $[q(D) - L, q(D) + L]$, where L is half length of the interval.

We extend the above approach to other types of queries, such as *mean*, *sum*, *min* and *max*. The main difference is the predicate condition (see Line 3 in Algorithm 2).

Algorithm 2 Inference for *Min*, *Max* or *Sum* Query

Input: $\mathcal{A}(q(D)) = x + q(D)$
Output: Success or Failure
/* Lap(μ, b): the scale factor b, and the location u */
1: The adversary generates a noise x' variable, which is governed by Laplace distribution $f(\mu, b)$
2: $y \leftarrow x + q(D) - x'$
3: **if** $y \in [q(D) - L, q(D) + L]$ **then**
4: **return** Success
5: **else**
6: **return** Failure
7: **end if**

Theorem 3. For *min* (or extended to *max*, *sum*, etc.) queries, Algorithm 2 can correctly infer whether a certain victim is in interval $[q(D) - L, q(D) + L]$ with probability $1 - (1 + \frac{L \cdot \epsilon}{2 \cdot \Delta}) \cdot e^{-\frac{\epsilon \cdot L}{\Delta}}$

Note that Theorem 3 can be directly derived from Theorem 1.

4.3. Choice of ϵ

From the discussions in Section 4.2, we can make the conclusion to a certain extent. Theorem 1 can be used to estimate the risk of disclosing presence/absence of any individual in the database (by the interpretive inference model), which is given in the following corollary.

Corollary 1. Let ρ be the probability of being identified as present in the database, then parameter ϵ on the adversary's probability can be utilized to enforce requirements constrained by the following formulas:

- for *count* queries: $\rho \geq 1 - \frac{(1 + \frac{\epsilon}{4}) \cdot e^{-\frac{\epsilon}{2}}}{2}$
- for *min* (or extended to *max*,*sum*) *queries*: $\rho \geq 1 - (1 + \frac{L \cdot \epsilon}{2 \cdot \Delta}) \cdot e^{-\frac{\epsilon \cdot L}{\Delta}}$.

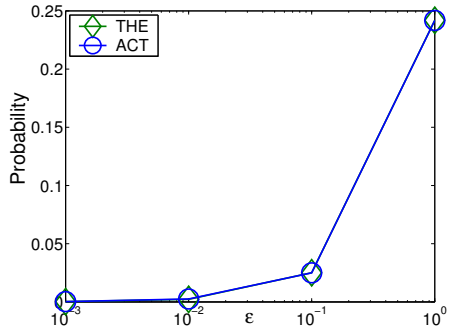
Thus, we illustrate how to choose ϵ to prevent the adversary from successfully identifying an individual (for the maximum inference probability ρ). Note that it is challenging to directly calculate the inverse functions of the above formulas directly. Consider that the formulas in Corollary 1 are functions of monotone decreasing with ϵ. Then, we can approximate it with a binary search strategy: we pre-compute some pairs of $< \rho_i, \epsilon_i >$. Given a ρ, if we can find a $\rho_i = \rho$, then we return ϵ_i. Otherwise, we need to find ϵ_k and ϵ_{k+1} where $\rho_k \leq \rho \leq \rho_{k+1}$. Next, we compute the pair $< \rho_j, \epsilon_j >$ using formulas in Corollary 1 with $\frac{\epsilon_k + \epsilon_{k+1}}{2}$. If $\rho_j > \rho$, we compute ρ_j with $\frac{\epsilon_j + \epsilon_{k+1}}{2}$; otherwise, we compute ρ_j with $\frac{\epsilon_j + \epsilon_k}{2}$. The above procedure terminates as ρ_j approximates ρ within a marginal error where the time complexity is $\log(n)$.

