General Confidentiality and Utility Metrics for Privacy-Preserving Data Publishing Based on the Permutation Model

Josep Domingo-Ferrer, *Fellow, IEEE,* Krishnamurty Muralidhar and Maria Bras-Amorós

Abstract—Anonymization for privacy-preserving data publishing, also known as statistical disclosure control (SDC), can be viewed under the lens of the permutation model. According to this model, any SDC method for individual data records is functionally equivalent to a permutation step plus a noise addition step, where the noise added is marginal, in the sense that it does not alter ranks. Here, we propose metrics to quantify the data confidentiality and utility achieved by SDC methods based on the permutation model. We distinguish two privacy notions: in our work, anonymity refers to subjects and hence mainly to protection against record re-identification, whereas confidentiality refers to the protection afforded to attribute values against attribute disclosure. Thus, our confidentiality metrics are useful even if using a privacy model ensuring an anonymity level *ex ante*. The utility metric is a general-purpose metric that can be conveniently traded off against the confidentiality metrics, because all of them are bounded between 0 and 1. As an application, we compare the utility-confidentiality trade-offs achieved by several anonymization approaches, including privacy models (k-anonymity and ϵ-differential privacy) as well as SDC methods (additive noise, multiplicative noise and synthetic data) used without privacy models.

Index Terms—Privacy, anonymity, confidentiality, utility, data anonymization, statistical disclosure control, permutation model

1 Introduction

Since the turn of the century, we are fully immersed in the information society. Most human activities leave digital traces that someone collects and stores. Social media, the internet of things, bank transactions, purchases at stores are just a few ways of gathering data on people. Such a massive data collection has many advantages: increased business opportunities, better and more rigorous research and, in general, rosier prospects of improving the well-being of the human race.

Yet, accumulating, sharing and publishing personally-identifiable information (PII) has also a disquieting side, as it invades the privacy of the subjects to whom PII relate; a famous example is the teenager pregnancy guess reported in [14]. Data protection legislation, epitomized by the EU General Data Protection Regulation [21], tries to protect citizens by restricting the accumulation of PII. Anonymizing PII, *i.e.* turning them into data that are not personally identifiable but still retain substantial analytical utility, is a way to enable data analysis, sharing and even publishing without violating the data protection laws.

Anonymization for privacy-preserving data publishing is also known as statistical disclosure control (SDC, [25]). The usual setting in anonymization is for a *data controller* (the entity that manages and releases the data, and often owns them) to hold the original data (with the original responses by the *subjects*) and modify them to reduce the disclosure risk. Then the controller publishes the anonymized data or shares them with *users*, typically data analysts or researchers —who expect the anonymized data to be still useful. It may occur that some of the users behave as *intruders* and try to perform disclosure attacks on the anonymized data. Disclosure can be of two types:

- *Re-identification disclosure*, whereby the intruder determines the subject to whom an anonymized data item corresponds;
- *Attribute disclosure*, in which the anonymized data help the intruder to estimate the value of a confidential attribute for a certain subject.

Data at the individual level, such that each record corresponds to one individual subject (person, enterprise, etc.), are known as microdata. From microdata, other formats can be derived, such as tables (the traditional output of national statistical institutes) and online queryable databases (that answer statistical queries on an underlying microdata set). Here, we will focus on microdata.

The traditional approach to anonymization, still dominant among statistical agencies, can be called "utility-first": the controller runs an SDC method [25] with a heuristic parameter choice and with suitable utility preservation properties on the microdata set. After that, the controller measures the risk of disclosure, which she can do empirically by attempting record linkage between

J. Domingo-Ferrer and Maria Bras-Amorós are with the UNESCO Chair in Data Privacy, CYBERCAT-Center for Cybersecurity Research of Catalonia, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, e-mail {josep.domingo,maria.bras}@urv.cat.
K. Muralidhar is with the Price College of Business, University of Oklahoma, 307 West Brooks, Adams Hall Room 10, Norman OK 73019-4007, USA, e-mail krishm@ou.edu.

the original and the anonymized data sets [37], or analytically by using generic metrics (*e.g.* [28]) or metrics tailored to a specific SDC method (*e.g.* [18] for sampling-based SDC). If the controller deems the remaining risk too high, she re-runs the anonymization method with more confidentiality-stringent parameters and probably more utility sacrifice.

Whereas most utility-first SDC methods obtain each anonymized record by masking a certain original record, synthetic data are an exception. In this case, the anonymized data set consists of synthetic/simulated data that preserve a set of utility characteristics of the original data set. Since there is no direct mapping between original and synthetic records, synthetic data are often regarded as the safest utility-first approach. Unfortunately, this lack of mapping also makes it difficult to quantify the confidentiality actually achieved, because many confidentiality metrics need to compare each anonymized record with its corresponding original record.

