TRANSACTIONS ON DATA PRIVACY 3 (2010) 199–215

An information theoretic approach for privacy metrics

Michele Bezzi

SAP Labs.

F-06560, Mougins, France

E-mail: michele.bezzi@sap.com

Abstract. Organizations often need to release microdata without revealing sensitive information. To this scope, data are anonymized and, to assess the quality of the process, various privacy metrics have been proposed, such as k-anonymity, ℓ-diversity, and t-closeness. These metrics are able to capture different aspects of the disclosure risk, imposing minimal requirements on the association of an individual with the sensitive attributes. If we want to combine them in a optimization problem, we need a common framework able to express all these privacy conditions. Previous studies proposed the notion of mutual information to measure the different kinds of disclosure risks and the utility, but, since mutual information is an average quantity, it is not able to completely express these conditions on single records. We introduce here the notion of one-symbol information (i.e., the contribution to mutual information by a single record) that allows to express and compare the disclosure risk metrics. In addition, we obtain a relation between the risk values t and ℓ, which can be used for parameter setting. We also show, by numerical experiments, how ℓ-diversity and t-closeness can be represented in terms of two different, but equally acceptable, conditions on the information gain.

Keywords. Database Applications-Statistical databases; Systems and Information Theory; Anonymization

1 Introduction

Governmental agencies and corporates hold a huge amount of data containing information on individual people or companies (micro data). They have often to release part of these data for research purposes, data analysis or application testing. However, these data contain sensitive information and organizations are hesitant to publish them. To reduce the risk, data publishers use masking techniques (*anonymization*) for limiting disclosure risk in releasing sensitive datasets. In many cases, the data publisher does not know in advance the data mining tasks performed by the recipient, or even it does not know who the recipients are (so called *privacy preserving data publishing* [18]). Therefore, the data publisher tasks are basically to anonymize (under some privacy constraints) and to release the data.

[0]Some parts of this paper has been presented at the International Conference on Privacy in Statistical Databases, Lausanne, Switzerland [6]. The theoretical framework has been extended, in particular new constraints on the mutual information has been added, Eq. 4, an upper bound for ℓ derived, and relation between t and ℓ proposed, Eq. 12. The experimental work (Section 4) and the corresponding discussion are all new work. Introduction and Conclusions have also been revised.

In other scenarios, the data publishers may know the data mining activities in advance (privacy preserving data mining scenario), therefore they can customize the anonymization to preserve some statistical properties. In this paper we will mainly focus on the first scenario, privacy preserving data publishing, which is relevant in many real use cases, such as creating test data for application testing, outsourcing the development of new mining algorithms [2], and it is also pushed by regulatory frameworks [9].

Typically, data are contained in tables, and the attributes (columns) in the original table can be categorized, from disclosure perspective, in the following types:

- *Identifiers*. Attributes that explicitly identify individuals. E.g., Social Security Number, passport number, complete name.
- *Quasi-identifiers (QIs) or key attributes* [10]. Attributes that in combination can be used to identify an individual E.g., Postal code, age, gender, etc ... .
- *Sensitive attributes*. Attributes that contain sensitive information about an individual or business, e.g., salary, diseases, political views, etc ...

Various anonymization methods can be applied to obfuscate the sensitive information, they include: generalizing the data, i.e., recoding variables into broader classes (e.g., releasing only the first two digits of the zip code) or rounding numerical data, suppressing part of or entire records, randomly swapping some attributes in the original data records, permutations or perturbative masking, i.e., adding random noise to numerical data values. These anonymization methods increase protection, lowering the disclosure risk, but, clearly, they also decrease the quality of the data and hence its utility [14]. There are two types of disclosure: identity disclosure, and attribute disclosure. Identity disclosure occurs when the identity of an individual is associated with a record containing confidential information in the released dataset. Attribute disclosure occurs when an attribute value may be associated with an individual (without necessarily being able to link to a specific record). Anonymizing the original data, we want to prevent both kinds of disclosures. In the anonymization process Identifiers are suppressed (or replaced with random values), but this is not sufficient, since combining the quasi-identifiers values with some external source information (e.g., a public register) an attacker could still be able to re-identify part of the records in the dataset. To reduce the risk some of the masking techniques described above are applied on the quasi-identifiers. To assess the quality of the anonymization process, there is the need to measure the disclosure risk in the anonymized dataset and its utility. Typically, disclosure risk metrics are set to define the privacy constraints, and then we look for the masking mechanisms that maximize the utility. Both disclosure risk and utility metrics are hard to define in general, because they may depend on context variables, e.g., data usage, level of knowledge of the attacker, amount of data released, etc..., and many possible definitions have been proposed so far.

We focus here on disclosure risk measures. Various models have been proposed to capture different aspects of the privacy risk (see [18] for a review). From the data publisher point of view, it would be desirable to have these privacy models expressed in terms of semantically "similar" measures, so he could be able to compare their impact and optimize the trade off between the different privacy risks.

In Ref. [1], the authors proposed an information theoretic framework to express average disclosure risk using mutual information. The advantages of mutual information formulation are twofold: first, it allows to express the different risk measures, and associate thresholds, in a common framework, with well defined units; second, it permits applying a

wide range of well established information theory tools to risk optimization (e.g., privacy-distortion trade off problem [24]). Note that, mutual information (or, similarly, information loss) has been also proposed as utility measure [11]. Although, we will use this definition in our simulations (see Sect. 4), the main focus of our paper is investigating how we can compare the different privacy metrics among them, independently on the utility metrics used.

In this paper, we extend the information theoretic formulation of disclosure risk measures. In particular, existing privacy metrics (k-anonymity, ℓ-diversity and t-closeness metrics [25, 22, 20]) define minimal requirements for each entry (or QI group) in the dataset, but because mutual information is an average quantity, it is not able to completely express these conditions on single entries. In fact, as pointed out in [21], *privacy is an individual concept and should be measured separately for each individual*, accordingly average measures, as mutual information, are not able to fully capture privacy risk.