5. Experimental Evaluation

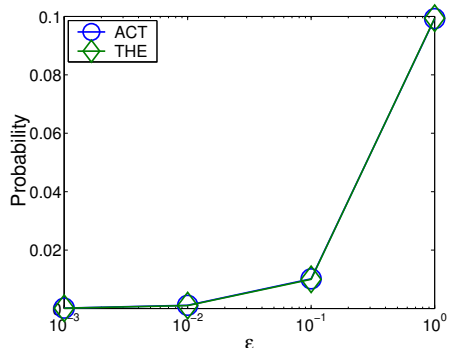
In this section, we conduct experiments to evaluate our proposed approach. Note that the propopsed approach is experimentally incomparable with [15] and [24] since three different approaches have different sets of parameters (which may result in biased comparisons).

Specifically, with comparison of the experimental and theoretical values of ϵ, we will show that our model is fully consistent with the actual tests. In the experiments, our theoretical model is denoted by THE and the actual tests is shortened as ACT. The ACT is conducted by randomly generating n pairs of noise $x_1, x_1', x_2, x_2', ..., , x_n, x_n'$. Denoting the number of noise

pairs that satisfy inequality $|x_i - x'_i| < L, 1 \leq i \leq n$ as m, the probabilities of successful inferences can be measured by the ratio: $\frac{m}{n}$. On the other hand, the THE is directly computed from Theorem 1.



(a) L=0.5



(b) L=0.2

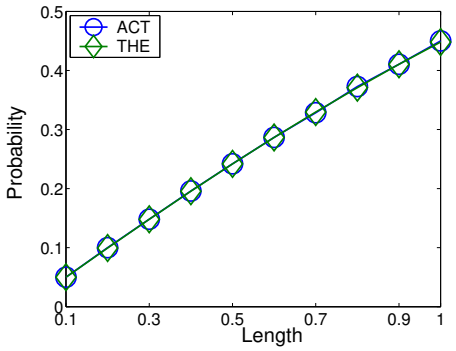
Figure 2. Probability of Successful Inference vs. ϵ

5.1. Results on Varying ϵ

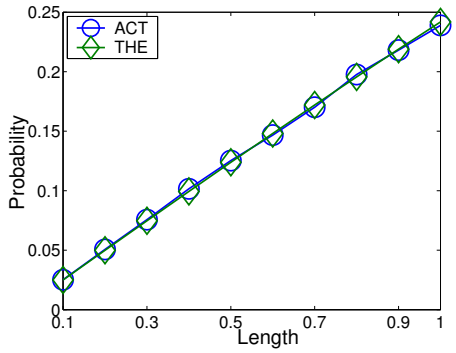
This set of experiments are mainly designed to study the influence of ϵ on the probabilities of successful inferences in THE and ACT. The results with length=0.5 and length=0.2 are shown in Figure 2(a) and 2(b). From the results, we can see that in all the experiments, the probabilities produced by our model are very close to the actual tests. To be specific, varying ϵ from 0.001 to 1, the differences between the two approaches (THE and ACT) are 0.00003, -0.00026, 0.00004, 0.000073 with length=0.5 (see Figure 2(a)). When the length equals 0.2, the differences are -0.00004, -0.00011, -0.000151, 0.000146, as presented in Figure 2(b), respectively. These results further confirm the correctness of our interpretive inference model.

5.2. Results on Varying Length of Interval L

Then we consider comparing THE and ACT by varying the length of intervals from 0.1 to 1. Figure 3(a) and 3(b) show the two probabilities of successful inferences by THE and ACT with $\epsilon = 1$ and $\epsilon = 0.5$, respectively. In general, the probabilities will increase along with the increase of length L. This is because with larger L, it is easier for the adversary to obtain a successful inference. In these figures, we can see that the probabilities generated by both THE and ACT have minor differences, which is consistently achieved, especially when L lies between 0.1 to 1.



