# Privacy-Preserving Characterization and Data Publishing

Jian Ren and Tongtong Li
Department of ECE, Michigan State University, East Lansing, MI 48824-1226. Email:{renjian, tongli}@msu.edu

***Abstract*—The increasing interest in collecting and publishing large amounts of data to public has raised significant privacy concerns. Many Privacy-Preserving Data Publishing (PPDP) techniques have been proposed to address these concerns. However, they often lack proper privacy assurance. In this paper, we first present a novel multi-variable privacy characterization and quantification model and analyze both prior and posterior adversarial belief about individual attribute values and the sensitivity of any identifier in privacy characterization using this model. We show that privacy is nearly impossible to be characterized by a single metric. Instead, we propose two metrics to quantify privacy leakage: distribution leakage and entropy leakage. Applying our framework and the proposed metrics, we can reveal limitations in existing PPDP schemes regarding privacy characterization and leakage. This research contributes to a better understanding and evaluation of these techniques, providing a foundation for the design and analysis of PPDP schemes.**

***Index Terms*—Data privacy, data security, data publishing, privacy quantification, privacy leakage.**

## I. Introduction

Datasets are considered a valuable source of information for the medical research, market analysis and economical measures. While the shared dataset gives useful societal information to researchers, it also creates security risks and privacy concerns to the individuals whose data are published. To avoid possible identification of individuals from records in published data, uniquely identifying information are generally removed from the published data table. While the obvious personal identifiers are removed, the quasi-identifiers may still be used to uniquely identify a significant portion of the population since the released data makes it possible to infer individuals information.

The spate of privacy related incidents has spurred a long line of research in privacy notions for data publishing and analysis [1]–[4]. Unfortunately, no existing scheme is sufficient to prevent attribute disclosure.

In this paper, we introduce a novel data publishing framework. First, we model attributes in a dataset as a multi-variable model. Based on this model, we are able to re-define the prior and posterior adversarial belief about attribute values of individuals. Then we characterize privacy of these individuals based on the privacy risks attached with combining different attributes.

For a given dataset, before it is released, we need to determine to what extent we can achieve privacy. Therefore, we introduce a new set of privacy quantification metrics to measure the gap between prior information belief and posterior information belief of an adversary, from both local and global perspectives. Specifically, we introduce two privacy leakage measurements: *distribution leakage* and *entropy leakage*. We discuss the rationale for these two measurements and illustrate their advantages through examples.

The rest of this paper is organized as follows. In Section II, we analyze the existing PPDP techniques. Our privacy characterization framework and quantification metrics are proposed in Section III and IV, respectively. In Section V, we provide the simulation results. We conclude the paper in Section VI.

## II. Analysis of The Existing PPDP Schemes

In this section, we analyze some representative PPDP schemes.

*$k$-Anonymity:* A table satisfies $k$-anonymity if every record in the table is indistinguishable from at least $k-1$ other records with respect to every set of *quasi-identifier* attributes, which requires the original table is generalized forming *equivalence class* $[C]$ that has at least $k$ records and share values of QIDs before data is published. Unfortunately, it has been shown that $k$-anonymity does not provide sufficient protection against attribute linkage [5], [6].

*$l$-diversity:* $l$-diversity was introduced to address the limitations of $k$-anonymity. An equivalence class is said to have $l$-diversity if there are at least $l$ *well-represented* values for the sensitive attribute. However, it is susceptible to skewness and similarity attacks [4], meaning that when the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information.

*$t$-closeness:* $t$-closeness was designed to combat similarity attack [4]. An equivalence class is said to have $t$-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold $t$, where the value $t$ is merely an abstract distance between two distributions, that could have different meanings in different contexts. Unfortunately, $t$-closeness does not offer the flexibility of having different protection levels for different sensitive attribute values [7]. The function used to measure the distance between distributions is not suitable for protection against attribute linkage on numerical sensitive attributes [8]. Moreover, enforcing $t$-closeness requires the distribution of sensitive attribute values to be the same in all $q$ equivalence classes, which would greatly affect the data utility and significantly damage the correlation between the set of *quasi-identifiers* QID and sensitive attributes. Finally, and most importantly, the distance $t$ is unreliable for quantifying the amount of privacy leakage. This is because, for two published tables $T_1'$ and $T_2'$ satisfying $t_1 < t_2$ does not necessarily imply that $T_2'$ is more privacy-preserving than $T_1'$ [6].