Thus, we introduce here two types of one-symbol information (i.e., the contribution to mutual information by a single record), and define the disclosure risk metrics in terms of information theory (see Sect. 3). By introducing one-symbol information we are able to express and compare different risk concepts, such as k-anonymity, ℓ-diversity and t-closeness, using the same units. In addition, we obtain a set of constraints on the mutual and one-symbol information for satisfying ℓ-diversity and t-closeness. We also derive a relation between the risk parameters t and ℓ, which allows to assess t in terms of the more intuitive ℓ value.

We present a simple example, to point out that in presence of a constant average risk, the records at risk may depend on the information metric used (Sect. 3.1). In Sect. 4, we test our framework, using a census dataset, to show the relevant differences between risk estimation based on average measures, as mutual information, and record specific metrics, as one-symbol information. We also show how focusing on the information contribution of a single record or group, we can minimize the information loss during anonymization. Lastly, we discuss our results and introduce some directions for future work.

In summary, this paper does not provide a unique answer to what disclosure risk is, but it gives the necessary theoretical ground for expressing and comparing different risk measures, and provides some useful relations for setting risk parameters.

Before entering in details about the proposed model, in the following sections we will introduce some background on disclosure risk metrics (Sect. 2.1) and information theory (Sect. 2.2).

2 Preliminaries

2.1 Privacy Metrics

Let us consider a dataset containing identifiers, quasi-identifiers (QIs), X, and sensitive attributes, W (for example as in Table 2). We create an anonymized version of such data, removing identifiers, and anonymizing quasi-identifiers ($\tilde{X}$), for example generalizing them in classes (see Table 3).

To estimate the disclosure risk in the anonymized data, various metrics have been proposed so far.

k-Anonymity [25] condition requires that *every* combination of key attributes (QI group) is shared by at least k records in the anonymized dataset. A large k value indicates that the anonymized dataset has a low identity disclosure risk, because, at best, an attacker has a

probability $1/k$ to re-identify a record, but it does not necessarily protect against attribute disclosure. In fact, a QI group (with minimal size of k records) could also have the same confidential attribute, so even if the attacker is not able to re-identify the record, he can discover the sensitive information.

To capture this kind of risk ℓ-diversity was introduced [22]. ℓ-diversity condition requires that for *every* combination of key attributes there should be at least ℓ "well represented" values for each confidential attribute. In the original paper, a number of definitions of "well represented" were proposed. Because we are interested here in providing an information-theoretic framework, the more relevant for us is in terms of entropy, i.e.,

$$H(W|\tilde{x}) \equiv -\sum_{w \in W} p(w|\tilde{x}) \log_2 p(w|\tilde{x}) \geq \log_2 \ell$$

for every QI group $\tilde{x}$, and with $\ell \geq 1$. For example, if each QI group have n equally distributed values for the sensitive attributes, the entire dataset will be n-diverse. Note that if ℓ-diversity condition holds, also the k-anonymity condition (with $k \leq \ell$) automatically holds, since there should be at least ℓ records for each group of QIs.

Although, ℓ-diversity condition prevents the possible attacker to infer exactly the sensitive attributes, he may still learn a considerable amount of probabilistic information. In particular if the distribution of confidential attributes within a QI group is very dissimilar from the distribution over the whole set, an attacker may increase his knowledge on sensitive attributes (*skewness attack*, see [20] for details). t-closeness estimates this risk by computing the distance between the distribution of confidential attributes within the QI group and in the entire dataset. The authors in [20] proposed two ways to measure the distance, one of them has a straightforward relationship with mutual information (see Eq. 7 below), as we discuss in the next section.

Another notion for assessing the privacy risk is differential privacy [15]. The basic idea is that the removal, addition or replacement of a single personal information (a record) in the original database should not significantly impact the outcome of any statistical analysis. This privacy model requires that the data publisher knows in advance the exact set of queries/analysis that need to be performed on the released data. However, this does not fit the requirements of the *privacy preserving data publishing* scenario we are addressing in this paper. We will discuss possible extension to include differential privacy in our framework in the last section.

These measures provide a quantitative assessment of the different risks associated to data release, but they have also major limitations. First, they impose strong constraints on the anonymization, resulting in a large utility loss; second, it is often hard to find a computational procedure to achieve a pre-defined level of risk; third, since they capture different features of disclosure risks, they are difficult to compare and optimize at the same time. To address the last point, we propose in the next sections a common framework, one-symbol information, for expressing these three risk measures. We will discuss the first two issues, in the last Section.

2.2 Information theory

Let us consider two random variables X and Y (e.g., the tuple of quasi-identifiers or sensitive attributes), which take values x and y. Let us denote the corresponding probability density or probability mass functions $p_X(x)$ ($p(x)$ in short) and $p_Y(x)$ ($p(y)$). In the context of data anonymization, $p(x)$ and $p(y)$ may be estimated in terms of frequency. Be $p(x, y)$

Definition	Positive definite	Chain rule X	Chain rule Y	Average MI
I_1	Yes	Yes	No	Yes
I_2	No	Yes	Yes	Yes
I_3	No	Yes	No	Yes
I_4	Yes	Yes	Yes/No	No

Table 1: Main properties of the four definitions of one-symbol information. I_1 and I_2 are discussed in the main text, $I_3(x,Y) \equiv \sum_{y\in Y} p(y|x)[H(X)-H(X|y)]$ [8] is a definition based on weighted average of reduction of uncertainty, $I_4(x,Y) \equiv I(\{x,\bar{x}\};Y)$ [3] is the mutual information between Y and a set composed by two elements: x and its complement in X: $\bar{x} \equiv X\backslash x$. For more details, see [5].

and $p(x|y)$ the corresponding joint and conditional probability functions. Following Shannon [26], we can define the mutual information $I(X;Y)$ as:

$$\begin{aligned} I(X;Y) &= \sum_{x\in X, y\in Y} p(x,y)\ \log_2\left[\frac{p(x,y)}{p(x)p(y)}\right] \\ &= \sum_{x\in X, y\in Y} p(y)p(x|y)\ \log_2\left[\frac{p(x|y)}{p(x)}\right], \end{aligned} \tag{1}$$

(with conditional probability $p(x|y) = p(x,y)/p(y)$ according to Bayes' rule) or, equivalently, introducing the entropy of a probability distribution: $H(Y) = -\sum_{y\in Y} p(y)\log_2 p(y)$,

$$I(X;Y) = H(Y) - H(Y|X) = \sum_{x\in X} p(x)[H(Y) - H(Y|x)] \tag{2}$$

where $H(Y|X) \equiv \sum_{x\in X} p(x)H(Y|x)$ is the conditional entropy.