(a) ϵ=1.0



(b) ϵ=0.5

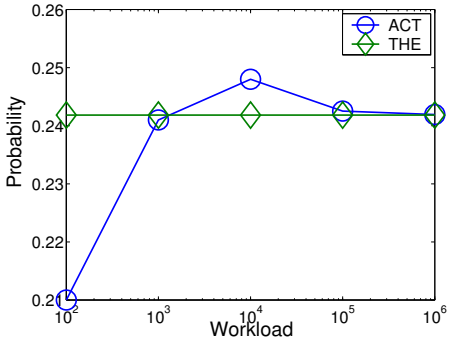
Figure 3. Probability of Successful Inference vs. Length L

5.3. Results on Varying Workload

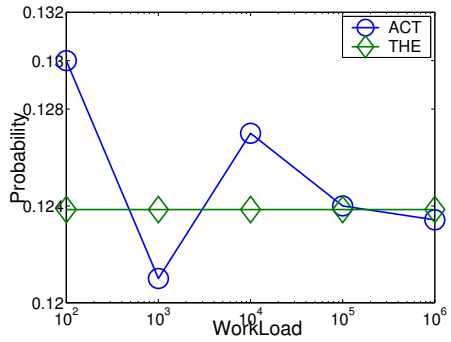
Finally, we demonstrate that how workload affects THE and ACT. More specifically, we consider the workload of testing varies from 100 to 1,000k. Figure 4(a) and Figure 4(b) plot THE and ACT as a function of workload, with $\epsilon = 1$ and $\epsilon = 0.5$, respectively. We can see that the test values become increasingly closer to theoretical values with the growth of workload. This observation can be attributed to the fact that more tests lead to more accurate results. Similarly, the theoretical values are consistently close to the actual tests for large workloads.

5.4. Summary

With the demonstrated experimental results (Figure 2-4), we can confirm that the main result of our model, i.e., Theorem 1, is correct. Our model generates only negligible errors compared with actual tests. Notice



(a) ϵ=1



(b) ϵ=0.5

Figure 4. The Probability of Success vs. Workload

that, the experimental results are derived from noisy queries (e.g., *count* and *sum*): comparing the actual noise generated for queries with the theoretical results (i.e., derived from Theorem 1). Thus, the experimental results are independent of the experimental datasets. In other words, the results would be consistent with the results derived from any dataset (using the same noise).

6. Related Work

Since Dwork [4] proposes the seminal mechanism of ϵ-differential privacy, there has been a large body of work on it. The literature of differential privacy can be classified into three main categories.

The first category aims at studying the properties of differential privacy and its variants. For example, a natural relaxation of differential privacy (ϵ, δ) [6] was proposed where better accuracy (a smaller magnitude of added noise) and generally more flexibility can often be achieved. Authors of [20] reported the design and implementation of the Privacy Integrated Queries (PINQ) platform for privacy-preserving data analysis. Complement to the Laplace mechanism, McSherry and Talwar [22] proposed the exponential mechanism, which works for any queries whose output spaces are discrete. This enables differentially private solutions for various interesting problems where the outputs are not real numbers.

The second category studies new differentially private methods with improved accuracy, without compromising privacy preservation. Privelet [33] was developed as a data publishing technique that ensures ϵ-differential privacy while providing accurate answers for range-count queries, i.e., *count* queries where the predicate on each attribute is a range. The core of their solution is a framework that applies wavelet transforms on the data before adding noise to it. The technique in [1] was designed for releasing marginals, i.e., the projections of a frequency matrix on various subsets of the dimensions. iReduct [31] was designed to compute answers with reduced relative errors. The basic idea of iReduct is to inject different amounts of noise to different query results, thus smaller (larger) values are more likely to be injected with less (more) noise.

The third category includes algorithms for enforcing ϵ-differential privacy in the publication of various types of data, such as relational tables [4, 11, 33], data mining results [8, 21, 29], and histogram publication [35]. Specifically, Xu et al. [35] investigated how the counts in one-dimensional histograms can be released in a differentially private manner. Barak et al. [1] proposed a solution for releasing marginals, each of which contains the counts of pertinent to a projection of the original dataset onto a subset of its attributes. Rastogi and Nath [26] studied the publication of time-series in a distributed setting. In some contexts (e.g., search logs [10, 13, 14]), besides ϵ-differential privacy, a relaxed notion of (ϵ, δ)-differential privacy have been proposed to bound the probabilities (by δ) that the output generated from one of two neighboring inputs D and D' cannot be generated from the other one (since the zero probability cannot be the denominator to be bounded by e^{ϵ}). Similar to ϵ-differential privacy, Mohammady et al. [23] has proposed a privacy notion ϵ-indistinguishability for different views of the outsourced network trace data.

