Privacy Characterization and Quantification in Data Publishing

M.H. Afifi, Kai Zhou and Jian Ren

Abstract—The increasing interest in collecting and publishing large amounts of individuals' data to public for purposes such as medical research, market analysis and economical measures has created major privacy concerns about individual's sensitive information. To deal with these concerns, many Privacy-Preserving Data Publishing (PPDP) techniques have been proposed in literature. However, they lack a proper privacy characterization and measurement. In this paper, we first present a novel multi-variable privacy characterization and quantification model. Based on this model, we are able to analyze the prior and posterior adversarial belief about attribute values of individuals. We can also analyze the sensitivity of any identifier in privacy characterization. Then we show that privacy should not be measured based on one metric. We demonstrate how this could result in privacy misjudgment. We propose two different metrics for quantification of privacy leakage, distribution leakage and entropy leakage. Using these metrics, we analyzed some of the most well-known PPDP techniques such as k-anonymity, l-diversity and t-closeness. Based on our framework and the proposed metrics, we can determine that all the existing PPDP schemes have limitations in privacy characterization. Our proposed privacy characterization and measurement framework contributes to better understanding and evaluation of these techniques. Thus, this paper provides a foundation for design and analysis of PPDP schemes.

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

1 Introduction

Nowadays, datasets are considered a valuable source of information for the medical research, market analysis and economical measures. These datasets can include information about individuals that contain social, medical, statistical, and customer data. Many organizations, companies and institutions publish privacy related datasets. While the shared dataset gives useful societal information to researchers, it also creates security risks and privacy concerns to the individuals whose data are in the table. To avoid possible identification of individuals from records in published data, uniquely identifying information such as names and social security numbers are generally removed from the table. While the obvious personal identifiers are removed, the quasi-identifiers such as zip-code, age, and gender may still be used to uniquely identify a significant portion of the population since the released data makes it possible to infer or limit the available options of individuals than would be possible without releasing the table. In fact, [1] showed that by correlating this data with the publicly available side information, such as information from voter registration list for Cambridge Massachusetts, medical visits about many individuals could be easily identified [2]. This study estimated that 87% of the population of the United States could be uniquely identified using quasi-identifiers through side information based attacks, including the medical records of the governor of Massachusetts in the medical data.

The spate of privacy related incidents has spurred a long line of research in privacy notions for data publishing and analysis, such as k-anonymity, l-diversity and t-closeness, to name a few [1], [3]–[9]. A table satisfies k-anonymity if each quasi-identifier attribute in the table is indistinguishable from at least $k-1$ other quasi-identifier attributes; such a table is called a k-anonymous table. While k-anonymity protects identity disclosure of individuals by linking attacks, it is insufficient to prevent attribute disclosure with side information. By combining the released data with side information, it makes it possible to infer the possible sensitive attributes corresponding to an individual. Once the correspondence between the identifier and the sensitive attributes is revealed for an individual, it may harm the individual and the distribution of the entire table. To deal with this issue, ℓ-diversity was introduced in [4]. ℓ-diversity requires that the sensitive attributes contain at least ℓ *well-represented* values in each equivalence class. As stated in [5], ℓ-diversity has two major problems. One, is that it limits the adversarial knowledge, while it is possible to acquire knowledge of a sensitive attribute from generally available global distribution of the attribute. Another problem is that all attributes are assumed to be categorical, which assumes that the adversary either gets all the information or gets nothing for a sensitive attribute.

In [5], authors propose a privacy notion called t-closeness. They first formalize the idea of global background knowledge and propose the base model t-closeness. This model requires the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). This distance was introduced to measure the information gain between the posterior belief and prior belief through the Earth Mover Distance (EMD) metric [10], which is represented as the information gain for a specific individual over the entire population. However, the value t is an abstract

The authors are with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824-1226. Email:{afifi, zhoukai, renjian}@msu.edu.

distance between two distributions that does not have any intuitive relation with privacy leakage. Moreover, as we show in this paper, the distance between two distributions cannot be easily quantified by a single measurement. t-closeness also has many limitations that will be described later. The state of the art PPDP techniques will be further analyzed in more details in section 3.

Research on data privacy has purely been focused on privacy definitions, such as k-anonymity, l-diversity, and t-closeness. While these models only consider minimizing the amount of privacy leakage without directly measuring what the adversary may learn, there is a motivation to find consistent measurements of how much information is leaked to an adversary by publishing a dataset.

In this paper, we begin by introducing our novel data publishing framework. The proposed framework consists of two steps. 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. This model is indeed a more precise model to describe privacy risk of publishing datasets.

For a given dataset, before it is released, we want 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. We show how considering only one metric ignoring the effect of the other strongly contributes to the information leakage and in turn affects the privacy. An intuitive example for this problem is reviewing a blood work. The medical status of a patient cannot be determined based on only one measure even if this particular measure is the most sensitive one. Instead, a physician has to review the relation between combinations of all measures in the blood work. We show that a minimized distribution leakage between sensitive attribute values distributions of the original and the published datasets does not essentially achieve the minimum entropy leakage that an adversary could gain. In fact, we show that distribution and entropy leakage are two different measures. We believe that for a published dataset to achieve better privacy, both metrics have to be taken into consideration.

The rest of this paper is arranged as follows. In Section 2 we introduce the general data publishing and adversarial attacks on datasets. In Section 3, we conduct a qualitative analysis of the existing PPDP techniques. Our privacy characterization framework and quantification metrics are proposed in Section 4 and 5. In Section 6, we demonstrate our empirical analysis and introduce the simulation results. We finally draw our conclusion and suggest some open problems for future work in Section 7.

2 Data Publishing and Attacks on Datasets

Privacy-Preserving Data Publishing Datasets publishing naturally consists of two phases. Different parties first collect data from record owners in a phase known as the data collection phase. It is then managed by the data publisher and is released in a phase known as the data publishing phase. This data is published to a certain data recipient for the purpose of data mining or to the public for the purpose of providing useful societal information that could be utilized in different areas including research.

Data is commonly published in two models, untrusted and trusted model. In the untrusted model, the data publisher attempts to extract or manipulate sensitive information about record owners. To avoid such attempts, record owners apply cryptographic operations on the published data to prevent the publisher from accessing sensitive information. In the trusted model, the data publisher is assumed to be honest. In this model, record owners are not concerned about uploading their record to the publisher. However, when data is released to the public, the publisher guarantees that sensitive information or identity of the record owner is not revealed to any possible adversary.

Utility-privacy Trade off Data utility is in a natural conflict with data privacy. It is trivial that, from the perspective of data utility, it is best to publish a dataset as is, while from the perspective of data privacy, it is best to publish a mostly generalized dataset or even an empty one. Although this is easy to understand, as far as we know, including the information theoretic approaches proposed in [11] and [12], there is not yet a tight closed form relationship that fully model the utility-privacy trade off. We believe that the first step on the track of finding such a relationship is to better characterize and quantify both sides of the trade off. We note that the importance of studying data utility is undeniable and of great value as it definitely contributes to resolving the trade off modeling. In this paper, we focus on the data privacy side.

Data Disclosure Model Data is usually released in the form of tables, where the rows are the records of individuals and columns are their corresponding attributes. Some of the attributes are for information only and not sensitive, while others are sensitive. For the information that is not being viewed as sensitive, when multiple records or maybe side information are combined, the individual maybe potentially identified. These attributes are generally referred to as *quasi-identifiers* QID, which may include information such as $\mathsf{Zip-Code}$, Age, and Gender. The sensitive information may include attributes that can uniquely identify the individuals such as the social security or the driving license numbers. These attributes are called *explicit-identifiers*. Another type of information being considered sensitive may include information such as disease and salary. When datasets are published, all explicit-identifiers are removed. Sensitive attribute disclosure occurs when the adversary learns information about an individual's sensitive attribute. This form of privacy breach is different and incomparable to learning whether an individual is included in the database, which is the focus of differential privacy [13].