Mutual information summarizes the *average* amount of knowledge we gain about X by observing Y (or vice-versa); e.g.: in the trivial case, they are completely independent, $p(x,y) = p(x)p(y)$ and $I = 0$. Mutual information has some mathematical properties that agree to our intuitive notion of information. In particular, we expect that any observation does not decrease the knowledge we have about the system. So, mutual information has to be positive, as it can be easily shown starting from Shannon's definition. In addition, for two independent random variables $\{X^1, X^2\}$, we expect: $I(\{X^1,X^2\},Y\}) = I(X^1,Y) + I(X^2,Y)$. This additivity property is a special case of a more general property, known as chain rule [26].

Mutual information is an average quantity, for some applications (see Sect. 3 and e.g., [5] and references therein), it is important to know which is the contribution of a single symbol (i.e., a single value x or y) to the information. In his original formulation, Shannon did not provide any insights about how much information can be carried by a single symbol, such as a single tuple in our case. After Shannon's seminal work, to the author's knowledge, four different definitions of so called one-symbol specific information (sometimes also called *stimulus specific information*, because it has been used in the framework of neural response analysis) have been proposed. Ideally, this *specific information* should be proper information in a mathematical sense (non-negative, additive) and give mutual information as average.

In a previous work [5], we reviewed the four definitions proposed in literature (see Table 2.2), outlining their main characteristics, and, in particular, stressing that none of the

proposed definitions have these desired mathematical properties, but each of them can capture different aspects of information transmission. We also showed, in the context of brain processing, how each of them should be used according to the aspect of neural coding we are interested in studying.

In this paper, we want to express existing privacy metrics in terms of such specific informations, accordingly, we will focus on two of them, following [12] referred as I_1 and I_2, that can be directly linked to well-known risk metrics (see below). The other two measures, I_3 and I_4 are not considered here, because they have no easy interpretation as privacy metrics, and, as mentioned above, the scope of this study is not introducing new privacy metrics, but to redefine the existing ones in a common framework.

I_1, Surprise (j-measure)

Originally proposed by Fano [17], this definition can be immediately inferred from Eq. (1), simply taking the single symbol contribution to the sum:

$$I_1(x, Y) = Surprise(x) = \sum_{y \in \mathcal{Y}} p(y|x) \log_2 \frac{p(y|x)}{p(y)}$$

This quantity measures the deviation (Kullback-Leibler distance) between the marginal distribution $p(y)$ and conditional probability distribution $p(y|x)$. It clearly averages to the mutual information, i.e. $\sum_{x \in X} p(x) I_1(x; Y) = I(X; Y)$, and it is always non-negative: $I_1(x; Y) \geq 0$ for $x \in X$. Furthermore it is the only positive decomposition of the mutual information [7]. Since $I_1(x, Y)$ is large when $p(y|x)$ dominates in the regions where $p(y)$ is small, i.e., in presence of *surprising* events, this quantity is often referred to as "surprise" [12] (it is also called j-index [27]). Surprise lacks additivity, and this causes many difficulties when we want to apply it to a sequence of observations. Despite this main drawback, *surprise* has been widely used, for example for exploring the encoding of brain signals or for association rule discovery [27].

I_2, Specific Information (i-measure)

An entropy based definition was proposed by Blachman [7] and it may be derived from Eq. (2), extracting the single symbol contribution from the sum:

$$\begin{aligned} I_2(Y, x) &= H(Y) - H(Y|x) = \\ &= -\left[\sum_{y \in Y} p(y) \log_2 p(y) - p(y|x) \log_2 p(y|x) \right] \end{aligned} \tag{3}$$

Here, information is identified with the reduction of entropy between marginal distribution $p(y)$ and conditional probability $p(y|x)$. The $I_2(Y, x)$ measure is called *specific information* in [12], and *i-measure* in [27]. This quantity captures how diverse are the entries in Y for a given entry (tuple) x. Indeed, it expresses the difference of uncertainty between the a priori knowledge of Y, $H(Y)$, and the knowledge for a given symbol x, $H(Y|x)$. As shown in [12], this is the only decomposition of mutual information that is also additive, but, unlike mutual information, it can assume negative values.

Note that any weighted combination of I_1 and I_2 averages to mutual information, and it can represent a possible definition of one-symbol specific information. Thus, we have an

Original Dataset $\{X, W\}$		
Name	Height X	Diagnosis W
Timothy	166	N
Alice	163	N
Perry	161	N
Tom	167	N
Ron	175	N
Omer	170	N
Bob	170	N
Amber	171	N
Sonya	181	N
Leslie	183	N
Erin	195	Y
John	191	N

Table 2: Original dataset.

infinite number of plausible choices for a one-symbol decomposition of mutual information. But, as mentioned above, only I_1 is always non-negative and for I_2 only the chain rule is fulfilled. In addition, as we will see in the next section, only I_1 and I_2 have a straightforward interpretation as disclosure risk measures.

Anonymized Dataset $\{\tilde{X}, W\}$		
Name	Height $\tilde{X}$	Diagnosis W
******	[160-170]	N
*****		N
*****		N
*****		N
******	[170-180]	N
*****		N
*****		N
*****		N
******	[180-190]	N
*****		N
******	[190-200]	Y
*****		N

Table 3: Anonymized dataset.

3 Information theoretic Risk Metrics

Before entering in the details of the different privacy metrics, let us deduce a general relation between the information about the sensitive attributes and the quasi-identifiers before, $I(W, X)$, an after the anonymization process $I(W, \tilde{X})$. Because the anonymization of the

QIs can be generally represented as a function between X and $\tilde{X}$ [1] for the data processing inequality of the mutual information, we have:

$$I(W, \tilde{X}) \leq I(W, X) \tag{4}$$

Let us consider the different privacy metrics in terms of information theory.