Yang et al. [36] listed some open problems that we believe are important and deserved additional attention from researchers. The first problem is about the actual/physical meaning of privacy budget ϵ. The most relevant prior work (to ours) is [15]. They consider the probability of identifying any particular individual as being in the database, and demonstrate the challenge of setting proper values of ϵ given the goal of protecting individuals in the database with some fixed inference probability. The details of their techniques are discussed in Section 3.

Recently, the differential privacy models have been extended to local differential privacy [3] in which each user locally perturbs its data before disclosing to the untrusted data recipients. The state-of-the-art LDP techniques are proposed to sanitize statistical data to generate histograms/heavy hitters [7], social graphs [25] and function frequent itemset mining [30]. We

intend to extend our approach for local differential privacy in the future.

Finally, privacy bounds have been studied outside differential privacy community. For instance, Zhang et al. [37] studied the privacy bound for identifying which intermediate data sets need to be encrypted and which do not in cloud computing.

7. Conclusion

Although the mechanism of differential privacy has received considerable attention in the past decade, few efforts have been dedicated to studying the practical implications of its given privacy bound (e.g., ϵ) and applying it in practice. In addition, despite its apparent importance in real world, the choice of an appropriate value of ϵ (based on a required quantitative probability that any individual can be identified from the input data) has not been well studied in the literature.

Prior works suffer from some limitations. To address these deficiencies, we have presented a novel interpretive inference model to convert the differential privacy bound ϵ to the probability of identifying any individual from the input database. In addition, it is also possible to determine an appropriate value of the privacy bound ϵ from our inference model for any desired privacy guarantee (i.e., given a limited probability of identification). We have also shown that the upper bound ϵ for differential privacy suggested by prior models is too large – this makes the prior interpretive inference models vulnerable to our inferences performed by the adversaries. We have theoretically and experimentally validated the effectiveness of our model.

Acknowledgments

This work is partially supported by the National Science Foundation of China under Grants No. 61672303 and U1509213. It is also partially supported by the National Science Foundation under Grant No. CNS-1745894 and the WISER ISFG grant.

References

[1] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. 2007. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 273–282.

[2] Avrim Blum, Katrina Ligett, and Aaron Roth. 2013. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)* 60, 2 (2013), 12.

[3] Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. 2018. Privacy at Scale: Local Differential Privacy in Practice. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*. 1655–1658. https://doi.org/10.1145/3183713.3197390

[4] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, Shai Halevi and Tal Rabin (Eds.). Lecture Notes in Computer Science, Vol. 3876. Springer Berlin Heidelberg, 265–284. https://doi.org/10.1007/11681878_14

[5] Cynthia Dwork and Moni Naor. 2008. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality* 2, 1 (2008), 8.

[6] Cynthia Dwork and Adam Smith. 2010. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality* 1, 2 (2010), 2.

[7] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*. 1054–1067. https://doi.org/10.1145/2660267.2660348

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

[9] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.* 42, 4 (2010), 14:1–14:53. https://doi.org/10.1145/1749603.1749605

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

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

[12] Yuan Hong, Jaideep Vaidya, Haibing Lu, Panagiotis Karras, and Sanjay Goel. 2015. Collaborative Search Log Sanitization: Toward Differential Privacy and Boosted Utility. *IEEE Trans. Dependable Sec. Comput.* 12, 5 (2015), 504–518. https://doi.org/10.1109/TDSC.2014.2369034

[13] Yuan Hong, Jaideep Vaidya, Haibing Lu, and Mingrui Wu. 2012. Differentially private search log sanitization with optimal output utility. In *15th International Conference on Extending Database Technology, EDBT '12, Berlin, Germany, March 27-30, 2012, Proceedings*. 50–61. https://doi.org/10.1145/2247596.2247604

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

[15] Jaewoo Lee and Chris Clifton. 2011. How much is enough? choosing ε for differential privacy. In

Information Security. Springer, 325–340.

