Measuring Risk and Utility of Anonymized Data Using Information Theory

Josep Domingo-Ferrer
Universitat Rovira i Virgili
Department of Computer Engineering and Maths
UNESCO Chair in Data Privacy
Av. Països Catalans 26
E-43007 Tarragona, Catalonia
josep.domingo@urv.cat

David Rebollo-Monedero
Technical University of Catalonia
Telematics Engineering Department
C. Jordi Girona 1-3
E-08034 Barcelona, Catalonia
david.rebollo@entel.upc.es

ABSTRACT

Before releasing anonymized microdata (individual data) it is essential to evaluate whether: i) their utility is high enough for their release to make sense; ii) the risk that the anonymized data result in disclosure of respondent identity or respondent attribute values is low enough. Utility and disclosure risk measures are used for the above evaluation, which normally lack a common theoretical framework allowing to trade off utility and risk in a consistent way. We explore in this paper the use of information-theoretic measures based on the notion of mutual information.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*statistical databases*; H.1.1 [**Models and Principles**]: Systems and Information Theory; K.4.1 [**Computers and Society**]: Public Policy Issues—*privacy*

General Terms

Privacy, Information theory

1. INTRODUCTION

When a set of microdata (individual data) containing the values of some attributes for individuals or companies is to be released by a statistical office for research or general use, anonymization is more often than not a legal requirement [6]. Anonymization means processing original microdata so as to obtain masked microdata that can be released in such a way that the respondent subjects to whom the microdata records refer cannot be re-identified (identity disclosure) and particular attribute values cannot be associated or disassociated with a particular respondent subject (attribute disclosure). In order to reduce the disclosure risk, anonymization distorts the data in some way (*e.g.* by perturbing them or reducing their detail); as distortion increases, disclosure risk can be expected to decrease. However, distortion should not be so severe that the anonymized data become useless, *i.e.*, that all information contained in the data is lost. The problem of optimizing the trade-off between disclosure risk and information loss is known as the statistical disclosure control (SDC) problem.

Information loss measures in SDC of microdata are usually based on the relative discrepancy between some statistics or models computed on the original data and on the masked data [4]. A critique to the above measures is that, for continuous attributes, relative discrepancies are unbounded[1] and difficult to combine with disclosure risk; the latter is most often measured as a probability of re-identification and is thus bounded between 0 and 1 [14].

Probabilistic information loss measures yielding a figure between $[0, 1]$ which can be readily compared to disclosure risk have been proposed [9]. Let θ be a population parameter (on the original data) and let $\hat{\Theta}$ be the corresponding sample statistic (on the masked data). If the number n' of records of the original data is large (> 100), then

$$Z = \frac{\hat{\Theta} - \theta}{\sqrt{\hat{\Theta}}}$$

can be assumed to follow a $N(0, 1)$ distribution.

A probabilistic information loss measure $pil(\theta)$ for parameter θ is the probability that the absolute value of the discrepancy Z is $\leq$ the actual discrepancy in the masked data:

$$pil(\theta) = 2 \cdot P(0 \leq Z \leq \frac{|\hat{\theta} - \theta|}{\sqrt{Var(\hat{\Theta})}})$$

Clearly, the more different is $\hat{\Theta}$ from θ, the greater is $pil(\theta)$.

However, getting information loss and disclosure risk measures bounded in the same $[0, 1]$ does not imply that their semantics are entirely comparable. Indeed, probabilistic information loss measures are actually distortions mapped to

[1] *E.g.* a statistic whose value is 0 on the original data and 0.1 on the masked data has an infinite relative discrepancy $(0.1 - 0)/0$.

$[0, 1]$, whereas disclosure risk is normally computed as a re-identification probability. If information loss and disclosure risk could be expressed with semantically similar measures, boundedness would be irrelevant: both measures could be unbounded without any problem.

1.1 Contribution and plan of this paper

The original contribution of this paper, developed in the following sections, is to explore the use of information-theoretic measures based on the notion of mutual information in view of providing a unified framework embracing information loss and disclosure risk. Section 2 motivates the use of information-theoretic measures. Section 3 describes information-theoretic measures for information loss. Section 4 describes information-theoretic measures for disclosure risk. Based on the previous measures, Section 5 presents models to trade off information loss against disclosure risk. Section 6 lists conclusions and issues for future research.

2. MOTIVATION

Unbounded loss measures based on relative discrepancies are very easy to understand, but rather difficult to trade off against a bounded risk. Probabilistic loss measures have the following strong points:

- They can be applied to the same usual statistics θ (means, variances, covariances, etc.) as measures based on relative discrepancies.
- They are bounded within $[0, 1]$, so they easily compare to disclosure risk.