From these discussions, we can see that fully characterizing privacy information and determining potential privacy leakages using a single metrics is challenging, if not impossible.

979-8-3503-7099-7/24/$31.00 ©2024 IEEE

TABLE I: Original Salary/Disease

(a) Original Dataset

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

(b) A 3-diverse Version of Salary/Disease

| | ZipCode | Age | Salary | Disease |
|---|---|---|---|---|
| 1 | 476** | 2* | 3K | gastric ulcer |
| 2 | 476** | 2* | 4K | gastritis |
| 3 | 476** | 2* | 5K | stomach cancer |
| 4 | 4790* | $\geq 40$ | 6K | gastritis |
| 5 | 4790* | $\geq 40$ | 11K | flu |
| 6 | 4790* | $\geq 40$ | 8K | bronchitis |
| 7 | 476** | 3* | 7K | bronchitis |
| 8 | 476** | 3* | 9K | pneumonia |
| 9 | 476** | 3* | 10K | stomach cancer |

TABLE II: (0.167, 0.278)-Closeness for (Salary, Disease)

| | Zip Code | Age | Salary | Disease |
|---|---|---|---|---|
| 1 | 4767* | $\leq 40$ | 3K | gastric ulcer |
| 2 | 4767* | $\leq 40$ | 5K | stomach cancer |
| 3 | 4767* | $\leq 40$ | 9K | pneumonia |
| 4 | 4790* | $\geq 40$ | 6K | gastritis |
| 5 | 4790* | $\geq 40$ | 11K | flu |
| 6 | 4790* | $\geq 40$ | 8K | bronchitis |
| 7 | 4760* | $\leq 40$ | 4K | gastritis |
| 8 | 4760* | $\leq 40$ | 7K | bronchitis |
| 9 | 4760* | $\leq 40$ | 10K | stomach cancer |

## III. Proposed Privacy-Preserving Characterization and Data Publishing

The sensitivity of an attribute comes from combining it with other attributes, instead the attribute itself. To obtain a meaningful privacy characterization, it is necessary to determine the knowledge that the adversary gains about sensitive attributes by considering the combinational relation of different attributes from observing the published dataset.

We employ a multi-dimensional scheme of privacy risk analysis attached with combining different attributes. Assume each individual in a given table $T$ only owns one record represented as a function of multi-variables $v = \{v_1, v_2, \cdots, v_l\}$, where $v$ corresponds to the set of attributes $\mathcal{A} = \{A_1, A_2, \cdots, A_l\}$ in the original dataset. The order of each variable $v_i$, denoted as $\mathrm{ord}(v_i)$, is the number of all possible attribute values. Privacy-preserving techniques apply some generalizations and suppressions to the quasi-identifiers QID to avoid linking individuals to records in the table.

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

- *Value Mapping*: $\forall v_i \in T$ and $v'_i \in T'$, any value $u[v_i]$ in $T$ is mapped to $u'[v'_i]$ in $T'$.
- *Record Mapping*: For the two sets $v = \{v_1, v_2, \cdots, v_l\} \in T$ and $v' = \{v'_1, v'_2, \cdots, v'_l\} \in T'$, any record $u[v]$ in $T$ is mapped to $u'[v']$ in $T'$.
- For any variable $v_i$ and its generalization $v'_i$, it always holds that $\mathsf{ord}(v_i) \geq \mathsf{ord}(v'_i)$.
- After generalization, different combinations of $v'_i$'s in the published table $T'$ naturally divide the table into a set $\mathcal{C} = \{[C_1], [C_2], \cdots, [C_q]\}$ of $q$ equivalence classes.

Publishing a table $T'$ gives different privacy risks for each combination of the generalized variables $\langle v'_i, v'_j \rangle$. As the number of combined variables increases, the privacy risk of an individual increases and it would be easier for an adversary to identify an individual of interest from the published table.

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

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

*Estimated Conditional Distribution of Attributes*: $\forall v_i \in v$, an estimate of the conditional distribution $a_{v_i,v_j}$ of any combination of random variables is previously known by an adversary and is defined as

$$a_{v_i,v_j} = \tilde{P}(v_i \,|\, v_j),\ i = 1, \cdots, \mathsf{ord}(v'_i), j = 1, \cdots, \mathsf{ord}(v'_j),$$

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

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

*Published Conditional Distribution of Attributes*: $\forall v_i \in v$, the conditional distribution $x_{v_i,v_j}$ of any combination of random variables is defined as

$$x_{v_i,v_j} = P(v_i \,|\, v_j), i = 1, \cdots, \mathsf{ord}(v'_i), j = 1, \cdots, \mathsf{ord}(v'_j),$$

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

The goal of any privacy-preserving technique is to minimize the privacy loss between prior and posterior belief as much as possible while maintaining a sufficient level of published data utility. We define this loss as the conditional privacy leakage.