Generalization and Anonymization As the original dataset contains abundant information that could help an adversary link records to certain individuals, datasets are not published before being modified. Modifications could be accomplished in many ways. Basically, all modifications are listed under the anonymization operations. These operations might be in the form of generalization, suppression, anatomization, permutation, or perturbation. In generalization and suppression [1], [14]–[19], values of quasi-identifiers are somehow relaxed in case of generalization, or suppressed in case of suppression, to increase the range of individuals that carry the same quasi-identifier values and therefore increase the uncertainty of a possible adversary about certain individual's record. On the other hand, anatomization and permutation operations achieve anonymization by dissociation of quasi-identifiers and sensitive attributes [20], [21]. Perturbation mainly adds some noise to the whole dataset based on the statistical properties of the original data [22]–[25].

However, unlike statistical databases [21], [26], publishing individuals' data, also known as micro-data, requires that data remains intact after being released. Therefore not all the previously mentioned techniques are good candidates for anonymization of micro-data. To keep data intact, and as much useful as possible, it is obvious that only generalization and suppression operations could be applied in privacy-preserving micro-data publishing techniques.

Attacks on Datasets Generally, there are two types of attacks on datasets, *record linkage* and *attribute linkage*. The record linkage occurs when some values of quasi-identifier attributes can lead to the identification of a smaller number of records in the published dataset. In this case, an individual having these attribute values is vulnerable to being linked to a limited number of records. On the other hand, attribute linkage occurs if some sensitive values are predominate in a group, where an attacker has no difficulty to infer such sensitive values for the record owner belonging to this group.

Attribute linkage mainly consists of two types, *homogeneity and background knowledge attacks*. In homogeneity attacks, protection model may create groups that leak information due to lack of diversity in the sensitive attribute. In fact, some protection process is based on generalizing the quasi-identifiers but does not address the sensitive attributes that can reveal information to an attacker. In background knowledge attacks, an attacker can have prior knowledge that enables him to guess sensitive data with high confidence. These kinds of attacks depend on other information available to an attacker. Using this background knowledge, an adversary can disclose information in two ways, *positive and negative disclosure.* In positive disclosure, an adversary can correctly identify the value of a sensitive attribute with high probability. On the other hand, in negative disclosure, the adversary can correctly eliminate some possible values of sensitive attribute with high probability. We also note that a background knowledge attack is difficult to prevent as compared to homogeneity attack.

In the next section we introduce a thorough analysis of the existing privacy-preserving data publishing techniques that attempt to combat these types of attacks on privacy.

3 Analysis of The Existing PPDP Schemes

In this section, some representative PPDP schemes will be analyzed.

TABLE 1: Patient Table with the Original Distribution Maintained

(a) Original Table

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	49	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

(b) A 3-anonymous Version

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

3.1 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; such a table is called a k-anonymous table. To satisfy this condition, before being published, the original table is generalized forming groups that share values of QIDs. Each group, named as an *equivalence class* $[C]$, shares the same combination of quasi-identifiers and has at least k records. The idea of k-anonymity was proposed to combat record linkage attacks. In [20], [27], [28], authors show that k-anonymity does not provide sufficient protection against attribute linkage.

Example 1 (Homogeneity and Background Knowledge Attacks)**.** Table 1a represents the original data table and Table 1b is an anonymized version of it satisfying 3-anonymity. The Disease attribute is sensitive. Suppose Alice knows that Bob is a 27-year old man living in $\mathsf{Zip}-\mathsf{Code} = 47678$ and Bob's record is in the table. From Table 1b, Alice can conclude that Bob is the owner of one of the first three records, and thus, must have $\mathsf{Heart}-\mathsf{Disease}$. This is the homogeneity attack. For an example of the background knowledge attack, suppose that by knowing Carl's Age and $\mathsf{Zip}-\mathsf{Code}$, Alice can conclude that Carl corresponds to a record in the last equivalence class in Table 1b. Furthermore, suppose that Alice knows that Carl has a very low risk for $\mathsf{Heart}-\mathsf{Disease}$. This background knowledge enables Alice to conclude that Carl most likely has cancer.

To address the limitations of k-anonymity, [27] introduced l-diversity as a stronger notion of privacy.

3.2 l-diversity

An equivalence class is said to have l-diversity if there are at least l *well-represented* values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity. l-diversity represents an important step beyond k-anonymity in protecting against attribute linkage. However, it is susceptible to attacks such

TABLE 2: Original Salary/Disease

(a) Original Dataset

	Zip Code	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

	Zip Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

as skewness and similarity attacks. As shown in [5], when the overall distribution is skewed, satisfying the l-diversity does not prevent attribute linkage. Consider the following example:

Example 2 (Skewness Attack)**.** Suppose that the original dataset has only one sensitive attribute, which is the test result for a particular virus. The virus takes two values either positive or negative. For a table that has 10,000 records, with 99% of them being negative and only 1% being positive. To satisfy distinct 2-diversity, any equivalence class $[C]$ must carry the two attribute values. If one of the equivalence classes has an equal number of positive and negative records, although it is 2-diverse, it presents a serious privacy risk. Any individual in this class has probability 50% to be infected compared to a 1% of the whole original population. Now, consider another extreme case. An equivalence class that has 49 positive records and only 1 negative record. Any individual in the equivalence class is 98% positive, compared to 1% of the whole original population.

When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information.

Example 3 (Similarity Attack)**.** In the original Table 2a and an anonymized version satisfying distinct 3-diversity Table 2b, consider Salary and Disease as the two sensitive attributes. An adversary is interested in finding the sensitive attribute value of an individual u. Based on the quasi-identifier values of u, the adversary is able to determine that the individual belongs to the first equivalence class. Therefore he knows that his salary is in the range $[3K, 5K]$. This also applies to categorical attribute such as the Disease. The adversary would also know that the individual of interest indeed has a stomach-related disease. This leakage of sensitive information occurs because l-diversity does not take into account the semantical closeness of attribute values.

To prevent such attacks, Li *et al.* [5] proposed a privacy model, known as t-closeness.

TABLE 3: 0.167-Closeness w.r.t. Salary and 0.278-Closeness w.r.t. Disease

	Zip Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
2	4767*	≤ 40	5K	stomach cancer
3	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	4760*	≤ 40	4K	gastritis
8	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

3.3 t-closeness

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. A table is said to have t-closeness if all equivalence classes have t-closeness. The distance used in this publishing technique is the earth mover's distance. EMD is simply the minimal amount of work needed to transform one distribution to another by moving distribution mass between each of them. Table 3 shows another anonymized version of Table 2a. It has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease. The similarity attack is prevented in Table 3. Revisiting Example 3, Alice can neither infer that Bob has a low salary nor he has stomach-related diseases.

In k-anonymity and l-diversity, given a k or l value, the data publisher will have some intuition on its practical meaning in the real application and hence can effectively choose k and l values to process the dataset. Unlike k-anonymity and l-diversity, in t-closeness, the value t is merely an abstract distance between two distributions, that could have different meanings in different contexts. t-closeness also has several other limitations [29]. First, it does not offer the flexibility of having different protection levels for different sensitive attribute values. Second, the EMD function, used to measure the distance between distributions, is not suitable for protection against attribute linkage on numerical sensitive attributes [30]. Third, as the case for l-diversity, enforcing t-closeness would greatly affect the data utility where it requires the distribution of sensitive attribute values to be the same in all q equivalence classes. This would significantly damage the correlation between the set of *quasi-identifiers* QID and sensitive attributes. Finally, and the most important, we believe that the distance t measured as the EMD is unreliable to quantify the amount of privacy leakage. More specifically, if we have two published tables T_1' and T_2' with $t_1 < t_2$, then table T_1' is not necessarily more privacy-preserving than T_2'. In other words, two published classes might have the same EMD distance relative to an original distribution, however they correspond to different levels of privacy leakage. Consider the following example:

Example 4. A medical dataset has the Disease as a sensitive attribute. Distribution of attribute values Cancer, *Heart-Disease* and Flu in the original table is $(0.1, 0.5, 0.4)$. The published table is divided into two equivalence classes, denoted as $[C_1]$ and $[C_2]$. In $[C_1]$, distribution of attribute values is given as $(0.2, 0.4, 0.4)$, while in $[C_2]$ the distribution is $(0, 0.6, 0.4)$. This table achieves an 0.1-closeness w.r.t. Disease. Although the EMD in the two equivalence classes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2797092, IEEE Transactions on Knowledge and Data Engineering

is the same, it is obvious that attribute values of individuals in $[C_2]$ are more prone to be inferred.