- k-anonymity. In case of suppression and generalization, we have that a single QI group in the anonymized database $\tilde{x}$ can correspond to a number, $N_{\tilde{x}}$ of records in the original table X. Accordingly, the probability of re-identifying a record x given $\tilde{x}$ is simply: $p(x|\tilde{x}) = 1/N_{\tilde{x}}$, and k-anonymity reads:

  $$H(X|\tilde{x}) \geq \log_2 k \tag{5}$$

  for each $\tilde{x} \in \tilde{X}$. In terms of one-symbol specific information I_2, it reads

  $$I_2(X, \tilde{x}) \equiv H(X) - H(X|\tilde{x}) \leq \log_2 \frac{N}{k} \tag{6}$$

  where N is the number of tuples in the original dataset X (assumed different). $I_2(X, \tilde{x})$ measures the identity disclosure risk for a single record. Eq. 5 holds also in case of perturbative masking [4], therefore I_2 can be used for any kind of masking transformations.

  Averaging Eq. 6 over $\tilde{X}$ we get:

  $$I(X, \tilde{X}) \leq \log_2 \frac{N}{k}$$

  So, the mutual information can be used as a risk indicator for identity disclosure [13], but we have to stress that this condition does not guarantee the k-anonymity QI group $\tilde{x}$, i.e., it is necessary but not sufficient.

- t-closeness condition requires:

  $$D(p(w|\tilde{x})||p(w)) \equiv \sum_{w \in W} p(w|\tilde{x}) \log_2 \frac{p(w|\tilde{x})}{p(w)} \leq t \tag{7}$$

  for each $\tilde{x} \in \tilde{X}$. This is equivalent to the one-symbol specific information I_1 (surprise), i.e.,

  $$I_1(W, \tilde{x}) \equiv \sum_{w \in W} p(w|\tilde{x}) \log_2 \frac{p(w|\tilde{x})}{p(w)} \leq t \tag{8}$$

  $I_1(W, \tilde{x})$ is a measure of attribute disclosure risk for a QI group $\tilde{x}$, as difference between the prior belief about W from the knowledge of the entire distribution $p(w)$, and the posterior belief $p(w|\tilde{x})$ after having observed $\tilde{x}$ and the corresponding sensitive attributes. Averaging over the set $\tilde{X}$ we get an estimation of the disclosure risk (based on t-closeness) for the whole set [24],

  $$I(W, \tilde{X}) \equiv \sum_{\tilde{x} \in \tilde{X}} p(\tilde{x}) \sum_{w \in W} p(w|\tilde{x}) \log_2 \frac{p(w|\tilde{x})}{p(w)} \leq t \tag{9}$$

[1] Strictly speaking, Eq. 4 is true if $\tilde{X}$ is a function, deterministic or random, of X only, i.e., $\tilde{X} = G(X)$. This is the case of generalization or suppression, and, in most of the cases, of perturbative masking. Eq. 4 is not valid any more when $\tilde{X}$ is not conditionally independent of W, i.e., $p(\tilde{x}|x, w) \neq p(\tilde{x}|x)$.

Again, this is necessary but not a sufficient condition to have t-closeness table, since this condition requires to have t-closeness for each $\tilde{x}$.

- ℓ-diversity condition, in terms of entropy, reads:

$$H(W|\tilde{x}) \geq \log_2 \ell$$

for each QI group $\tilde{x} \in \tilde{X}$. It can be expressed in terms of one-symbol specific information I_2,

$$I_2(W, \tilde{x}) \equiv H(W) - H(W|\tilde{x}) \leq H(W) - \log_2 \ell \tag{10}$$

$I_2(W, \tilde{x})$ is a measure of attribute disclosure risk for a QI group $\tilde{x}$, as reduction of uncertainty between the prior distribution and the conditional distribution.

Averaging over the set $\tilde{X}$ we get an estimation of the average disclosure risk for the whole set [24].

$$I(W, \tilde{X}) \equiv H(W) - H(W|\tilde{X}) \leq H(W) - \log_2 \ell \tag{11}$$

This is the ℓ-diversity condition on average. Again, this is necessary but a not sufficient condition to satisfy ℓ-diversity for each $\tilde{x}$. Note that, since the mutual information is a non-negative quantity, $I(W, \tilde{X}) \geq 0$, from Eq. 11 immediately follows that $H(W)$ is an upper bound for $\log_2 \ell$, i.e.,

$$\log_2 \ell \leq H(W)$$

or equivalently $\ell \leq \ell_{max} \equiv 2^{H(W)}$.

Eqs. 9, 11 suggest a way to compare the two risk parameters ℓ and t. Indeed, if we equalize the maximal contribution to information of ℓ and t, we can derive the following relation:

$$\ell_t = 2^{H(W)-t} \tag{12}$$

ℓ_t tells us, for a given t, what the *equivalent* value ℓ is, i.e., the value of ℓ that has the same impact on the information. The advantage of Eq. 12 is that it allows us to express the value of t parameter, which it is often hard to set, in terms of ℓ, that has a much more intuitive meaning.

In summary, for any anonymized dataset which satisfies ℓ-diversity and t-closeness, the following average conditions are *necessary*:

$$\begin{cases} I(W, \tilde{X}) \leq I(W, X) \\ I(W, \tilde{X}) \leq t \\ I(W, \tilde{X}) \leq H(W) - \log_2 \ell \end{cases} \tag{13}$$

whereas the *necessary* and *sufficient* conditions are:

$$\begin{cases} I_1(W, \tilde{x}) \leq t \\ I_2(W, \tilde{x}) \leq H(W) - \log_2 \ell \end{cases} \tag{14}$$

for each $\tilde{x} \in \tilde{X}$.

For setting the risk parameters, we derived lower and upper bounds for ℓ,

$$1 \leq \ell \leq \ell_{max} \equiv 2^{H(W)} \tag{15}$$

and the ℓ_t equivalent to t,

$$\ell_t = 2^{H(W)-t}$$

which allows to express t in terms of the more intuitive diversity parameter ℓ.