**Definition 4** (Conditional Privacy Leakage)**.** The privacy loss of an individual $u$ belonging to an equivalence class $[C_u]$ with respect to an attribute $v_i$ is the amount of information gained by an adversary represented as the change of the belief after publishing the table $T'$. This leakage $L(v_i \,|\, [C_u])$ is typically the change of an adversarial belief about an attribute's distribution from $a_{v_i,[C_u]}$ to $x_{v_i,[C_u]}$.

## IV. Our Proposed Privacy Quantification

The state-of-the-art approaches to measure privacy [9] can be mainly sub-categorized into uncertainty, information gain or loss, similarity and diversity, and indistinguishability metrics.

Our privacy quantification approach depends on understanding when information leakage happens and how this leakage could be measured. Let $U = \{u_1, u_2, \cdots, u_n\}$ be a finite set of $n$ individuals participating in the data table $T$, $\mathcal{A} = \{A_1, A_2, \cdots, A_l\}$ be the set of $l$ attributes and $u[A_i]$ denote the value of attribute $A_i$ for individual $u$. Define the sensitive attribute $S \subset \mathcal{A}$ as the attribute of interest for an adversary. Two individuals $u_i$ and $u_j$ having the same values of *quasi-identifiers* are assumed to be QID-equivalent if $u_i[Q]$ = $u_j[Q]$, i.e., they share the same equivalence class $[C]$. As previously mentioned, we are generally interested in $a_{v_i,[C_i]}$

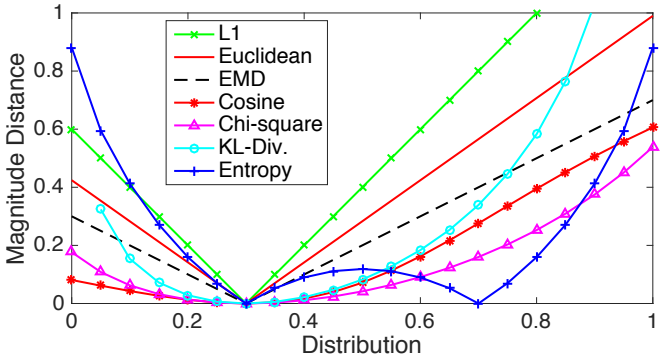


Fig. 1: Comparison of different metrics.

and $x_{v_i,[C_i]}$. Hence, for ease of notations, throughout the rest of this paper, we denote $a_{v_i,[C_i]}$ as $a$ and $x_{v_i,[C_i]}$ as $x$.

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

$$p(u_i[A_j]) = p(u_i[A_j] \,|\, T'), i = 1, \cdots, n, j = 1, \cdots, l,$$

where $p(u_i[A_j])$ denotes the probability of $u_i$ on attribute $A_j$ and $p(u_i[A_j] \,|\, T')$ denotes the conditional probability of $u_i[A_j]$ after the table $T'$ is published.

Nearly any publishing technique could result in some privacy leakage, which makes it necessary to determine the information leakage of any given data publishing schemes. To find suitable metrics, we seek the distance measures based on two criteria. First, the metric should be *sensitivity* and able to capture minor variations in the distributions. Second, the metrics should be *independence*. As shown in Fig. 1, the $L_1$ and the Euclidean distances are the most sensitive metrics in comparison to others. However, the $L_1$ distance has the problem of not being robust under simple transformations such as rotation of the coordinate system. Therefore, it is not a good metric so we choose the Euclidean distance as our first distance metric. We can also see from this figure that entropy distance is the only metrics that is independent of the other ones. This qualifies it to be the second metric.

Let $S = \{s_1, s_2, \cdots, s_m\}$ be the set of all $m$ attribute values of a sensitive attribute $S$. The estimated initial distribution of $S$ given an equivalence class $[C_u]$ is denoted as $a = \{a_1, a_2, \cdots, a_m\}$. The published distribution of $S$ given an equivalence class $[C_u]$ is denoted as $X = \{x_1, x_2, \cdots, x_m\}$.