In section 6, supported by an analyzed example on t-closeness, we show how our privacy metrics enable us to deliberately characterize and quantify this leakage.

4 Our Proposed Publishing Model and Privacy Characterization

All previous approaches to characterize and quantify privacy have only investigated the privacy risk of publishing a sensitive attribute by focusing only on the change of belief of an adversary about the probability distribution of this attribute. However, we believe that any attribute by itself is not sensitive. The sensitivity of an attribute comes from combining it with other attributes. For example, cancer in a medical records dataset, high or low salaries in an employees dataset, are not sensitive unless they are linked to a certain geographical area, age-range or race. To obtain a meaningful definition of data privacy, it is necessary to characterize and quantify the knowledge about sensitive attributes that the adversary gains from observing the published dataset taking into consideration the combinational relation of different attributes. In our approach to characterize privacy, we employ a multi-dimensional scheme of privacy risk analysis attached with combining different attributes. Thus, we introduce the following combinational characterization of privacy.

In our privacy characterization we assume that any individual in a given table T only owns one record. Thus we, interchangeably, use the notion u to represent the record or the record owner. We also assume that any record is 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 order of the corresponding attribute A_i, that is the number of all possible attribute values. As previously mentioned, to maintain the utility of a published table T', different attributes of all individuals are expected to be disclosed. However, privacy-preserving techniques apply some generalizations and suppressions to the quasi-identifiers QID to avoid linking individuals to records in the table. For any record, each value is typically generalized before being published. Any value in the original table is basically mapped to a generalized value in the published table following a certain mapping function that will be defined below. Records are generalized to be represented as functions of multi-variables $v' = \{v'_1, v'_2, \cdots, v'_l\}$, where v' is the generalization of v. The order of each generalized variable v'_i is defined as $\mathsf{ord}(v'_i) = |v'_i|$.

Consider two tables (T, T'), their corresponding attributes (v, v') and a mapping function $f :\ T \to T'$. We define table generalization as follows:

Definition 1 (Table Generalization). For (T, T') and (v, v'), table generalization is a mapping $f :\ T \to T'$ that maps any table T to a table T'. This mapping function implies 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.

Table generalization, represented in the mapping function, is the tool that controls privacy level of individuals and data utility of the published dataset. This mapping function is the key for designing any data publishing technique. Furthermore, privacy leakage is directly linked to the combination of different variables. Hence, any publishing technique should consider the privacy risk attached with the combination of any of these variables.

Publishing a table T' gives different privacy risks for each combination of the generalized variables $< v'_i, v'_j >$. For example, $< \mathsf{Age}, \mathsf{Disease} >$ mapped to $< v'_1, v'_2 >$ is a combination of two variables that represents privacy risk of individuals of specific Age (age-range) and suffering from a specific $\mathsf{Disease}$, while $< \mathsf{Zip-Code}, \mathsf{Salary} >$ mapped to $< v'_3, v'_4 >$ is a combination that represents privacy risk of individuals living at a certain geographical area and are paid certain salary. Similarly, $< \mathsf{Age}, \mathsf{Zip-Code}, \mathsf{Salary}, \mathsf{Disease} >$ mapped to $< v'_1, v'_3, v'_4, v'_2 >$ represents the risk of individuals with certain $Age\ v'_1$, living at certain location with $\mathsf{Zip-Code}\ v'_3$, suffering from $\mathsf{Disease}\ v'_2$ and are paid an annual $\mathsf{Salary}\ v'_4$. 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. The order of any combination of variables could be easily derived as $\prod_{i=1}^{l} |v'_i|$, where l is the number of combined variables.

The adversary is given the published table T' generated from an original table T, and assumed to know quasi-identifier values $u[QID]$ of an individual u of interest. The individual of interest is assumed to be in the table with probability 1. Hence, the membership disclosure problem, i.e., learning whether a given individual is present in the published dataset, is a different, incomparable privacy property and is out of the scope of this paper. In our approach of characterizing privacy, an adversary is generally assumed to be aware of all the public information that might be available. Therefore, an adversary is believed to possess the original distribution of all the attributes. Moreover, for a dataset with l attributes, while some attributes are entirely independent, others could be correlated. Thus, an adversary possibly has an estimate of the joint distributions of these attributes. We now introduce the definition of the adversarial prior belief, that is the general public belief of all the distributions of attributes combinations.

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

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

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

For example, the distribution of a population over attributes such as Gender, Age and Disease is publicly available. Typically, within any geographical location, information such as percentage of males and females, percentages of individuals lying in a specific age-range and percentage of population suffering from a specific disease, are considered as adversarial prior information. Moreover, based on general trivial information, an adversary could have a very good estimate of joint distributions of some attributes. For instance, individuals suffering from a disease such as Breast Cancer are generally much more likely to be females, while individuals suffering from diseases such as Alzheimer and Arthritis are more likely to be above 60. Similarly, individuals living in a richer neighborhood are more likely to be paid higher salaries. We believe that any adversarial model should take such information into consideration. Consequently, any privacy quantification approach that ignores this adversarial knowledge is not precise and lacks sufficiency. Any further information gained by an adversary after observing a published table is considered privacy loss and is represented as the adversarial posterior belief and is defined as follows,

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

$$x_{v_i,v_j} = P(v_i \mid v_j), i = 1, \cdots, \mathsf{ord}(v_i'),\ j = 1, \cdots, \mathsf{ord}(v_j'),$$

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

As any published table T' is eventually formed of a subset of all possible combinations of generalized attributes v_i', each of these combinations represents an equivalence class $[C_i]$. While a_{v_i,v_j} and x_{v_i,v_j} represent the prior and posterior belief of an adversary about an attribute v_i given an attribute v_j. We are generally interested in $a_{v_i,[C_i]}$ and $x_{v_i,[C_i]}$, that are the prior and posterior beliefs of an adversary about an attribute v_i given a combination of generalized attributes represented in the specific class $[C_i]$ they form.

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 $L(v_i \mid [C_u])$ is typically the change of an adversarial belief about an attribute's distribution from $a_{v_i,[C_u]}$ to $x_{v_i,[C_u]}$.

Consider, an original table T having only 100 records described over two attributes, Age and Disease. If $P(5^*) = \frac{1}{4}$, $P(\mathsf{Cancer}) = \frac{1}{10}$ and an adversary has an estimate of their joint probability to be $\tilde{P}(\mathsf{Cancer} \cap 5*) = \frac{4}{100}$, then the estimated conditional probability $\tilde{P}(\mathsf{Cancer} \mid 5*) = \frac{\tilde{P}(\mathsf{Cancer} \cap 5*)}{P(5*)} = \frac{16}{100}$. However, in the published table T', the adversary observes that the published joint probability $P(\mathsf{Cancer} \cap 5*) = \frac{7}{100}$, which gives the published conditional probability $P(\mathsf{Cancer} \mid 5*) = \frac{P(\mathsf{Cancer} \cap 5*)}{P(5*)} = \frac{28}{100}$. Now an adversarial belief about individuals of the age-range (5∗) and suffering from Cancer has changed from a prior belief of 16/100 to a posterior belief of 28/100. This change of belief is the amount of information gained by an adversary. That is, the amount of privacy loss of individuals in a specific class (5∗) and having a certain attribute value (Cancer). Similarly we can find the privacy loss of individuals having other attribute values (other diseases) within the same class. One of our goals is to quantify this loss. In the next section, we propose two privacy metrics that are able to measure privacy leakage from two different perspectives.

5 Our Proposed Privacy Quantification

There is an immense amount of existing privacy loss quantification metrics in literature [7]. The state-of-the-art approaches to measure privacy can be mainly sub-categorized into uncertainty, information gain or loss, similarity and diversity, and indistinguishability metrics. *Uncertainty metrics* measure the uncertainty in the adversarial estimate. The more uncertain the adversary is, the higher the achieved privacy in the published dataset. *Information gain or loss metrics* quantify the amount of information gained by the adversary, or the amount of information lost by users after data publishing. High adversarial gain and high user's loss of information corresponds to low privacy. *Similarity and diversity metrics* measure the similarity or diversity between the original and the published dataset. High similarity or low diversity between the two datasets corresponds to low privacy. *Indistinguishability* measures the ability of an adversary to distinguish between two outcomes of a privacy preserving data publishing technique. Privacy is high if it is hard for an adversary to distinguish between any pair of outcomes.