[16] David Leoni. 2012. Non-interactive Differential Privacy: A Survey. In *Proceedings of the First International Workshop on Open Data (WOD '12)*. ACM, New York, NY, USA, 40–52. https://doi.org/10.1145/2422604.2422611

[17] Jiexing Li, Yufei Tao, and Xiaokui Xiao. 2008. Preservation of proximity privacy in publishing numerical sensitive data. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, New York, NY, USA, 473–486. https://doi.org/10.1145/1376616.1376666

[18] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*. 106–115. https://doi.org/10.1109/ICDE.2007.367856

[19] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. *L*-diversity: Privacy beyond *k*-anonymity. *TKDD* 1, 1 (2007), 3. https://doi.org/10.1145/1217299.1217302

[20] F.D. McSherry. 2009. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 35th SIGMOD international conference on Management of data*. ACM, 19–30.

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

[22] Frank McSherry and Kunal Talwar. 2007. Mechanism Design via Differential Privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS '07)*. IEEE Computer Society, Washington, DC, USA, 94–103. https://doi.org/10.1109/FOCS.2007.41

[23] Meisam Mohammady, Lingyu Wang, Yuan Hong, Habib Louafi, Makan Pourzandi, and Mourad Debbabi. 2018. Preserving Both Privacy and Utility in Network Trace Anonymization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*. 459–474. https://doi.org/10.1145/3243734.3243809

[24] Maurizio Naldi and Giuseppe D'Acquisto. 2015. Differential Privacy: An Estimation Theory-Based Method for Choosing Epsilon. *Computer Science* (2015).

[25] Zhan Qin, Ting Yu, Yin Yang, Issa Khalil, Xiaokui Xiao, and Kui Ren. 2017. Generating Synthetic Decentralized Social Graphs with Local Differential Privacy. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*. 425–438. https://doi.org/10.1145/3133956.3134086

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

[27] Latanya Sweeney. 2002. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570. https://doi.org/10.1142/S0218488502001648

[28] Yufei Tao, Hekang Chen, Xiaokui Xiao, Shuigeng Zhou, and Donghui Zhang. 2009. ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication. *IEEE Transactions on Knowledge and Data Engineering* 21 (2009), 1073–1087. https://doi.org/10.1109/TKDE.2009.65

[29] Jaideep Vaidya, Basit Shafiq, Anirban Basu, and Yuan Hong. 2013. Differentially Private Naive Bayes Classification. In *2013 IEEE/WIC/ACM International Conferences on Web Intelligence, WI 2013, Atlanta, GA, USA, November 17-20, 2013*. 571–576. https://doi.org/10.1109/WI-IAT.2013.80

[30] Tianhao Wang, Ninghui Li, and Somesh Jha. 2018. Locally Differentially Private Frequent Itemset Mining. In *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*. 127–143. https://doi.org/10.1109/SP.2018.00035

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

[32] Xiaokui Xiao and Yufei Tao. 2006. Anatomy: simple and effective privacy preservation. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases*. VLDB Endowment, 139–150.

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

[34] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. 2006. Utility-based anonymization using local recoding. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, 785–790. https://doi.org/10.1145/1150402.1150504

[35] Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, Ge Yu, and Marianne Winslett. 2013. Differentially private histogram publication. *The VLDB JournalÂąÂłThe International Journal on Very Large Data Bases* 22, 6 (2013), 797–822.

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

[37] Xuyun Zhang, Chang Liu, Surya Nepal, Suraj Pandey, and Jinjun Chen. 2013. A Privacy Leakage Upper Bound Constraint-Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud. *IEEE Trans. Parallel Distrib. Syst.* 24, 6 (2013), 1192–1202. https://doi.org/10.1109/TPDS.2012.238