Expressing privacy metrics as one-symbol information

Michele Bezzi
SAP Labs
F-06560, Mougins, France
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 the disclosure risk metrics. We also show, with a simple example, how *l*-diversity and t-closeness can be represented in terms of two different, but equally acceptable, conditions on the information gain.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Statistical databases; H.1.1 [**Models and Principle**]: Systems and Information Theory

1. INTRODUCTION

Governmental agencies and corporates hold a huge amount of data containing information on individual people or companies (microdata). 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. Typically, data are contained in tables, and the attributes (columns) in the original table can be categorized, from disclosure perspective, in the following types:

PAIS'10, March 22–26, 2010, Lausanne, Switzerland.

- *Identifiers.* Attributes that explicitly identify individuals. E.g., Social Security Number, passport number, complete name.
- *Quasi-identifiers (Key attributes).* 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 ...

To reduce the risk, data holders use masking techniques (*anonymization*) for limiting disclosure risk in releasing sensitive datasets, such as 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 [7]. 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 key attributes. To assess the quality of the anonymization process, there is the need to measure the disclosure risk in the anonymized dataset and its utility (or equivalently the information loss). Both these quantities 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. Ideally, we should be able to express these metrics in terms of semantically "similar" measures, so we can easily combine them in a optimization problem. In Ref. [12], the authors proposed an information theoretic

framework to express average disclosure risk and information loss, using mutual information. The advantages of mutual information formulation are twofolds: 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 tradeoff problem [12]).

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 [13, 11, 9]) define minimal requirements (*worst-case scenarios*) for each entry (or combination of keys) 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 [10], *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 two types of one-symbol information (i.e., the contribution to mutual information by a single record), and express the disclosure risk metrics in terms of information theory (see Sect. 3). We also present a simple example, to point out that in presence of a constant average risk, the records at the risk may depend on the information metric used (Sect. 3.1). 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. 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. BACKGROUND

2.1 Disclosure Risk Metrics

Let us consider a dataset containing identifiers, key attributes, X, and sensitive attributes, W (for example as in Table 2). We create an anonymized version of such data, removing identifiers, and anonymizing key attributes ($\tilde{X}$), 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 [13] condition requires that *every* combination of key attributes 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 group (with minimal size of k records) sharing the same combination of keys 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 [11]. *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.

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, $I_3(x,Y) \equiv \sum_{y\in Y} p(y|x)[H(X) - H(X|y)]$ [4] is a definition based on weighted average of reduction of uncertainty, $I_4(x,Y) \equiv I(\{x,\bar{x}\};Y)$ [1] is the mutual information between Y and a set composed by two elements: x and the its complement in X: $\bar{x} \equiv X\backslash x$. For more details, see [3].

Because we are interested here in providing an information-theoretic framework, the more relevant for us is in terms of entropy,

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

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 group sharing the same key attributes is very dissimilar from the distribution over the whole set, an attacker may increase his knowledge on sensitive attributes (*skewness attack*, see [9] for details). t-closeness estimates this risk by computing the distance between the distribution of confidential attributes within the group and in the entire dataset. The authors in [9] proposed two ways to measure the distance, one of them has a straightforward relationship with mutual information (see Eq. 6 below), as we discuss in the next 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 risks; 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 key 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 $p(x,y)$ and $p(x|y)$ the corresponding joint and conditional probability functions. Following Shannon [14], we can define the mutual information $I(X;Y)$ as:

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

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

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

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 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 [14].

Mutual information is an average quantity, for some applications (see Sect. 3 and e.g., [3] 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. Unfortunately none of the proposed definitions have these properties (see Table 2.2), but each of them can capture different aspects of information transmission. In this paper, we will focus on two of them, following [5] referred as I_1 and I_2, which have applications for risk metrics in data anonymization, we refer the reader to [5, 3] for a more detailed analysis on all these quantities.

I_1, *Surprise*

Originally proposed by Fano [8], 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 (for the proof, see Appendix 2 in Ref. [5]). 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". 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.

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.

I_2, *Specific Information*

An entropy based definition has been proposed by De Weese and Meister [5] and it may be derived from Eq. (2), extracting the single stimulus contribution from the sum:

$$I_2(x;Y) = H(Y) - H(Y|x) = -\left[\sum_{y\in Y} p(y)\log_2 p(y) - p(y|x)\log_2 p(y|x)\right] \quad (3)$$

Here, information is identified with the reduction of entropy between marginal distribution $p(y)$ and conditional probability $p(y|x)$. This quantity captures how diverse are the entries in Y for a given entry 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 [5], 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 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

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

- k-anonymity. In case of suppression and generalization, we have that a single combination of keys 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:

$$H(X|\tilde{x}) \geq \log_2 k \tag{4}$$

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

$$I_2(X, \tilde{x}) \equiv H(X) - H(X|\tilde{x}) \leq \log_2 \frac{N}{k} \tag{5}$$

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. 4 holds also in case of perturbative masking [2], therefore I_2 can be used for any kind of masking transformations.

Averaging Eq. 5 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 [6], but we should be remind that this condition does not guarantee the k-anonymity for every $\tilde{x}$, i.e, it is necessary but not sufficient.

- t-closeness condition requires:

$$D(p(w|\tilde{x})||p(w)) \equiv \sum_{w \in W} p(w|\tilde{x}) \log_2 \frac{p(w|\tilde{x})}{p(w)} \leq t \tag{6}$$

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

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

$I_1(W, \tilde{x})$ is a measure of attribute disclosure risk for a combination of keys $\tilde{x}$, 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 ([12]),

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

Again, this 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:

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

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

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

$I_2(W, \tilde{x})$ is a measure of attribute disclosure risk for a combination of keys $\tilde{x}$, 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 [12].

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

This is the l-diversity condition on average. Again this is necessary but a not sufficient condition to satisfy l-diversity for each $\tilde{x}$.

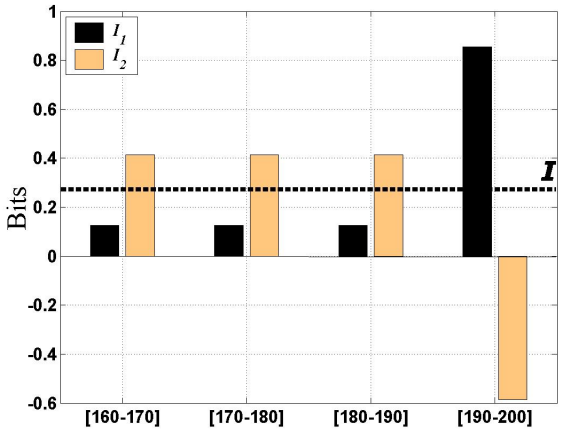

Figure 1: Values of attribute disclosure risk metrics: I_1 (Surprise) and I_2 for the different entries in the anonymized database. Dashed line indicated mutual information $I(\tilde{X}, W)$, i.e., the average of I_1 and I_2.

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, entries 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.

4. DISCUSSION AND CONCLUSIONS

In the original t-closeness paper [9], 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 finding the best tradeoff between privacy and 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. Clearly, this is an important question to address in the near future, and in particular if it is possible to find realistic cases where this problem is numerically tractable.

Acknowledgments

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[1].

5. REFERENCES

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

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

[3] 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).

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

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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