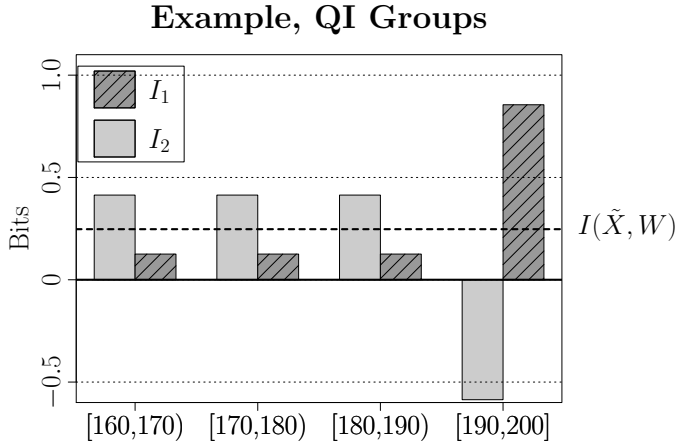

Figure 1: Values of information-based disclosure risk metrics: I_1 (t-closeness) and I_2 (ℓ-diversity related) for the different entries in the anonymized database. Dashed line indicated mutual information $I(\tilde{X}, W)$, i.e., the average of I_1 and I_2.

3.1 Example

To illustrate the qualitative and quantitative differences in the behavior of t-closeness and ℓ-diversity based risk metrics, I_1 and I_2, let us consider a simple example. This example is not realistic, but the aim is to show some basic features of I_1 and I_2 without any additional complexity. Let us take a medical database $\{X, W\}$ (Table 2) containing three fields only: a unique identifier (name), a quasi-identifier (Height) and a sensitive attribute (Diagnosis). In the released, anonymized dataset $\{\tilde{X}, W\}$, Table 3, names are removed, the Height generalized in broader classes, and the sensitive attribute unchanged. Let us say that after this anonymization process, we have reached an acceptable level of identity and attribute disclosure risk as measured by $I(X, \tilde{X})$ and $I(W, \tilde{X})$. But, if we analyze the contribution to this risk of single entries in $\tilde{X}$ in terms of symbol specific informations I_1, I_2 (see Fig. 1), we observe:

- The distribution of risk shows large fluctuations, so the average is not a good representation of the risk level.

- The entries at risk (say, well above the average) depends on the risk measures used (I_1 or I_2). In other words, set of tuples (QI groups) largely at risk according I_2 (ℓ-diversity based) have low value of I_1 (so they are acceptable from t-closeness point of view), and vice versa.

In short, this simple example shows that, although, on average the two risk metrics are equal, their impact on single entries can be the opposite. We will show a more extended analysis on a larger dataset in the next section.

UCI Adult Dataset		
Attribute	Values	Generalization
Age	74	ranges-25-50
Gender	2	Suppression
Race	5	Suppression
Education	16	Suppression
Occupation	14	Suppression
Marital Status	7	*Sensitive*
Salary Class	2	*Sensitive*

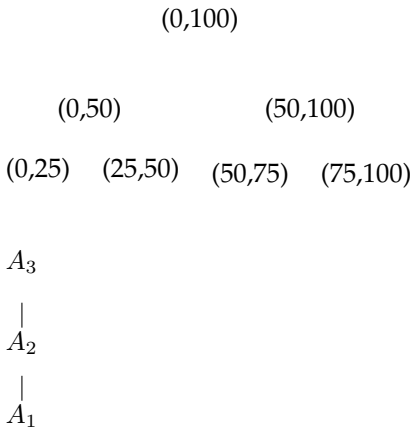

Table 4: Left. Summary of the anonymization methods. Right. Generalization hierarchy for age attribute: first level, A_1, Age is generalized in 25 years range, A_2 in 50 years and A_3 fully generalized. Un-generalized Age (A_0) not shown.

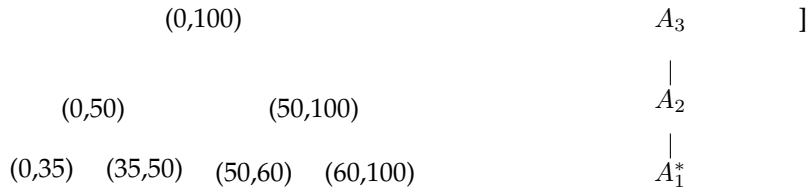

Table 5: Generalization hierarchy for age attribute: first level, A_1^*, Age is generalized in ranges of variable size, for minimizing information loss. A_2 in 50 years and A_3 fully generalized. Un-generalized Age (A_0) not shown.

4 Experiments

For testing our framework, we performed some numerical analysis on a relatively large dataset widely used in the research community. We ran our experiments on the Adult Database [2] from the UCI Machine Learning Repository, which contains 32561 tuples from US Census data with 15 demographic and employment related variables. We removed the tuples with missing values, ending with 30162 usable tuples.

The choice of the identifiers, QIs, and sensitive attribute set, typically depends on the specific domain. In particular, for QIs, they should include the attributes a possible attacker is more likely to have access to (e.g., using a phonebook or a census database) and for sensitive attribute to the application the anonymized data are used for. Generally speaking increasing the number of QIs increases the risk, or results in strong anonymization impacting final data quality. In our experiments we limited the QIs to four attributes: $QI \equiv \{Age, Gender, Race, Education\}$. The generalizations applied are summarized in Table 4. In the census data the salary is typically chosen as sensitive attribute. In this dataset,