**Definition 6** (Distribution Leakage)**.** For an individual $u$ belonging to an equivalence class $[C_u]$, the distribution leakage of attribute $S$ given an equivalence class $[C_u]$ is defined as the Euclidean distance between the two distributions $a$ and $x$

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

The distribution leakage $\mathscr{L}_D(S, [C_u])$ defined above can be viewed as a measure of the overall divergence of attribute values distribution from one state to the other. Distribution leakage is closely related to the standard deviation [6].

**Theorem 1.** An individual $u$ belonging to an equivalence class $[C_u]$ has probability distribution on $S$ of $X$. The distribution leakage of an attribute $S$ in the published table $T'$ with respect to uniform distribution is

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

where $\sigma_a$ is the standard deviation of $a$.

When publishing a table $T$, it is optimum to maintain the same original distribution over the set of equivalence classes. However it is natural that the distance between distributions will change, which causes privacy leakage. To minimize privacy leakage, we need to keep the distribution leakage among equivalence classes below a predetermined level.

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

$$\max(\mathscr{L}_D(S, [C_i])) \leq \varepsilon, \;\; i = 1, 2, \cdots, q.$$

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

$$\mathscr{L}_E(\mathscr{A}, \mathscr{B}) = \left| \sum_{i=1}^{m} a_i \log_2 \frac{1}{a_i} - \sum_{i=1}^{m} b_i \log_2 \frac{1}{b_i} \right|. \tag{3}$$

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

**Definition 9** (Entropy Leakage)**.** For an individual $u$ belonging to an equivalence class $[C_u]$, the entropy leakage is defined as

$$\mathscr{L}_E(S, [C_u]) = \left| \sum_{i=1}^{m} a_i \log_2 \frac{1}{a_i} - \sum_{i=1}^{m} x_i \log_2 \frac{1}{x_i} \right|. \tag{4}$$

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

$$\max(\mathscr{L}_E(S, [C_i])) \leq \alpha, \;\; i = 1, 2, \cdots, q.$$

Distribution leakage and entropy leakage are two different metrics since when the attribute distribution is a permutation of the original distribution, unless the original distribution is uniform, whatever the distribution leakage is, the entropy leakage will always be zero.

It remains an open problem how many metrics would be sufficient to quantify privacy. Nevertheless, we believe that any independent metrics can be added to the proposed framework to enhance privacy assessment.

## V. Empirical Analysis and Simulation Results

### A. Empirical Analysis

We use examples to show that the distribution and the entropy leakages are two different measures of privacy leakage.

TABLE III: Impatient Micro-data

(a) Original Dataset

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

(b) 4-anonymous, 2-diverse Dataset

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

**Example 1.** Consider a dataset $T$ with sensitive attribute $S$ containing $m = 3$ attribute values. The original attribute values distribution of $S$ is given as $\left(\frac{7}{12}, \frac{3}{12}, \frac{2}{12}\right)$. The published table $T'$ is divided into a set of $q = 3$ equivalence classes with attribute values distributions of $\left(\frac{3}{4}, \frac{1}{4}, 0\right)$, $\left(\frac{3}{4}, \frac{1}{4}, 0\right)$ and $\left(\frac{1}{4}, \frac{1}{4}, \frac{2}{4}\right)$. The distribution leakage $\mathscr{L}_D(S, [C])$ and the entropy leakage $\mathscr{L}_E(S, [C])$ for the attribute values are $\left[\frac{\sqrt{8}}{12}, \frac{\sqrt{8}}{12}, \frac{\sqrt{26}}{12}\right]$ and $[0.57, 0.57, 0.11]$ respectively. We notice that $[C_3]$ has the highest distribution leakage however it provides the least entropy leakage. This motivates us to further think of the implication of the large distribution leakage of $[C_3]$. The third attribute value is fully represented in this class. Therefore, an adversary has a 100% confidence that any individual that has the third attribute value is in $[C_3]$.

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

**Example 3.** Consider the original impatient dataset given in Table IIIa. The 4-anonymous, 2-diverse, 0.67-closeness impatient dataset is given in Table IIIb. For these two tables, the original probability distribution for the three diseases is $\left(\frac{7}{12}, \frac{3}{12}, \frac{2}{12}\right)$. In this case the EMD is $\left[\frac{1}{3}, \frac{1}{3}, \frac{2}{3}\right]$, distribution leakage $\mathscr{L}_D(S, [C])$ for individuals within the first, second and third equivalence classes is $\left[\frac{\sqrt{8}}{12}, \frac{\sqrt{8}}{12}, \frac{\sqrt{26}}{12}\right]$, while the entropy leakage is $[0.57, 0.57, 0.11]$, respectively.