An alternative anonymization approach can be termed "privacy-first" and is based on the notion of *privacy model*, which is a condition dependent on a parameter that guarantees an upper bound on the risk of reidentification disclosure and perhaps also on the risk of attribute disclosure by an intruder. Well-known privacy models include k-anonymity [35] and its extensions, as well as ϵ-differential privacy [16]. The controller can enforce a certain privacy model using one or several specific SDC methods whose parameters are a function of the model parameters. For example, k-anonymity can be attained with a combination of generalization and suppression or with microaggregation [12]; ϵ-differential privacy is normally attained via noise addition. There may be two issues with the privacy-first approach: on the one side, if the controller chooses too stringent a parameter for the privacy model, the utility of the anonymized data may be too low; on the other side, if she chooses too relaxed a parameter, the protection given by the privacy model may be insufficient.

Thus, no matter whether the controller follows the utility-first or the privacy-first approaches, she needs to measure the utility and the protection provided by a certain anonymization method. However, SDC methods for microdata rely on a diversity of principles [25], and this makes it difficult to analytically compare their utility and data protection properties [15]; this is why one usually resorts to empirical comparisons [11].

Contribution and plan of this paper

In this paper, we present new *confidentiality* and *utility* metrics for anonymized data. Let us briefly define these notions. We use utility in the customary sense of preserving the statistical properties of the original data. On the other hand, confidentiality is one of the two main privacy notions in statistical disclosure control, the other being anonymity. Whereas anonymity refers to subjects and hence mainly to protection against record re-identification, confidentiality refers to the protection afforded to attribute values against attribute disclosure.

Specifically, we exploit the unified view of anonymization afforded by the permutation model [8] to derive bounded confidentiality metrics for microdata anonymization that are based on the relative amounts of permutation undergone by the different attributes of a data set. We then give a bounded utility metric that can be used to evaluate the trade-off between utility and confidentiality and also to compare this trade-off among different utility-first SDC methods as well as among different privacy models.

In Section 2, we give background on anonymity vs confidentiality, on the permutation model and on canonical correlation, a primitive that we will use to construct our confidentiality metrics. Section 3 makes the case for using permutation to assess confidentiality and utility. In Section 4 we present the confidentiality metrics and in Section 5 we present the utility metric. Empirical work on the operation of the new metrics is described in Section 6; we compare the utility-confidentiality trade-offs attained by several anonymization approaches, including privacy models (k-anonymity and differential privacy) and utility-first SDC methods (additive noise, multiplicative noise and synthetic data). Section 7 reviews related work. Finally, conclusions and lines of future research are summarized in Section 8.

2 Background

2.1 Anonymity vs confidentiality

As mentioned above, we use anonymity to refer to protection against record re-identification and confidentiality to protection against attribute disclosure. In fact, a given anonymity level can co-exist with different levels of confidentiality:

- In k-anonymity, if the original data set happens to be already k-anonymous (this is quite unlikely, but possible), then no SDC masking is needed for anonymity. Even though the anonymity level is k, confidentiality is zero, because the original attribute values are not modified. On the other hand, if attributes need to be heavily masked to attain k-anonymity, confidentality is in general nonzero.
- In ϵ-differential privacy, if the original data set or query have very low sensitivity (they depend very little on the absence or presence of any single subject in the data), very little noise needs to be added to satisfy the model. Thus in this case, anonymity is inversely proportional to ϵ, but confidentiality is low. Conversely, if sensitivity is high, a lot of noise is added and confidentiality is high even though ϵ is the same.
- No matter whether k-anonymity or ϵ-differential privacy are used, for the same anonymity level confidentiality can be expected to grow with the number of attributes. Indeed, in k-anonymity, the

more attributes, the less homogeneous the records in the k-anonymous classes, and the higher the distortion when generalizing the records in a class or replacing them by the centroid record. On the other hand, in ϵ-differential privacy, the privacy budget ϵ needs to be split among the attributes, which means that the more attributes, the less budget per attribute and the more noise needs to be added to each attribute to achieve ϵ-differential privacy for the overall data set; hence, confidentiality increases.

2.2 The permutation model

In [8], we introduced the permutation model of anonymization. Consider an original attribute $X = \{x_1, x_2, \cdots, x_n\}$ and its anonymized version $Y = \{y_1, y_2, \cdots, y_n\}$. Assume X and Y can be ranked (even categorical nominal attributes can be ranked, using a semantic distance [9]). For $i = 1$ to n: compute $j = \text{Rank}(y_i)$ and let $z_i = x_{(j)}$, where $x_{(j)}$ is the value of X of rank j. Then call attribute $Z = \{z_1, z_2, \cdots, z_n\}$ the *reverse-mapped* version of X. For example, if original value $x_1 \in X$ is anonymized as $y_1 \in Y$, and y_1 is, say, the 3rd smallest value in Y, then we take z_1 to be the 3rd smallest value in X.

If there are several attributes in the original data set $\mathbf{X}$ and anonymized data set $\mathbf{Y}$, the previous reverse-mapping procedure is conducted for each attribute; call $\mathbf{Z}$ the data set formed by reverse-mapped attributes. Note that: (i) a reverse-mapped attribute Z is a permutation of the corresponding original attribute X; (ii) the rank order of Z is the same as the rank order of Y. Therefore, *any microdata anonymization technique is functionally equivalent to permutation* (from $\mathbf{X}$ into $\mathbf{Z}$) *followed by residual noise addition* (from $\mathbf{Z}$ into $\mathbf{Y}$). The noise added is residual, because the ranks of $\mathbf{Z}$ and $\mathbf{Y}$ are the same. See Figure 1.