Our approach to quantify privacy mainly depends on understanding when information leakage happens and how this leakage could be measured. To have a better understanding of when leakage occurs, we revisit the two states of knowledge of an adversary before and after a table T is published. At the first state of knowledge, based on public information of sensitive attribute's distribution, an adversary has some prior belief about the attribute value of an individual. This prior belief is in the form of probability distributions of attributes and joint distributions of their combinations. After publishing the table, an adversary moves to the second state of knowledge to gain some more information about the individual. This amount of information is the leakage that we need to capture where it enables us to measure the extent to which this data-publishing model minimizes privacy leakage. We now analyze this leakage and find a set of appropriate metrics that contribute to a better quantification of privacy represented in the amount of uncertainty an adversary has about an individual's sensitive attribute value after a table is published.

We fix a finite set $U = \{u_1, u_2, \cdots, u_n\}$ of n individuals participating in the data table T. Let $\mathcal{A} = \{A_1, A_2, \cdots, A_l\}$ be the set of l attributes and $u[A_i]$ denotes the value of attribute A_i for individual u. In addition, we define the sensitive attribute $S \subset \mathcal{A}$ as the attribute of interest for an adversary. We fix a finite set $S = \{s_1, s_2, \cdots, s_m\}$, representing the set of m possible values for the sensitive attribute (e.g. disease-name in a medical dataset). We denote the set of *quasi-identifiers* as $QID \subset \mathcal{A}$. Two individuals u_i and u_j having the same values of *quasi-identifiers* are assumed to be QID-equivalent if $u_i[QID] = u_j[QID]$, i.e., they share the same equivalence class $[C]$. As previously mentioned, we are generally interested in $a_{v_i,[C_i]}$ 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. We note that all of the concepts in this paper are easily explained in the single sensitive attribute setting, but can also be extended to multiple sensitive attributes.

We reemphasize that in our privacy quantification approach we only consider the leakage between the two knowledge states. For example, if 10% of the individuals in a medical record's table T have HIV, then it should not be counted as leakage. That is, if at the second state of knowledge, the adversary finds that an individual u has HIV with probability 10%, the information leakage for this scheme should be zero since this sensitive attribute's distribution is always considered public. Based on this, we introduce a generic definition of privacy-preserving data publishing as follows.

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 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 each $u_i \in U$ represents an individual from the population, $p(u_i[A_j])$ 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.

In this definition, to be considered as privacy-preserving, the publishing technique strictly prohibits any privacy leakage in the published data. While this conservative definition is practically impossible to achieve, any publishing technique should attempt to be as close as possible to achieve it. Privacy leakage should, therefore, be quantified to be able to decide how far any given published data is from being privacy-preserving. Before we introduce our two proposed privacy quantification metrics, namely, distribution and entropy leakage, we show the reason behind adopting these metrics.

The intuitive expectation of the proposed metrics is to compute the change in the data user's belief about an individual's sensitive attribute value before and after data disclosure. To find suitable metrics, we seek the distance measures based on two criteria. First is the *sensitivity* meaning that the metric should be sensitive to variations in the distributions. Second is the *independence* meaning that the metrics should be independent. That is, if two metrics independently measure the distance between two distributions, they both contribute to two different types of leakages. 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. From Fig. 1 we can also see that entropy distance is the only distance that is independent of the other metrics. This qualifies it to be the second metric.

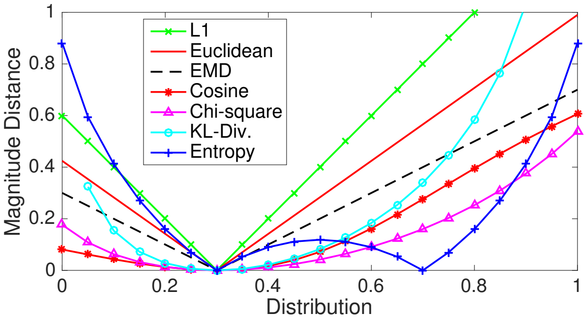


Fig. 1: Comparison of different metrics.

We now introduce these two proposed privacy metrics and show how they are able to quantify privacy leakage from two different perspectives.

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 given as $a = (a_1, a_2, \cdots, a_m)$. The published distribution of S given an equivalence class $[C_u]$ is given 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

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

Since it is a Euclidean distance function, the distribution leakage $\mathscr{L}_D(S, [C_u])$ defined above is indeed a distance metric, i.e. it satisfies all metric conditions. The distribution leakage could be viewed as a measure of the overall divergence of attribute values distribution from one state to the other. Generally, any privacy-preserving publishing technique modifies the original dataset into a set of equivalence classes. Leakage is measured between the original distribution of sensitive attribute values in the original and the published dataset for each given equivalence class.

As some privacy-preserving publishing techniques, falsely, assume that a uniform published distribution of attribute values achieves optimal privacy. It is interesting to find the distribution leakage for this specific scenario. We find that distribution leakage is closely related to the standard deviation.

Theorem 1. Let $S = \{s_1, s_2, \cdots, s_m\}$ be the set of all attribute values of a given dataset and $a = (a_1, a_2, \cdots, a_m)$ be the corresponding probability distribution. An individual u, belonging to an equivalence class $[C_u]$, has probability distribution on S of $x = (x_1, x_2, \cdots, x_m)$. The distribution leakage of an attribute S in the published table T' with respect to uniform distribution is

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

where σ_a is the standard deviation of a.

Proof: While the distribution leakage is given as

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

$$\sigma_a = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(a_i[S] - \frac{\sum_{i=1}^{m} a_i[S]}{m} \right)^2}.$$

Since $\sum_{i=1}^{m} a_i[S] = 1$, then,

$$\sigma_a = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(a_i[S] - \frac{1}{m} \right)^2}.$$

Therefore we have

$$\mathscr{L}_D(S, [C_u]) = \sigma_a \sqrt{m}. \qquad \square$$

$\mathscr{L}_D(S, [C_u])$ reaches a minimum value of 0 when $a_i[S] = \frac{1}{m}$. It reaches a maximum value of $\sqrt{\frac{m-1}{m}}$ when $a_i[S] = 1$ for some attribute i and 0 for other attributes. It is obvious that the distribution leakage depends on the standard deviation of the original distribution. Thus, a uniform distribution of published attribute values is not essentially optimal. We believe that matching the published distribution to the original estimated distribution would indeed achieve better privacy.

When publishing a table T, it is optimum to maintain the same original distribution over the set of equivalence classes. That is, distribution leakage $\mathscr{L}_D(S, [C])$ is desired to be zero. However it is natural that the distance between distributions will change. This change contributes to the privacy leakage. Therefore, an objective of minimizing privacy leakage is to keep the distribution leakage among equivalence classes below a predetermined level.

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

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

While the distribution leakage captures the amount by which privacy of an attribute is leaked, it does not give a sufficient implication about privacy leakage of individuals carrying different attribute values. Specifically, a small distribution leakage in the published table might lead to a critical decrease in the amount of uncertainty of an adversary about the attribute value of a certain individual of interest. This motivates us to think of an information theoretic metric that would capture this change of adversarial uncertainty before and after table publishing. Hence, we propose the following privacy metric.

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

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

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. We now give the following theorem about entropy distance.

Theorem 2 (***Triangle Inequality***)**.** For the proposed entropy distance, the triangle inequality holds true, that is

$$\mathscr{L}_E(\mathscr{A}, \mathscr{B}) + \mathscr{L}_E(\mathscr{B}, \mathscr{C}) \geq \mathscr{L}_E(\mathscr{A}, \mathscr{C}).$$

Proof: We split the proof into four cases.