Let's show how EMD in $t$-closeness is unreliable and insufficient to measure privacy leakage.

**Example 4.** For the original impatient dataset given in Table II [4], an 0.167-closeness w.r.t salary impatient dataset is given in Table Ib. The original distribution for the salaries in the published table $T'$ are given as $\{3K, 5K, 9K\}$, $\{6K, 11K, 8K\}$ and $\{4K, 7K, 10K\}$ for the three equivalence classes $[C_1], [C_2]$ and $[C_3]$. The EMD for the three classes is given as $[0.167, 0.167, 0.083]$. The EMD is a semantic metric. It gives a weight to the attribute values based on their sensitivity in the original distribution. However, it fails to give a correct measurement of the privacy leakage. For example, consider two possible equivalence classes. Assuming that the sensitive attribute values for two classes are $\{7K, 7K, 7K\}$ and $\{3K, 4K, 5K\}$, with EMD $[0.375, 0.278]$, which implies that $[C_2]$ achieves better privacy level than $[C_1]$. However, it is obvious that the adversarial general belief about the attribute values has changed more dramatically in $[C_2]$ compared to $[C_1]$. This change of belief is properly characterized in distribution leakage metric $\mathscr{L}_D(S, [C])$ which is given as $[0.22, 0.89]$. Furthermore, we can easily notice that $[C_2]$ suffers from a higher privacy leakage. An adversary would know the attribute value of an individual in this class with probability 1. This leakage is very well represented in our entropy leakage metric $\mathscr{L}_E(S, [C])$ which is given as $[0.158, 3.16]$.

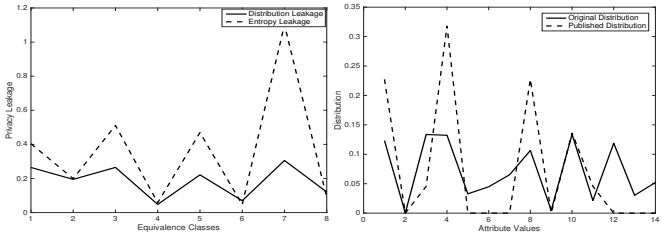
TABLE IV: Description of Adults Database

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

### *B. Simulation Results*

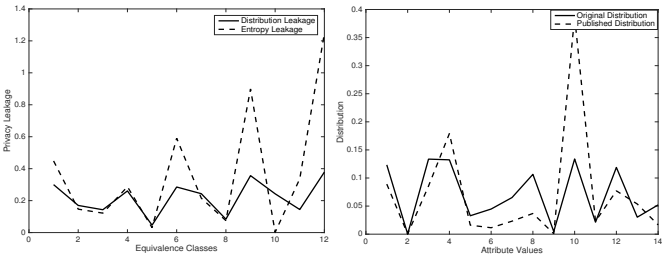
Simulations are done on a sample of the US census dataset from the UC Irvine machine learning repository [10]. After eliminating records with missing values, we have a total of 30,162 records. Following the work in [3], as shown in Table IV, we utilize only 9 attributes, 7 of which form the set of possible quasi-identifiers while Occupation and Salary form the set of possible sensitive attributes. We adopt the incognito algorithm [11] for generating the anonymized tables that satisfy the privacy measures of different PPDP techniques. Throughout the simulations, we take the Occupation as the sensitive attribute. The number of quasi-identifiers QIDs is represented by the variable $n$ that takes values from 1 to 7 with the same order in Table IV.

We consider a published table satisfying 0.5-closeness, 6-diversity, and $k \geq 6$-anonymity at $n = 2$. Quasi-identifiers are chosen to be Age and WorkClass where QID $= (1, 2)$. From Fig. 2a, an observed instance has a considerably high entropy leakage at $[C_7]$. This clearly identifies a major privacy leakage in the published table for users in class Age $= [75, 100]$, WorkClass $=$ Gov. Fig. 2b shows the original versus the published distribution of the sensitive attribute. It is obvious that $[C_7]$ has some missing attribute values. Hence, an observer can eliminate these values and thus gains an increased confidence about the sensitive attribute value of the user of




(a) Distribution and Entropy Leakage (b) Distribution of a Specific Class

Fig. 2: Evaluation of a table satisfying 0.5-closeness, 6-diversity, and $k \geq 6$-anonymity at $n = 2$