[2] Available at http://archive.ics.uci.edu/ml/datasets/Adult

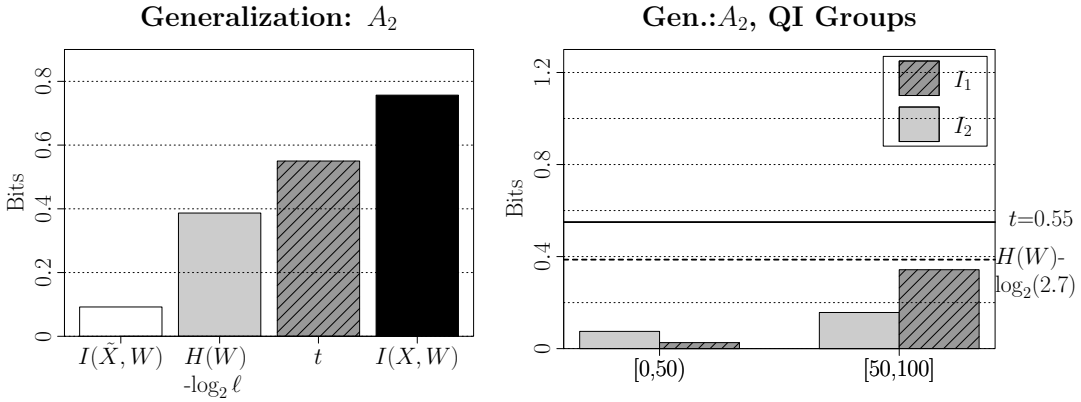

Figure 2: Left: Mutual information $I(\tilde{X}, W)$ between the anonymized QIs and the sensitive attribute (black bar), the t value (shaded dark gray bar), ℓ-diversity threshold for information (light gray bar), and the mutual information between the original QIs and the sensitive attributes (white bar). Right: information contributions in terms of I_1 (shaded dark gray bar) and I_2 (light gray bar) for the A_2 QI groups. Straight line indicates the t threshold, for I_1, with $t = 0.55$. Dashed line indicates the ℓ-diversity threshold for I_2 ($\ell = 2.7$).

salary attribute can assume two values $> 50k$ and $< 50k$, and the entropy of this attribute is quite small ($H \approx 0.80$), limiting its effectiveness for demonstrating the different features of the privacy metrics. Therefore, for our tests, we used as sensitive attribute the Marital Status, which has 7 possible values and a larger entropy ($H \approx 1.85$), which corresponds to a maximum value of $\ell_{max} \approx 3.53$.

Let us consider that we want to release this dataset, assuring that the privacy risk is under a certain threshold value, as measured by t and ℓ.

We set these thresholds as following: $\ell = 2.7$ and $t = 0.55$. For setting the ℓ threshold, we had to find a compromise between having a sufficient level of privacy, and the same time not impacting too much the quality of data. In particular ℓ values lower than 2 do not clearly provide much privacy, whereas, considering the maximum value of $\ell_{max} \approx 3.53$, values larger than 3 need a strong anonymization, which removes most of the information. We set t in a way that its information contribution does not differ too much from the ℓ contribution, for our analysis we chose $t = 0.55$, which corresponds to a value of $\ell_t = 2.42$ so differing of $\simeq 10\%$ from the chosen value of ℓ.

For the anonymization step, we used as anonymizaton engine the UT Dallas Anonymization Toolbox [19], which implements Incognito algorithm for k-anonymity and ℓ-diversity. This toolbox comprises an implementation of Incognito for t-closeness, too, but it measures t value against the Earth Moving distance (whereas we use Eq. 7), so we did not use it, and we verified the t value, according to our metric, on the anonymized dataset.

We run the anonymization engine using the transformations listed in Table 4(Left), and we obtained that the attributes $\{Gender, Race, Education\}$ were suppressed (or fully generalized) and Age generalized in two groups, A_2 (see Table 4(Right)).

In Fig. 2(Left) we plot the values of the mutual information $I(\tilde{X}, W)$ between the QI groups (anonymized QIs) and the sensitive attribute, the ℓ threshold in terms of information (Eq. 11), the t value, and the mutual information between the original QIs and the sensitive attributes. We observe that, as expected, the necessary conditions, Eqs. 13, on the

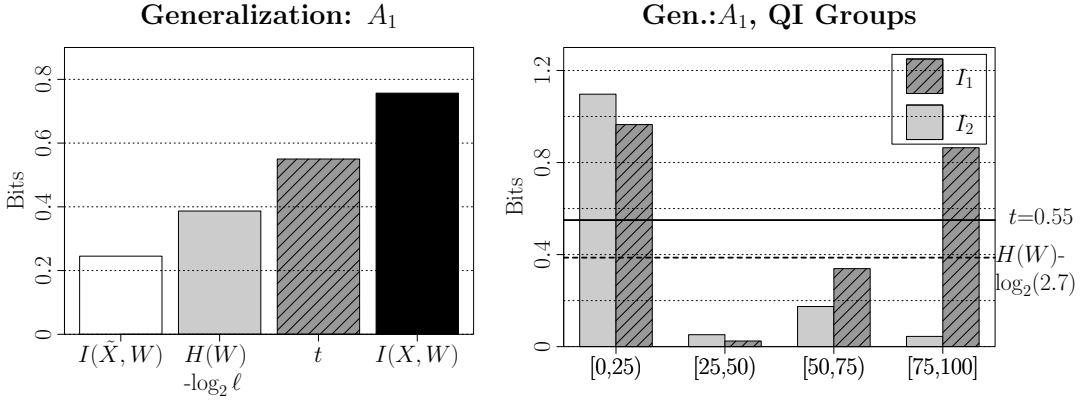

Figure 3: Left: Original mutual information $I(\tilde{X},W)$ (black bar), the t value (shaded dark gray bar), ℓ-diversity threshold (light gray bar), and the mutual information for anoynmized dataset (white bar). Right: information contributions A_1 QI groups. For details see the caption of Fig. 2

average information are fulfilled, and $I(\tilde{X},W)$ (white bar) is lower than the other three bars. In Fig. 2(Right), we plot the information contributions in terms of I_1 and I_2 for the two QI groups. Both of them are lower than the corresponding thresholds for t (continuous line) and ℓ, fulfilling the conditions expressed by Eqs. 14. In other words, for this anonymized dataset 2.7-diversity and 0.55-closeness are fulfilled.

On the other hand, the information loss is quite relevant $\approx 88\%$ (see white bar vs. black bar in Fig. 2(Left)), and it is typically desirable to preserve as much information as possible in the anonymization [3]. To this scope, we consider a weaker generalization on $\{Age\}$ attribute, i.e., one single level A_1 instead of two A_2 of the previous case.