Case 1: $\sum_{i=1}^{m} a_i \log_2 a_i \leq \sum_{i=1}^{m} b_i \log_2 b_i$ and $\sum_{i=1}^{m} b_i \log_2 b_i \leq \sum_{i=1}^{m} c_i \log_2 c_i$. Then we have $\sum_{i=1}^{m} a_i \log_2 a_i \leq \sum_{i=1}^{m} c_i \log_2 c_i$, and

$$\begin{aligned}
&\mathscr{L}_E(\mathscr{A}, \mathscr{B}) + \mathscr{L}_E(\mathscr{B}, \mathscr{C}) \\
&= -\sum_{i=1}^{m} a_i \log_2 a_i + \sum_{i=1}^{m} b_i \log_2 b_i - \sum_{i=1}^{m} b_i \log_2 b_i + \sum_{i=1}^{m} c_i \log_2 c_i \\
&= \left| -\sum_{i=1}^{m} a_i \log_2 a_i + \sum_{i=1}^{m} c_i \log_2 c_i \right| = \mathscr{L}_E(\mathscr{A}, \mathscr{C}).
\end{aligned}$$

Case 2: $\sum_{i=1}^{m} a_i \log_2 a_i \geq \sum_{i=1}^{m} b_i \log_2 b_i$ and $\sum_{i=1}^{m} b_i \log_2 b_i \geq \sum_{i=1}^{m} c_i \log_2 c_i$. Then we have $\sum_{i=1}^{m} a_i \log_2 a_i \geq \sum_{i=1}^{m} c_i \log_2 c_i$, and

$$\begin{aligned}
&\mathscr{L}_E(\mathscr{A}, \mathscr{B}) + \mathscr{L}_E(\mathscr{B}, \mathscr{C}) \\
&= \sum_{i=1}^{m} a_i \log_2 a_i - \sum_{i=1}^{m} b_i \log_2 b_i + \sum_{i=1}^{m} b_i \log_2 b_i - \sum_{i=1}^{m} c_i \log_2 c_i \\
&= \left| -\sum_{i=1}^{m} a_i \log_2 a_i + \sum_{i=1}^{m} c_i \log_2 c_i \right| = \mathscr{L}_E(\mathscr{A}, \mathscr{C}).
\end{aligned}$$

Case 3: $\sum_{i=1}^{m} a_i \log_2 a_i \leq \sum_{i=1}^{m} b_i \log_2 b_i$ and $\sum_{i=1}^{m} b_i \log_2 b_i \geq \sum_{i=1}^{m} c_i \log_2 c_i$. Then we have $\prod_{i=1}^{m} a_i^{a_i} \leq \prod_{i=1}^{m} b_i^{b_i}$, and $\prod_{i=1}^{m} b_i^{b_i} \geq \prod_{i-1}^{m} c_i^{c_i}$, and

$$\begin{aligned}
&\mathscr{L}_E(\mathscr{A}, \mathscr{B}) + \mathscr{L}_E(\mathscr{B}, \mathscr{C}) \\
&= -\sum_{i=1}^{m} a_i \log_2 a_i + \sum_{i=1}^{m} b_i \log_2 b_i + \sum_{i=1}^{m} b_i \log_2 b_i - \sum_{i=1}^{m} c_i \log_2 c_i \\
&= \log \left(\prod_{i=1}^{m} \frac{b_i^{b_i}}{a_i^{a_i}} \cdot \frac{b_i^{b_i}}{c_i^{c_i}} \right) \geq \log \left(\prod_{i=1}^{m} \frac{b_i^{b_i}}{a_i^{a_i}} \right) \geq \log \left(\prod_{i=1}^{m} \frac{c_i^{c_i}}{a_i^{a_i}} \right) \\
&= -\sum_{i=1}^{m} a_i \log_2 a_i + \sum_{i=1}^{m} c_i \log_2 c_i.
\end{aligned}$$

Similarly, we also have

$$\begin{aligned}
&\mathscr{L}_E(\mathscr{A}, \mathscr{B}) + \mathscr{L}_E(\mathscr{B}, \mathscr{C}) \\
&\geq \log \left(\prod_{i=1}^{m} \frac{a_i^{a_i}}{c_i^{c_i}} \right) = \sum_{i=1}^{m} a_i \log_2 a_i - \sum_{i=1}^{m} c_i \log_2 c_i.
\end{aligned}$$

Therefore, we have

$$\mathscr{L}_E(\mathscr{A},\mathscr{B})+\mathscr{L}_E(\mathscr{B},\mathscr{C})\geq\mathscr{L}_E(\mathscr{A},\mathscr{C}).$$

Case 4: $\sum_{i=1}^m a_i\log_2 a_i\geq\sum_{i=1}^m b_i\log_2 b_i$ and $\sum_{i=1}^m b_i\log_2 b_i\leq\sum_{i=1}^m c_i\log_2 c_i$. Then we have $\prod_{i=1}^m a_i^{a_i}\geq\prod_{i=1}^m b_i^{b_i}$, and $\prod_{i=1}^m b_i^{b_i}\leq\prod_{i-1}^m c_i^{c_i}$, and

$$\begin{aligned}&\mathscr{L}_E(\mathscr{A},\mathscr{B})+\mathscr{L}_E(\mathscr{B},\mathscr{C})\\&=\sum_{i=1}^m a_i\log_2 a_i-\sum_{i=1}^m b_i\log_2 b_i-\sum_{i=1}^m b_i\log_2 b_i+\sum_{i=1}^m c_i\log_2 c_i\\&=\log\left(\prod_{i=1}^m\frac{a_i^{a_i}}{b_i^{b_i}}\cdot\frac{c_i^{c_i}}{b_i^{b_i}}\right)\geq\log\left(\prod_{i=1}^m\frac{a_i^{a_i}}{b_i^{b_i}}\right)\geq\log\left(\prod_{i=1}^m\frac{a_i^{a_i}}{c_i^{c_i}}\right).\\&=\sum_{i=1}^m a_i\log_2 a_i-\sum_{i=1}^m c_i\log_2 c_i.\end{aligned}$$

Similarly, we also have

$$\begin{aligned}&\mathscr{L}_E(\mathscr{A},\mathscr{B})+\mathscr{L}_E(\mathscr{B},\mathscr{C})\\&=\log\left(\prod_{i=1}^m\frac{a_i^{a_i}}{b_i^{b_i}}\cdot\frac{c_i^{c_i}}{b_i^{b_i}}\right)\geq\log\left(\prod_{i=1}^m\frac{c_i^{c_i}}{b_i^{b_i}}\right)\geq\log\left(\prod_{i=1}^m\frac{c_i^{c_i}}{a_i^{a_i}}\right)\\&\geq-\sum_{i=1}^m a_i\log_2 a_i+\sum_{i=1}^m c_i\log_2 c_i.\end{aligned}$$

Therefore, we have

$$\mathscr{L}_E(\mathscr{A},\mathscr{B})+\mathscr{L}_E(\mathscr{B},\mathscr{C})\geq\mathscr{L}_E(\mathscr{A},\mathscr{C}).\qquad\square$$

Based on Theorem 2, we have the following theorem.

Theorem 3. Entropy Leakage is a distance metric and has the following properties:

1) **Non-negativity:** $\mathscr{L}_E(x,y)\geq 0$.
2) **Definiteness:** $\mathscr{L}_E(x,y)=0$ if and only if $x=y$
3) **Symmetry:** $\mathscr{L}_E(x,y)=\mathscr{L}_E(y,x)$.
4) **Triangle inequality:** $\mathscr{L}_E(x,z)\leq\mathscr{L}_E(x,y)+\mathscr{L}_E(y,z)$.

The proof of this theorem is straight forward. We note that maximum entropy of attribute values in the published dataset does not achieve the maximum privacy. The maximum entropy corresponds to the uniform distribution of attribute values. This kind of distribution can be optimum if the background information of an adversary is ignored. However, given that an adversary has some prior belief about original attribute values distributions, it is best to maintain the same entropy level after publishing. Therefore, we introduce the following metric.

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

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

We define the entropy leakage of an individual as the entropy leakage of the equivalence class that the individual belongs to. Note that the entropy leakage reaches maximum $\log_2 m$ when the original distribution is uniform and the published distribution is $x_i[S]=1$ for some attribute i and 0 for other attributes. This is easily explained as a transition in the adversarial belief, from a state where the uncertainty about the attribute value of an individual of interest in a given class is maximum, to a state where he becomes 100% confident about the attribute value of this individual.