Both relative-discrepancy and probabilistic loss measures lack an underlying theory allowing to optimize their trade-off with disclosure risk.

The mutual information $I(X;Y)$ between two random variables X and Y measures the mutual dependence of the two variables and is measured in bits. Mutual information can be expressed as a function of Shannon's entropy:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

$$= H(X) + H(Y) - H(X,Y)$$

where $H(X), H(Y)$ are marginal entropies, $H(X|Y), H(Y|X)$ are conditional entropies and $H(X,Y)$ is the joint entropy of X and Y.

Given a symmetric, positive definite matrix Σ, let $\Sigma^{1/2}$ denote the unique, symmetric, positive definite square root of Σ. The following can be stated about mutual information applied to random Gaussian data:

- If U and V are jointly Gaussian random vectors, and U' is the best linear estimate of U from V, then U' is a sufficient statistic, that is, $I(U';V) = I(U;V)$.
- If U and V are random, jointly Gaussian scalars with correlation coefficient ρ, then $I(U;V) = -\log\sqrt{1-\rho^2}$.
- More generally, if U and V are random, jointly Gaussian vectors with matrix correlation
$$P = \Sigma_U^{-1/2}\Sigma_{UV}\Sigma_V^{-1/2}$$
then
$$I(U;V) = -1/2\log\det(I - PP^t)$$
where P^t is the transpose of P, I the identity matrix and $\det(\cdot)$ is the determinant.

If mutual information can be used to express information loss or/and disclosure risk, then the machinery of information theory can be used to optimize the trade-off between both quantities.

3. INFORMATION-THEORETIC LOSS MEASURES

The attributes in a microdata set can be classified as:

- *Identifiers.* These are attributes that *unambiguously* identify the respondent. Examples are the passport number, social security number, name-surname, etc.
- *Key attributes.* These are attributes which identify the respondent with some degree of ambiguity. (Nonetheless, a combination of key attributes may provide unambiguous identification.) Examples are address, gender, age, telephone number, etc.
- *Confidential attributes.* These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.

We assume that the original microdata have been pre-processed to remove all identifiers from them. Let X, Y be, respectively, the key and confidential attributes in the original microdata set. Let X' be the key attributes in the masked microdata set (as in k-anonymization [13], we assume that only key attributes are masked). If we focus on the damage inflicted to key attributes [11], a possible information loss measure is the expected distortion $E(d(X,X'))$ where $d(x,x')$ is a distortion measure, *e.g.* $d(x,x') = ||x - x'||^2$.

A probably better option is to focus on how masking affects the statistical dependence between the key and confidential attributes. A possible measure for this is $I(X;Y) - I(X';Y)$.

LEMMA 1. *If X' is a randomized function of X, but not of Y, it holds that $I(X;Y) - I(X';Y) \geq 0$.*

Proof:. Given three random variables X_1, X_2 and X_3, define the conditional mutual information $I(X_1;X_2|X_3)$ as the expected value of $I(X_1;X_2)$ conditional to X_3, that is, $I(X_1;X_2|X_3) = E_{X_3}(I(X_1;X_2)|X_3)$. We have that

$$I(X';Y)+I(X;Y|X') = I((X,X');Y) = I(X;Y)+I(X';Y|X)$$

The hypothesis of the lemma implies that X' and Y are conditionally independent given X, that is, $I(X';Y|X) = 0$. Hence,

$$I(X';Y) + I(X;Y|X') = I(X;Y)$$

Since $I(X;Y|X') \geq 0$, we have that $I(X';Y) \leq I(X;Y)$. □

Let us now compare the mutual information with the more usual information loss measures based on the mean square error (MSE) and correlations.

3.1 Mutual information vs. MSE

The MSE $E(d(X,X')) = E(||X-X'||^2)$ seems better adapted than $I(X;X')$ to measuring how well statistical properties are preserved. For example:

- A zero MSE between X and X', that is, $E(d(X,X')) = 0$ implies $X = X'$.
- However, $I(X;Y) - I(X';Y) = 0$ only implies that X and X' are bijectively related.

Nonetheless, MSE and the mutual information are not that different, both belonging to the family of so-called Bregman divergences [10, 2].

3.2 Mutual information vs. correlations

The difference $I(X;Y) - I(X';Y)$ bears some resemblance to the relative discrepancy between correlation matrices proposed as an information loss measure in [4]. However, mutual information measures the general dependence between attributes, while the correlation measures only the linear dependence; thus the former seems superior [8]. It will be shown below that, under some assumptions, preserving mutual information preserves the covariance matrix up to a constant factor.