In Fig. 3(Left) we show the amount of information $I(\tilde{X},W)$ (white bar) left after the anonymization, and (as in the previous case) we compared it with Eq. 11, the t value, and the original information $I(X,W)$. We can see that $I(\tilde{X},W)$ is lower than the other bars, i.e., the conditions expressed by Eqs. 13 are fulfilled. These conditions are necessary but not sufficient for having ℓ-diversity and t-closeness enforced. Indeed, if we analyze the contribution of the level of QI group, Fig. 2(Right), we observe that two groups (labeled $[0,25)$ and $[75,100]$) do not satisfy the t closeness condition (continuous line), and one of them, $[0,25)$, does not satisfy ℓ-diversity condition (dashed line). Note, as in the example of Section 3.1, $[75,100]$ group has a very different risk profile depending whether we use t or ℓ, i.e., very risky from t-closeness point of view, and pretty safe for ℓ-diversity.

In short, although the conditions on the average are satisfied, the same is not true for each $\tilde{x}$, QI group, so the dataset is neither 2.7-diverse nor satisfies 0.55-closeness.

We identified the two QI groups that do not fulfill the privacy requirements, the next step is trying to increase the size of these outlier groups for reducing the privacy risk (both in terms of ℓ and t). To this scope, we modified the first level of the generalization A_1

[3]Note that here we try to maximize the information $I(\tilde{X},W)$, or minimize the information loss, whereas, as mentioned in the introduction, we should maximize the utility (once the privacy constraints are satisfied). In general, utility cannot be measured as mutual information [21], but utility metric strongly depends on the final application of the anonymized data. Still, for demostrating purpose, here, we limit our analysis to information maximization, which, in any case, is always benefical to preserve at maximum.

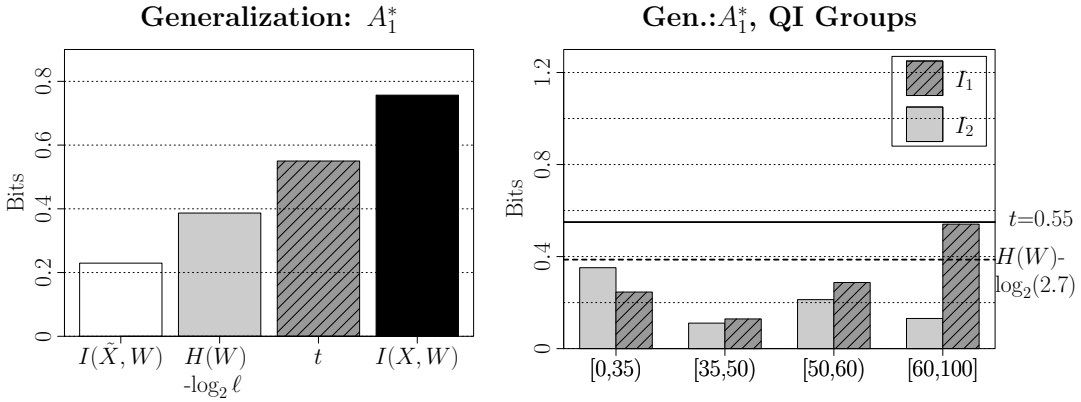

Figure 4: Left: Original mutual information $I(\tilde{X}, W)$ (black bar), the t value (shaded dark gray bar), ℓ-diversity threshold (light gray bar), and the mutual information for anoynmized dataset (white bar). Right: information contributions A_1^* QI groups. For details see the caption of Fig. 2

hierarchy for age attribute, in particular we run an heuristic search for the optimal size of the two groups, which maximizes the information $I(\tilde{X}, W)$, and satisfies the t-closeness and ℓ-diversity requirements. The resulted generalization A_1^* is reported in Table 5. In Fig. 4, we plotted, as in the previous cases, the mutual information $I(\tilde{X}, W)$ compared to ℓ and t thresholds and to the original information $I(X, W)$. On the right panel, we show I_1 and I_2 for each QI group compared to the risk thresholds. We can see as t-closeness and ℓ-diversity conditions are satisfied, and that we were able to preserve a larger amount of information compared to A_2, Fig. 2, (we do not consider A_1, since it does not satisfy our privacy requirements), indeed the information loss was reduced to $\approx 55\%$, compared to $\approx 88\%$ of A_2, and in terms of information we have $I(\tilde{X}, W) = 0.34$ for A_1^* vs. $I(\tilde{X}, W) = 0.09$ for A_2.

In summary in our numerical analysis, using a census dataset, we showed the main features of information-based privacy metrics, including their application for setting risk parameters, and how they can be used to reduce the information loss in the anonymization process.

5 Conclusions and Future Works

In the original t-closeness paper [20], the authors stated that *"Intuitively, privacy is measured by the information gain of an observer"*. The question is which metric we should use for measuring such information gain. In this paper, we showed that if we consider "information gain" as a reduction of uncertainty, the corresponding privacy metrics is similar to ℓ-diversity, whereas if we think to information gain as the *novelty* of the information, t-closeness is the corresponding metrics. Accordingly, the choice of the privacy risk metric depends on what kind of information we do not want to disclose, which in turn depends on the specific application, the tolerable level of information loss, and the attack model. The advantage of the proposed formulation in terms of information theory is that we can express all the different metrics using comparable units (bits), and, at least principle, use

all the tools of information theory for optimizing the privacy constraints, and, possibly, utility. The last point can be technically difficult in many cases, because expressing conditions on particular records largely increases the complexity of the optimization problem. Indeed, Eq. 14 must hold for each QI group, resulting in large number of constraints, so making difficult to obtain analytical results and even numerical solutions. Clearly, this is a relevant question to address in the near future, and, in particular, it is important to test the applicability of our approach to realistic cases.