Note that the entropy leakage metric is convex in our case since $\sum_{i=1}^m a_i\log_2\frac{1}{a_i}$ a fixed number that can be computed based on the prior knowledge ahead of time. Therefore, the optimal value always exists theoretically. The maximum value of equation (1) is $\left|\sum_{i=1}^m x_i\log_2\frac{1}{x_i}-Z\right|$ is $\log_2 m-Z$.

As previously mentioned, many PPDP techniques assume that a uniform published distribution of attribute values achieves optimal privacy. Thus, we also find the entropy leakage for this specific scenario. For a published uniform distribution x of a dataset, the entropy is $\log_2 m$. Using Definition 9, the entropy leakage between the original dataset distribution and the uniform distribution is given as $\log_2 m-\sum_{i=1}^m a_i\log_2\frac{1}{a_i}$.

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

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

We argue that distribution leakage and entropy leakage are two different metrics. To justify this argument, let us assume the case when the attribute values 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. More examples are presented in the next section to support our argument.

Given the knowledge of the distribution of sensitive attribute values of the original distribution, an adversary has a certain level of uncertainty about individuals attribute values. Any change in this level of uncertainty is considered a leakage. An objective of any data publishing technique would be minimizing this leakage.

Meanwhile, we do not know how many metrics would be sufficient to quantify privacy. However, we believe that any further proposed independent metrics that would contribute to reaching an optimum and provably sufficient set of measures, can be added to the proposed quantitative measurement framework.

6 Empirical Analysis and Simulation Results

This section is divided into two parts. In the first part, based on our findings, we introduce a wide set of empirical examples for different case scenarios that support our findings. The provided examples aim to help understand the implications of the proposed metrics and show how these metrics contribute to analyzing, comparing and evaluating the previously mentioned existing privacy-preserving data-publishing techniques. In the second part of this section, aided with our simulation results, we focus on instances where different PPDP techniques assume to achieve an intended privacy level. However, based on our proposed metrics, they fail to express, and therefore fail to avoid, a considerable amount of privacy leakage. Throughout this section, we assume that an adversary has no other side information about dataset statistics or the user of interest other than the determined quasi-identifier values.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2797092, IEEE Transactions on Knowledge and Data Engineering

6.1 Empirical Analysis

We begin by giving examples to show how the distribution and the entropy leakages are two different measures of privacy leakage.

Example 5. 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. It is obvious that a high distribution leakage does not necessarily provide a high entropy leakage and vice versa. 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 $[C_3]$.

While a published table with uniform attribute values distribution naturally has the highest output entropy (not entropy leakage), it is sometimes, falsely, assumed to be optimal. We believe that matching the published distributions with the original distribution would certainly achieve a better privacy protection. This can be justified using the following example.

Example 6. 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.

Another example that shows how a uniform distribution of published attribute values for an equivalence class might, at some cases, even give a large distance and a worse entropy leakage than other distributions.

Example 7. Consider a dataset T with sensitive attribute S containing $m = 4$ attribute values. The original attribute values distribution of S is given as $\left(\frac{9}{16}, \frac{3}{16}, \frac{2}{16}, \frac{2}{16}\right)$. The published table T' is divided into a set of $q = 4$ equivalence classes with attribute values distributions of $\left(\frac{4}{16}, \frac{4}{16}, \frac{4}{16}, \frac{4}{16}\right)$, $\left(\frac{8}{16}, \frac{4}{16}, \frac{4}{16}, 0\right)$, $\left(\frac{12}{16}, \frac{4}{16}, 0, 0\right)$, 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 $\left[\frac{\sqrt{34}}{16}, \frac{\sqrt{10}}{16}, \frac{\sqrt{18}}{16}, \frac{\sqrt{26}}{16}\right]$ and $[0.33, 0.16, 0.85, 0.85]$ respectively. We notice here that the first equivalence class having a uniform distribution has the highest distribution leakage and does not achieve the best entropy leakage.

Apparently, the distribution leakage maybe considered as a reflection of the extent to which privacy of attribute values are leaked. On the other hand, the entropy leakage is a reflection of the extent to which privacy of individuals within an equivalence class is leaked compared to the initial entropy from the original distribution. We stress that both leakages contribute to the overall privacy leakage of individuals in the published dataset.

TABLE 4: Impatient Micro-data

(a) Original Dataset

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Virus Infection
4	13053	23	American	Virus Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Virus Infection
8	14850	49	American	Virus Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

(b) 4-anonymous Impatient Micro-data

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	<30	*	Heart Disease
2	130**	<30	*	Heart Disease
3	130**	<30	*	Virus Infection
4	130**	<30	*	Virus Infection
5	1485*	≥40	*	Cancer
6	1485*	≥40	*	Heart Disease
7	1485*	≥40	*	Virus Infection
8	1485*	≥40	*	Virus Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(c) 3-diverse Impatient Micro-data

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤40	*	Heart Disease
2	1305*	≤40	*	Virus Infection
3	1305*	≤40	*	Cancer
4	1305*	≤40	*	Cancer
5	1485*	>40	*	Cancer
6	1485*	>40	*	Heart Disease
7	1485*	>40	*	Virus Infection
8	1485*	>40	*	Virus Infection
9	1306*	≤40	*	Heart Disease
10	1306*	≤40	*	Virus Infection
11	1306*	≤40	*	Cancer
12	1306*	≤40	*	Cancer

Based on the previous examples, it is obvious that while designing any publishing technique, to achieve a better level of privacy, the data publisher should not only consider the distribution leakage but also the entropy leakage. Now we use our proposed metrics to analyze some existing schemes.

Example 8. In the example from [4], the original impatient dataset is given in Table 4 and the 4-anonymous impatient dataset is given in Table 4b. For these two tables, the probability distribution for the three diseases is $\left(\frac{3}{12}, \frac{4}{12}, \frac{5}{12}\right)$. In this case the distribution leakage $\mathscr{L}_D(S, [C])$ for individuals within the first, second and third equivalence classes is $[0.5137, 0.2357, 0.7619]$, while the entropy leakage is $[0.5546, 0.0546, 1.5546]$ respectively. Our findings for Table 4b can be summarized as follows:

1) Patients under 30 have $\mathsf{Heart-Disease}$ or $\mathsf{Virus-Infection}$ with equal probability. The scheme

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2797092, IEEE Transactions on Knowledge and Data Engineering

provides distribution leakage $\mathscr{L}_D(S,[C_1]) = 0.5137$ and entropy leakage equals $\mathscr{L}_E(< 30) = 0.5546$.

2) For patients over 40, $\frac{1}{4}$ have Cancer, $\frac{1}{4}$ have Heart − Disease and $\frac{1}{2}$ have Virus − Infection. The scheme provides distribution leakage $\mathscr{L}_D(S,[C_2]) = 0.2357$ and entropy leakage is $\mathscr{L}_E(\geq 40) = 0.0546$.
3) Patients in their 30s, all have Cancer. The individual gets distribution leakage $\mathscr{L}_D(S,[C_3]) = 0.7619$ and entropy leakage $\mathscr{L}_E(30s) = 1.5546$.

Example 9. For the same original impatient dataset from last example given in Table 4, the 3-diverse impatient dataset is given in Table 4c. For these two tables, the probability distribution for the three diseases is $(\frac{3}{12}, \frac{4}{12}, \frac{5}{12})$. In this case the distribution leakage for individuals within the first, second and third equivalence classes is $[0.1179, 0.2357, 0.1179]$, while the entropy leakage is $[0.0546, 0.0546, 0.0546]$ respectively. Our findings for Table 4c can be summarized as follows:

1) Patients under 40 and living in Zip − Code = 1305$*$ have Heart − Disease, Virus − Infection or Cancer. Therefore, the scheme provides distribution leakage $\mathscr{L}_D(S,[C_1]) = 0.1179$ and entropy leakage $\mathscr{L}_E(1305 * \cap \leq 40) = 0.0546$.
2) For patients over 40 and living in Zip − Code = 1485$*$, having Heart − Disease, Virus − Infection or Cancer. The scheme provides distribution leakage $\mathscr{L}_D(S,[C_2]) = 0.2357$ and entropy leakage $\mathscr{L}_E(1485 * \cap > 40) = 0.0546$.
3) For patients under 40 and living in Zip − Code = 1306$*$, having Heart − Disease, Virus − Infection or Cancer. The individual gets distribution leakage $\mathscr{L}_D(S,[C_3]) = 0.1179$ and entropy leakage $\mathscr{L}_E(1306 * \cap \leq 40) = 0.0546$.