(a) Distribution and Entropy Leakage (b) Distribution of a Specific Class

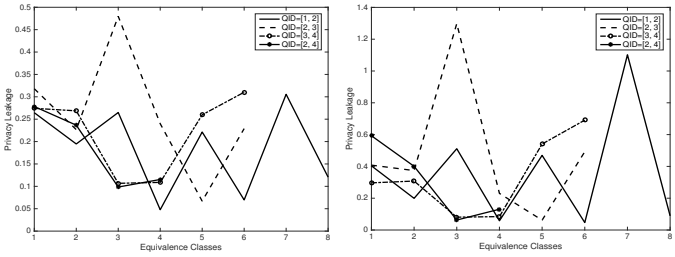
Fig. 3: Evaluation of a table satisfying 0.5-closeness, 7-diversity, and $k \geq 7$-anonymity at $n = 3$

interest. Specifically, an observer, a user falls in the age range has work class category $\mathsf{WorkClass} = \mathsf{Gov}$.

Based on the existing techniques explained earlier, a published table $T'$ satisfying 0.1-closeness, 13-diversity and $k \geq$ 13-anonymity at $n = 2$ is assumably privacy-preserving with these near optimum values of parameters for each PPDP technique. However, we find that there is a noticeable privacy leakage in $[C_2]$ since a user more than 50 years old will not have $\mathsf{Occupation}$ as $\mathsf{Armed\text{-}Forces}$. This privacy leakage could be captured use distribution and entropy leakage values of $[C_2]$, where $\mathscr{L}_D(S, [C]) = [0.0125, 0.0477]$ and $\mathscr{L}_E(S, [C]) = [0.0015, 0.0306]$. The increased distribution leakage is due to a fully non-represented attribute value in $[C_2]$ of the published table.

It is not necessarily an unrepresented attribute value that causes privacy leakage. Fig. 3b demonstrates the original versus the published distribution for $[C_6]$ of a published table satisfying 0.5-closeness, 7-diversity and $k \geq$ 7-anonymity at $n = 3$. We can see the accountable variation in published distributions of the $8^{th}$ and $10^{th}$ attribute values $[0.0369, 0.3871]$ compared to their original distribution $[0.1077, 0.1339]$.

In addition to comparing privacy leakage of different privacy levels of PPDP techniques, our work also provides a quite useful tool to compare data utility and privacy leakage of different combinations of chosen quasi-identifiers in PPDP techniques. In Fig. 4, we compare distribution and entropy leakages of the four tables at $n = 2$ while choosing a different combination of quasi-identifiers for each table, where quasi-identifiers are chosen to be $\mathsf{QID} = [(1, 2), (2, 3), (2, 4), (3, 4)]$. To satisfy the privacy conditions of the PPDP techniques, the anonymization process would decrease the number of classes in the published table and hence, the data utility decreases. Fig. 4 shows that anonymization process ended up with 8 classes at $\mathsf{QID} = (1, 2)$, 6 classes at $\mathsf{QID} = [(2, 3), (3, 4)]$, and




(a) Distribution Leakage (b) Entropy Leakage

Fig. 4: Leakage at different sets of QIDs

4 classes at $\mathsf{QID} = (2, 4)$. The figure illustrates the number of classes $q$ at each chosen combination and different levels of privacy represented in distribution and entropy leakage for each class. This tool gives an interesting option to adjust parameters by which a data publisher achieves the desirable privacy level with the requested data utility.

## VI. Conclusion

In this paper, we introduced comprehensive characterization and novel quantification methods of privacy in privacy-preserving data publishing. We presented data publishing as a multi-relational model. The proposed model and adversarial beliefs contribute to a more precise privacy characterization and quantification. Supported by insightful examples, we then showed that privacy could not be quantified based on a single metric. We proposed two different privacy leakage metrics. Based on these metrics, the privacy leakage of any given PPDP technique could be evaluated.

This work was supported in part by the National Science Foundation under Grant CCF-1919154 and Grant ECCS-1923409.

## References

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

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

[3] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "$\ell$-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, Mar. 2007.

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

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

[6] M. H. Afifi, K. Zhou, and J. Ren, "Privacy characterization and quantification in data publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, pp. 1756–1679, September 2018.

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

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

[9] I. Wagner and D. Eckhoff, "Technical privacy metrics: a systematic survey," *CoRR*, vol. abs/1512.00327, 2015.

[10] A. Asuncion and D. Newman, "Uci machine learning repository, http://www.ics.uci.edu/ mlearn/ml-repository.html, 2007,"

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