Another possible extension is considering other data release scenarios, and the corresponding privacy models. For example, if we examine the scenario where the data publisher knowns in advance the data mining tasks, we can consider the *differential privacy* notion [15]. In a nutshell, the idea of differential privacy model, called ϵ*-differential privacy*, is to guarantee that small changes (one record) in the original database has a limited impact on the outcome of any statistical analysis on the data. More formally, let us consider a database X_0 (we do not distinguish between identifiers and sensitive attributes, so we will use X_0 to indicate the whole original database), and a randomized function F that has as output a real number[4], for example composed by a query on the original database that returns a numerical value plus an appropriately chosen random noise. We say that the random function F ensures ϵ-differential privacy if for all the datasets X_0

$$\sup_{X\in\mathcal{X}}\sup_{s\in S}\left|\ln\frac{p(F=s|X_0)}{p(F=s|X)}\right|\leq\epsilon \tag{16}$$

where $\mathcal{X}\equiv\{X|\delta(X_0,X)\leq 1\}$, i.e., the set of databases differing for at most one record from X_0 (δ is the Hamming distance), $S\equiv$ Range(F), where Range(F) is the set of possible outputs of the randomized function F, and ϵ is the privacy parameter. The closer ϵ is to 0, the stronger privacy is guaranteed (for a discussion on ϵ see [16]).

A good candidate for expressing the ϵ-differential privacy as an information metric is the surprise, I_1, of the database X_0 (not of a single record, as used above), i.e.

$$I_1(F,X_0)\equiv\sum_{s\in S}p(s|X_0)\log_2\frac{p(s|X_0)}{p(s)}\leq\theta_\epsilon \tag{17}$$

with

$$p(s)=\sum_{X\in\mathcal{X}}p(s|X)p(X)$$

and θ_ϵ is a parameter related to the ϵ of Eq. 16

Note that the condition expressed by Eq. 17 is different from Eq. 16, since in the first case we compare the probability $p(s|X_0)$ of having a certain outcome s from the analysis of the database X, with the average outcome over "similar" databases, $p(s)$, and then we average over all the possible outcomes. Whereas, in the original ϵ-differential privacy, we compare $p(s|X_0)$ directly with $p(s|X)$ (so, not with the average), and we take the maximum over s and over all the neighbor databases. Therefore, the latter condition is more conservative.

The formal relationship between Eq. 17 and Eq. 16, and between the corresponding thresholds, θ_ϵ and ϵ, as well as a more detailed analysis of possible advantages of expressing differential privacy in terms of information theory have to be investigated in future works.

[4]For sake of simplicity, we consider here a single function F that takes values in $\mathbb{R}$, but the same analysis can be applied to a function F which takes values in $\mathbb{R}^n$ or to a set of functions $\{F_1,F_2,\ldots\}$. An extension to non-numeric output has also been proposed in [23].

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216483. I also thank Stuart Short for the careful reading of the manuscript.

References

[1] Dakshi Agrawal and Charu C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255, New York, NY, USA, 2001. ACM.

[2] J. Bennett and S. Lanning. The Netflix Prize. *KDD Cup and Workshop*, 2007.

[3] M. Bezzi, I. Samengo, S. Leutgeb, and S.J. Mizumori. Measuring information spatial densities. *Neural computation*, 14(2):405–420, 2002.

[4] Michele Bezzi. An entropy-based method for measuring anonymity. In *Proceedings of the IEEE/CreateNet SECOVAL Workshop on the Value of Security through Collaboration*, pages 28–32, Nice, France, September 2007.

[5] Michele Bezzi. Quantifying the information transmitted in a single stimulus. *Biosystems*, 89(1-3):4–9, May 2007 (http://arxiv.org/abs/q-bio/0601038).

[6] Michele Bezzi. Expressing privacy metrics as one-symbol information. In *EDBT '10: Proceedings of the 2010 EDBT Workshops*, pages 1–5, New York, NY, USA, 2010. ACM.

[7] N. Blachman. The amount of information that y gives about X. *IEEE Transactions on Information Theory*, 14(1):27–31, 1968.

[8] D.A. Butts. How much information is associated with a particular stimulus? *Network: Computation in Neural Systems*, 14(2):177–187, 2003.

[9] D.M. Carlisle, M.L. Rodrian, and C.L. Diamond. California inpatient data reporting manual, medical information reporting for california. Technical report, Technical report, Office of Statewide Health Planning and Development, 2007.

[10] T. Dalenius. Finding a needle in a haystack-or identifying anonymous census record. *Journal of Official Statistics*, 2(3):329–336, 1986.

[11] AG DeWaal and L. Willenborg. Information loss through global recoding and local suppression. *Netherlands Official Statistics*, 14:17–20, 1999.

[12] M.R. DeWeese and M. Meister. How to measure the information gained from one symbol. *Network: Comput. Neural Syst*, 10:325–340, 1999.

[13] J. Domingo-Ferrer and D. Rebollo-Monedero. Measuring Risk and Utility of Anonymized Data Using Information Theory. *International workshop on privacy and anonymity in the information society (PAIS 2009)*, 2009.

[14] GT Duncan, S. Keller-McNulty, and SL Stokes. Disclosure risk versus data utility: The RU confidentiality map. *Technical paper, Los Alamos National Laboratory, Los Alamos, NM*, 2001.

[15] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin / Heidelberg, 2006.

[16] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer Berlin / Heidelberg, 2006.

[17] R. M. Fano. *Transmission of Information; A Statistical Theory of Communications*. MIT University Press, New York, NY, USA, 1961.

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

[19] Ali Inan, Murat Kantarcioglu, and Elisa Bertino. Using anonymized data for classification. In *ICDE '09: Proceedings of the 2009 IEEE International Conference on Data Engineering*, pages 429–440, Washington, DC, USA, 2009. IEEE Computer Society.

[20] Ninghui Li, Tiancheng Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115, April 2007.

[21] T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–526. ACM New York, NY, USA, 2009.

[22] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06)*, page 24, Washington, DC, USA, 2006. IEEE Computer Society.

[23] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. *Foundations of Computer Science, Annual IEEE Symposium on*, 0:94–103, 2007.

[24] David Rebollo-Monedero, Jordi Forne, and Josep Domingo-Ferrer. From t-closeness-like privacy to postrandomization via information theory. *IEEE Transactions on Knowledge and Data Engineering*, 99(1), 2009.

[25] Pierangela Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.*, 13(6):1010–1027, 2001.

[26] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423,623–656, 1948.

[27] P. Smyth and R. M. Goodman. An information theoretic approach to rule induction from databases. *IEEE Trans. on Knowl. and Data Eng.*, 4(4):301–316, 1992.