Example 10. In this example, the original impatient dataset is given in Table 5a. The 4-anonymous, 2-diverse, 0.67-closeness impatient dataset is given in Table 5b. For these two tables, the original probability distribution for the three diseases is $(\frac{7}{12}, \frac{3}{12}, \frac{2}{12})$. In this case the EMD is $[\frac{1}{3}, \frac{1}{3}, \frac{2}{3}]$, 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.

We finally show how EMD in t-closeness is not reliable and insufficient to measure privacy leakage.

Example 11. For the original impatient dataset from [5] given in Table 3, an 0.167-closeness, w.r.t salary, impatient dataset is given in Table 2b. The original distribution for the salaries is $\{3K, 4K, 5K, 6K, 7K, 8K, 9K, 10K, 11K\}$. The distribution of attribute values 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 $[C_3]$. In this case the EMD for the three classes is given as $[0.167, 0.167, 0.083]$. The EMD, proposed in t-closeness, is a semantic metric. It gives a weight to the attribute values based on their sensitivity in the original distribution. However, as we will show, it fails to give a correct measurement of the privacy leakage. For example, consider a 27 records dataset with same attribute values having the same uniform distribution. After publishing, this dataset is divided into 9 equivalence classes. We consider two possible equivalence classes. Assuming that the sensitive attribute values in the first, and second classes are $\{3K, 4K, 5K\}$ and $\{7K, 7K, 7K\}$. The EMD for both classes is calculated as $[0.375, 0.278]$. Based on the EMD, $[C_2]$ achieves better privacy level among the two equivalence classes. However, it is obvious that the adversarial general belief about the attribute values before and after publishing has changed more dramatically in $[C_2]$ compared to $[C_1]$. This change of belief is properly characterized in the value of our 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 where all individuals have the same sensitive attribute value $(7K)$. Thus, an adversary would know the attribute value of an individual in this class with probability 1. This change of adversarial certainty about individual's attribute values is indeed a privacy leakage. 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 5: Impatient Micro-data

(a) Original Dataset

	Non-Sensitive		Sensitive
	Zip Code	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
	Zip Code	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

6.2 Simulation Results

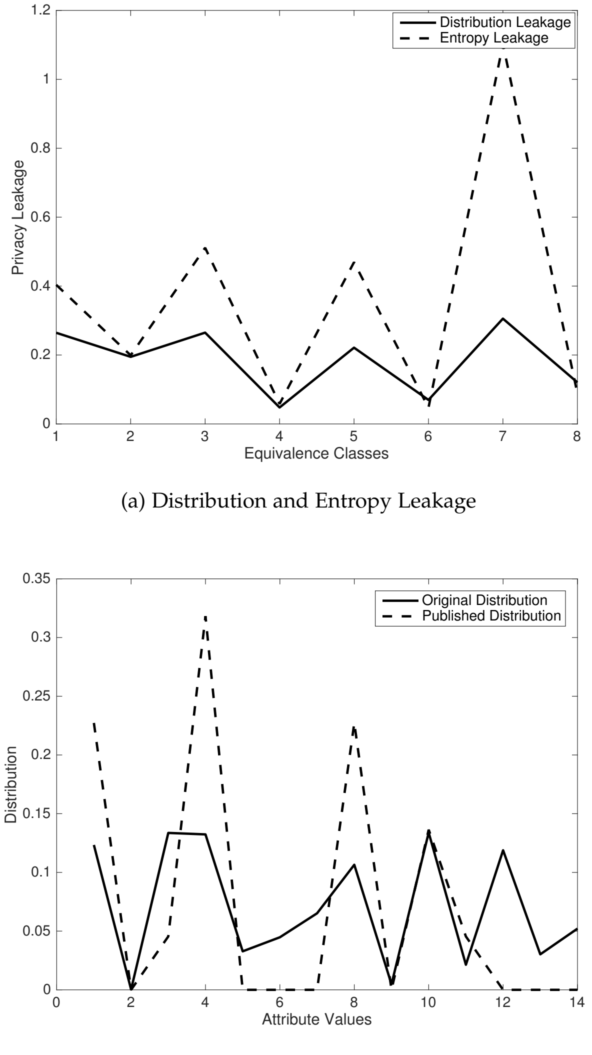
In our simulations, we investigate the effectiveness of different PPDP techniques based on our privacy metrics. Simulation results give us a more insightful understanding of privacy leakage. Specifically, our analysis gives a spotlight on several instances where published tables are believed to achieve privacy based on the PPDP techniques utilized, while based on our metrics, they do leak private valuable information about users in the datasets. We also show how our proposed metrics enable a data publisher to have more control over the privacy of a specific group of users having certain sensitive attribute values.

TABLE 6: 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	

Simulations are done on a sample of the US census dataset from the UC Irvine machine learning repository [31]. After eliminating records with missing values, we have a total of 30,162 records. Following the work in [4], as shown in Table 6, 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 [14] 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 6. While evaluating privacy of different PPDP techniques, it is essential to maintain the same level of data quality, i.e. unifying the level by which data is generalized to achieve the privacy constraint of the compared techniques.

We start by considering 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 the results shown in 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 this class $\mathsf{Age} = [75, 100], \mathsf{WorkClass} = \mathsf{Gov}$. To further understand the reason behind this leakage, we refer back to the distribution of the sensitive attribute at this specific class before and after publishing. 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 interest. Specifically, an observer, knowing that a certain user of interest falls in the age range $\mathsf{Age} = [75, 100]$ and work class category $\mathsf{WorkClass} = \mathsf{Gov}$, can eliminate 8 possible attribute values from the sensitive attribute domain.



(a) Distribution and Entropy Leakage

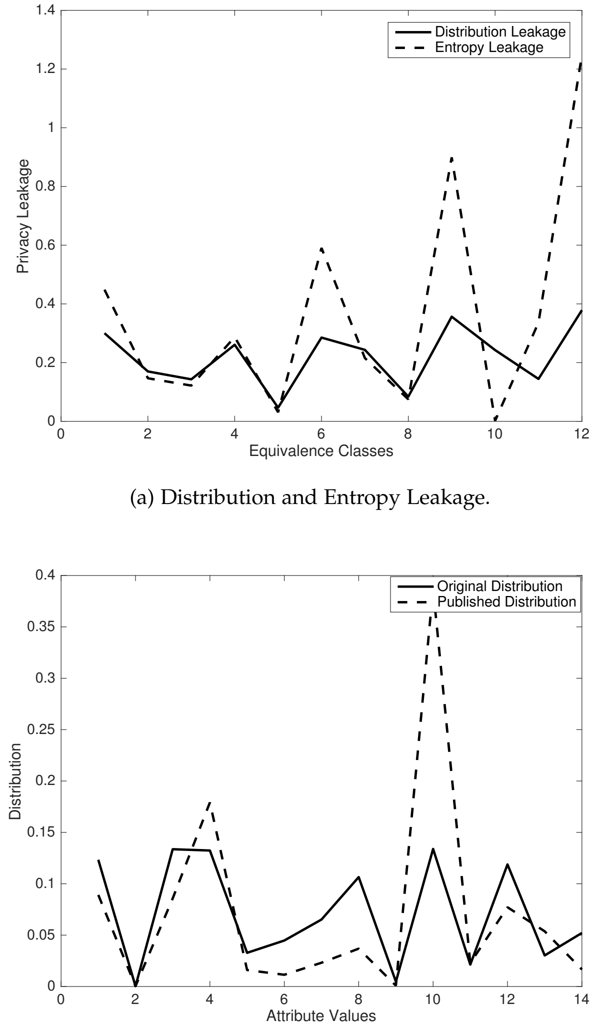
(b) Original Vs. Published Distribution of a Specific Class

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

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, observing the published table, we find that there is a noticeable privacy leakage in $[C_2]$. More specifically, an observer, with just knowing that the user of interest is more than 50 years old, will have 100% confidence that this user's Occupation is not $\mathsf{Armed-Forces}$. This privacy leakage could be noticed using our privacy metrics. The distribution and entropy leakage values of $[C_2]$ are relatively high, 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 $[0.1077, 0.1339]$. This is expressed in our distribution leakage metric shown in Fig. 3a where its value is relatively high for this specific class at 0.2853.