4. INFORMATION-THEORETIC RISK MEASURES

The mutual information $I(X';X)$ between the released and the original key attributes is a measure of *identity disclosure* (defined in the introduction above). Note that $I(X';X)$ was previously regarded as a possible information loss measure (which it is for key attributes).

The mutual information $I(X';Y)$ between the released key attributes and the confidential attributes is a measure of *attribute disclosure* (defined in the introduction above). Measuring risk as $I(X';Y)$ conforms to the t-closeness privacy property [7] requiring that the distance between the distribution of Y within records sharing each combination of values of X' and the distribution of Y in the overall dataset be no more than t.

5. LOSS-RISK OPTIMIZATION

Several combinations of the above loss and risk measures can be used when trying to optimize the trade-off of information loss and disclosure risk. Two approaches are conceivable:

- Place an upper-bound constraint on the information loss D and minimize the disclosure risk R.
- Place an upper-bound constraint on the disclosure risk R and minimize the information loss D (more natural in SDC).

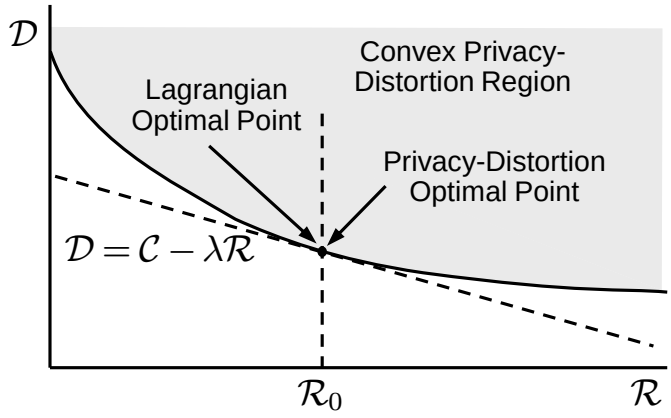

Figure 1: Risk-loss as Lagrangian rate-distortion optimization. Rate stands for risk in the privacy setting

By combining the above two approaches with the various loss and risk measures sketched above, several optimization models are obtained.

5.1 Model 1

In this model, disclosure risk is minimized while keeping information loss below a certain upper bound. Disclosure risk is measured using mutual information and information loss is measured as the expected distortion. Hence:

$$R(D) = \inf_{p_{X'|X}} I(X';Y)$$

$$\text{subject to } \; E(d(X,X')) \leq D$$

for a certain pre-specified maximum tolerable loss D.

Model 1 was related in [11] to the rate-distortion function optimization in information theory (see Figure 1): the risk R was assimilated to the rate and the loss D to the distortion. An optimal random perturbation $p_{X'|X}(x'|x)$ for key attributes was obtained. Let $d = D/\sigma_X^2$ be the normalized distortion. For the case of univariate Gaussian, real-valued X and Y, a closed form of the minimum was obtained. The idea is to compute X' as $X' = (1-d)X + dZ$, where the noise Z is distributed according to $N(\mu, \frac{1-d}{d}\sigma_X^2)$ independently of X and Y. This yields:

$$R_{inf} = -\frac{1}{2}\log(1-(1-d)\rho_{XY}^2)$$

$$p^{opt}_{X'|X}(x'|x) = N((1-d)x + d\mu_X, d(1-d)\sigma_X^2)$$

5.2 Model 2

If we maintain the same risk and loss measures, but take the more natural approach of minimizing D for a maximum tolerable risk R, we get

$$D(R) = \inf_{p_{X'|X}} E(d(X,X'))$$

$$\text{subject to } I(X';Y) \leq R.$$

This problem could be related to optimizing the distortion-rate function optimization in quantization. This again yields an optimal perturbation $p_{X'|X}$, which can be computed by solving the above model.

5.3 Models 3 and 4

Model 3 below results from miniziming disclosure risk while keeping information loss below a certain level, and using mutual information for measuring both the disclosure risk and the information loss:

$$R(D) = \inf_{p_{X'|X}} I(X;X')$$

$$\text{subject to } I(X;Y) - I(X';Y) \leq D.$$

Finally, Model 4 is obtained by maintaining the same measures, but minimizing the information loss while keeping disclosure risk below an upper bound:

$$D(R) = \inf_{p_{X'|X}} I(X;Y) - I(X';Y)$$

$$\text{subject to } I(X;X') \leq R.$$

5.4 Model 4 and synthetic data generation

Synthetic data generation, that is, generation of random simulated data preserving some properties of original data, can be viewed as a form of masking by perturbation [1]. If we want to generate synthetic key attributes X' in such a way that the connection between key attributes and confidential attributes is minimally affected, we can use Model 4 to compute $p_{X'|X}$. Synthetic X' can be generated by drawing from $p_{X'|X}$.