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. For example, let us consider four versions of a published table T' at $n = 2$. In Fig. 4, we compare distribution and entropy leakages of the four tables

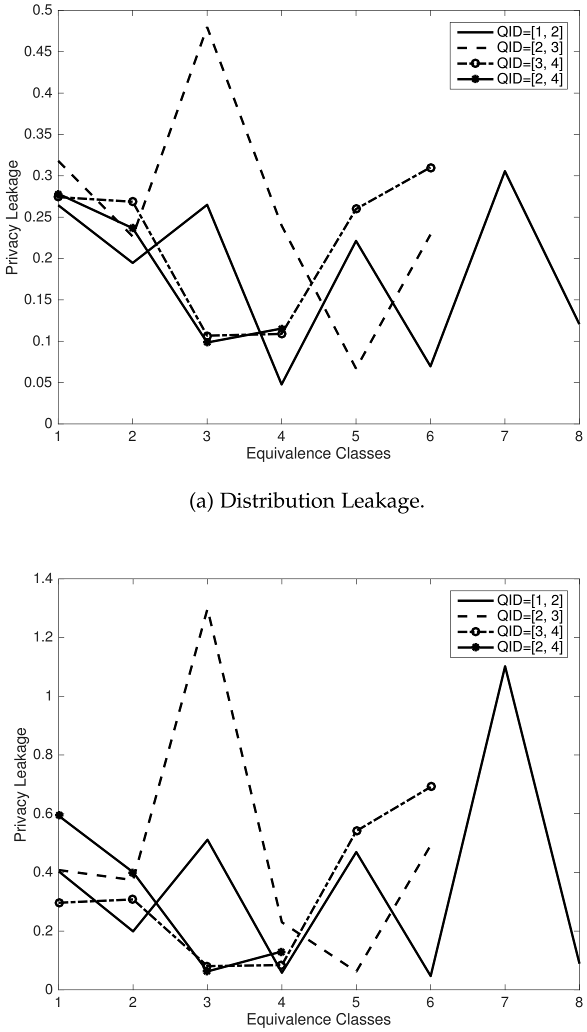


(a) Distribution and Entropy Leakage.

(b) Original and Published Distribution of a Specific Class.

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

while choosing a different combination of quasi-identifiers for each table, where quasi-identifiers are chosen to be $QID = [(1,2),(2,3),(2,4),(3,4)]$. 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. In particular, Fig. 4 shows that anonymization process ended up with 8 classes at $QID = (1,2)$, 6 classes at $QID = [(2,3),(3,4)]$, and 4 classes at $QID = (2,4)$. 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. Depending on the sensitivity of different classes formed by different combinations of quasi-identifiers, this tool gives an interesting option to adjust parameters by which a data publisher achieves the desirable privacy level with the requested data utility. Specifically, if a data publisher is more concerned about privacy of certain users that fall into $[C_3]$, then, as shown in Fig. 4, choosing $QID = [2,3]$ would leak too much private information about these users. Hence, according to our proposed metrics, a data-publisher can not only design the suitable data publishing technique for all users in a dataset, but also for a certain set of them.



(a) Distribution Leakage.

(b) Entropy Leakage.

Fig. 4: Leakage at different sets of QIDs.

7 Conclusion

In this paper, we introduced comprehensive characterization and novel quantification methods of privacy to deal with the problem of privacy quantification in privacy-preserving data publishing. In order to consider the privacy loss of combined attributes, we presented data publishing as a multi-relational model. We re-defined the prior and posterior beliefs of the adversary. 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. Our experiments demonstrate how we could gain a better judgment of existing techniques and help analyze their effectiveness in reaching privacy.

Our work opens doors to a wide range of research problems and questions including whether two metrics are sufficient to evaluate privacy or there exist other independent metrics that could help achieve better privacy quantification. Another open problem is the optimization of the original data generalization as to achieve maximum privacy

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2018.2797092, IEEE Transactions on Knowledge and Data Engineering

based on our proposed metrics. Typically, we believe that equivalence classes should be designed in such a way that keeps both the entropy leakage and the distribution leakage below a certain pre-determined level. This motivates us to think of a typical publishing scenario. We also leave as an open problem for further research, optimization of the chosen set of quasi-identifiers with an objective of minimizing distribution and entropy leakages within the published table or specific classes of higher privacy concerns.

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] L. Sweeney, "Uniqueness of simple demographics in the U.S. population," 2000.

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

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

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

[6] N. Li, W. Qardaji, D. S. Purdue, Y. Wu, and W. Yang, "Membership privacy: A unifying framework for privacy definitions," in *CCS*, (Berlin, Germany), 2013.

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

[8] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, (New York, NY, USA), pp. 1423–1434, ACM, 2014.

[9] M. Götz, S. Nath, and J. Gehrke, "Maskit: Privately releasing user context streams for personalized mobile applications," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, (New York, NY, USA), pp. 289–300, ACM, 2012.

[10] Y. Rubner, C. Tomasi, , L. J., and Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[11] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Trans. on Knowl. and Data Eng.*, vol. 22, pp. 1623–1636, Nov. 2010.

[12] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *Trans. Info. For. Sec.*, vol. 8, pp. 838–852, June 2013.

[13] C. Dwork., "Differential privacy," *ICALP*, 2006.

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

[15] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Transaction Knowledge Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.

[16] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," *In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE)*, pp. 217–228, 2005.

[17] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," *In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE)*, pp. 205–216, 2005.

[18] B. C. M. Fung, K. Wang, and P. S. Yu, "Anonymizing classification data for privacy preservation," *IEEE Trans. Knowl. Data Engin.*, vol. 19, no. 5, pp. 711–725, 2007.

[19] V. S. Iyengar, "Transforming data to satisfy privacy constraints," *In Proceedings of the 8th ACM SIGKDD*, pp. 279–288, 2002.

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

[21] N. Adam and J. Worthmann, "Security-control methods for statistical databases: A comparative study.," *ACM Computing Surveys*, 1989.

[22] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *SIGMOD*, 2000.

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

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

[25] I. Dinur and K. Nissim, "Revealing information while preserving privacy," *PODS*, 2003.

[26] J. Traub, Y. Yemini, and H. Wozniakowski, "The statistical security of a statistical database," *ACM Transactions on Database Systems*, 1984.

[27] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ICDE*, 2006.

[28] T. Truta and B. Vinay, "Privacy protection: p-sensitive k-anonymity property," *Proc. Int'l Workshop Privacy Data Management (ICDE Workshops)*, 2006.

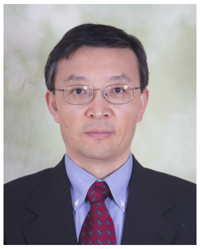
[29] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, pp. 14:1–14:53, June 2010.

[30] 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.

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

M. H. Afifi (S'17) received his B.S. and M.Sc. degrees in electrical engineering from department of Electronics and Communications, Arab Academy for Science, Technology (AAST), Alexandria, Egypt, in 2009 and 2012, respectively. He is currently a Research Assistant and a Ph. D. student at the department of Electrical and Computer Engineering, Michigan State University (MSU), East Lansing, Michigan, USA. His research interests include cybersecurity, data privacy, wireless communications, signal processing and wireless sensor networks. He is a student member of the IEEE.

Kai Zhou (S'16) received the B.S. degree in electrical engineering from Shanghai Jiao Tong University, China, in 2013. He is currently pursuing the Ph.D. degree in Electrical and Computer Engineering from Michigan State University. His research interests include applied cryptography, cloud security and privacy, coding theory, and secure communication. He is a student member of the IEEE.



Jian Ren (SM'09) received the BS and MS degrees both in mathematics from Shaanxi Normal University, and received the Ph.D. degree in EE from Xidian University, China. He is an Associate Professor in the Department of ECE at Michigan State University. His current research interests include network security, cloud computing security, privacy-preserving communications, distributed network storage, and Internet of Things. He is a recipient of the US National Science Foundation Faculty Early Career Development (CAREER) award in 2009. Dr. Ren is the TPC Chair of IEEE ICNC'17 and General Chair of ICNC'18. Dr. Ren is a senior member of the IEEE.