5.5 Mutual information vs. covariance preservation

We justify that preserving mutual information (that is, achieving $D(R) = 0$ in Model 4) preserves the covariance matrix (up to a constant factor).

Let X and Y be zero-mean, jointly Gaussian random variables, $\mathbb{R}$- and $\mathbb{R}^n$-valued, respectively. Let $X' = a^T Y$ be the best linear MSE estimate of X given Y, for $a \in \mathbb{R}^k$. Then $a = \Sigma_{XY}\Sigma_Y^{-1}$.

On the one hand, X' is a sufficient statistic for X given Y, that is, $I(X';Y) = I(X;Y)$ (see [12]).

On the other hand, the covariance matrix is preserved when replacing X by X' because

$$\Sigma_{X'Y} = \Sigma_{XY}\Sigma_Y^{-1}\Sigma_Y = \Sigma_{XY}.$$

6. CONCLUDING REMARKS AND FUTURE RESEARCH

Information loss measures based on relative discrepancies are awkward to combine with risk measures in order to optimize the risk-loss trade-off. Probabilistic loss measures are a step forward, but lack a theoretical framework. We have explored here loss and risk measures based on information theory, namely on mutual information. Models for optimizing the information-theoretic risk-loss trade-off when perturbing data and generating synthetic data have been presented. It has been shown that preserving mutual information offers covariance matrix preservation.

However, this paper is intended to be a starting point rather than a conclusive contribution. The information-theoretic measures and the models discussed above are just a first step. A number of issues for future research lie open ahead:

- Relate Model 2 with distortion-rate function optimization, a well-known problem in quantization. This should be done in a way analogous to the connection between Model 1 and rate-distortion function optimization established in [11].
- In the context of synthetic data generation, devise information-theoretic loss measures whose minimization is equivalent to preserving a given model.
- Whenever possible, find closed-form expressions for the optimal $p_{X'|X}$ transformations.
- If a closed-form expression is not possible, look for a convex optimization problem to be solved numerically as the next most attractive option.
- Investigate the connection of mutual information with information loss metrics other than MSE, like [5, 15, 3].

Acknowledgments and disclaimer

This work was partly funded by the Spanish Government through projects CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES" and TSI2007-65406-C03-01 "E-AEGIS". The first author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia; he holds the UNESCO Chair in Data Privacy, but his views do not necessarily reflect the position of UNESCO nor commit that organization.

7. REFERENCES

[1] J. M. Abowd and L. Vilhuber. How protective are synthetic data? In J. Domingo-Ferrer and Y. Saygin, editors, *Privacy in Statistical Databases*, volume 5262 of *Lecture Notes in Computer Science*, pages 239–246, Berlin Heidelberg, 2008. Springer.

[2] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math., Math. Phys.*, 7:200–217, 1967.

[3] J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 70–78, 2008.

[4] J. Domingo-Ferrer and V. Torra. Disclosure protection methods and information loss for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 91–110, Amsterdam, 2001. North-Holland.

[5] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 279–288, 2002.

[6] J. Lane, P. Heus and T. Mulcahy. Data access in a cyber world: making use of cyberinfrastructure. *Transactions on Data Privacy*, 1(1): 2–16, 2008.

[7] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: privacy beyond k-anonymity and l-diversity. In *Proceedings of the IEEE ICDE 2007*, 2007.

[8] W. Li. Mutual information functions vs correlation functions. *Journal of Statistical Physics*, 60:823–837, 1990.

[9] J. M. Mateo-Sanz, J. Domingo-Ferrer, and F. Sebé. Probabilistic information loss measures in confidentiality protection of continuous microdata. *Data Mining and Knowledge Discovery*, 11(2), 2005. 181-193.

[10] D. Rebollo-Monedero. *Quantization and Transforms for Distributed Source Coding*. PhD thesis, Stanford University, 2007.

[11] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer. From t-closeness to pram and noise addition via information theory. In *Privacy in Statistical Databases-PSD 2008*, volume 5262 of *Lecture Notes in Computer Science*, pages 100–112, Berlin Heidelberg, 2008.

[12] D. Rebollo-Monedero, S. Rane, A. Aaron, and B. Girod. High-rate quantization and transform coding with side information at the decoder. *Signal Processing*, 86(11):3160–3179, 2006.

[13] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[14] M. Trottini. *Decision models for data disclosure limitation*. PhD thesis, Carnegie Mellon University, 2003. http://www.niss.org/dgii/TR/Thesis-Trottini-final.pdf.

[15] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu. Utility-based anonymization using local recoding. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–790, 2006.