2.3 Canonical correlation

Correlations are range-independent metrics that assess the relationships between pairs of attributes. Canonical correlation analysis (CCA) is a multivariate statistics technique to measure the correlation between two vectors of random variables [24]. We will use CCA to assess the correlation between the original data set $\mathbf{X}$, that can be viewed as a sample of a vector $\mathbf{x}$ of random variables $X^1, \ldots, X^m$ (the original attributes), and the anonymized data set $\mathbf{Y}$, that can be viewed as a sample of a vector $\mathbf{y}$ of random variables $Y^1, \ldots, Y^m$ (the anonymized attributes).

Denote by $\mathbf{C}_{XX}$ and $\mathbf{C}_{YY}$ the respective covariance matrices of data sets $\mathbf{X}$ and $\mathbf{Y}$, and by $\mathbf{C}_{XY}$ the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$.

The canonical correlations between $\mathbf{X}$ and $\mathbf{Y}$ can be found by solving the eigenvalue equations

$$\begin{cases} \mathbf{C}_{XX}^{-1}\mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}\mathbf{C}_{YX}\mathbf{w}_X = \rho^2\mathbf{w}_X \\ \mathbf{C}_{YY}^{-1}\mathbf{C}_{YX}\mathbf{C}_{XX}^{-1}\mathbf{C}_{XY}\mathbf{w}_Y = \rho^2\mathbf{w}_Y \end{cases} \quad (1)$$

where the eigenvalues ρ^2 are the squared *canonical correlations* and the eigenvectors $\mathbf{w}_X$ and $\mathbf{w}_Y$ are the normalized canonical correlation *basis vectors*. If both $\mathbf{X}$ and $\mathbf{Y}$ have m attributes, there are m non-zero solutions of Equations (1), that is, there are m canonical correlations $\rho_1, \ldots, \rho_m$, where we have written them in non-increasing order $\rho_i \geq \rho_j$ if $i \leq j$.

Only one of the two equations (1) needs to be solved, say the first one, because $\mathbf{w}_X$ and $\mathbf{w}_Y$ are related as follows:

$$\begin{cases} \mathbf{C}_{XY}\mathbf{w}_Y = \rho\kappa_X\mathbf{C}_{XX}\mathbf{w}_X \\ \mathbf{C}_{YX}\mathbf{w}_X = \rho\kappa_Y\mathbf{C}_{YY}\mathbf{w}_Y \end{cases}$$

where

$$\kappa_X = \kappa_Y^{-1} = \sqrt{\frac{\mathbf{w}_Y^T\mathbf{C}_{YY}\mathbf{w}_Y}{\mathbf{w}_X^T\mathbf{C}_{XX}\mathbf{w}_X}}.$$

Canonical correlation ρ_1 turns out to be the correlation between $u_1 = \mathbf{x}^T\mathbf{w}_X^1$ and $v_1 = \mathbf{y}^T\mathbf{w}_Y^1$, where these linear combinations of $\mathbf{x}$ and $\mathbf{y}$ are the ones yielding the highest correlation. Then ρ_2 is the correlation between $u_2 = \mathbf{x}^T\mathbf{w}_X^2$ and $v_2 = \mathbf{y}^T\mathbf{w}_Y^2$, where these linear combinations yield the highest correlation among the combinations such that u_1 and u_2 are uncorrelated and v_1 and v_2 are uncorrelated. And so on with ρ_3 up to ρ_m.

See [23] for more details on CCA.

3 The Permutation Model and the Assessment of Confidentiality

3.1 Confidentiality and the permutation matrices

According to the permutation model, the protection offered by an anonymization method comes from two alterations of the original data $\mathbf{X}$: on the one hand, alteration of the *ranks* of attribute values (that is, permutation of $\mathbf{X}$ into $\mathbf{Z}$) and, on the other hand, addition of noise (to transform $\mathbf{Z}$ into $\mathbf{Y}$) such that it does not entail any further change in the ranks.

Hence, the main confidentiality protection principle turns out to be permutation. More precisely, let us consider the m permutation matrices that respectively represent the permutation undergone by each of the m attributes. The following holds:

- If and only if the m permutation matrices are identical, permutation is trivial in the sense that entire records are swapped, which provides no confidentiality.
- If permutation is non-trivial, it provides confidentiality as long as the intruder cannot accurately recreate the m permutation matrices.

The maximum-knowledge intruder assumed in the permutation model knows $\mathbf{X}$ and $\mathbf{Y}$. Thus, this intruder is stronger than any other prior intruder in the data set anonymization literature. Furthermore, he is purely malicious: even though he already knows the original data set, he wants to find the mapping between the original and the anonymized records, in order to recreate the permutation matrices and thereby discredit the controller having anonymized the data.

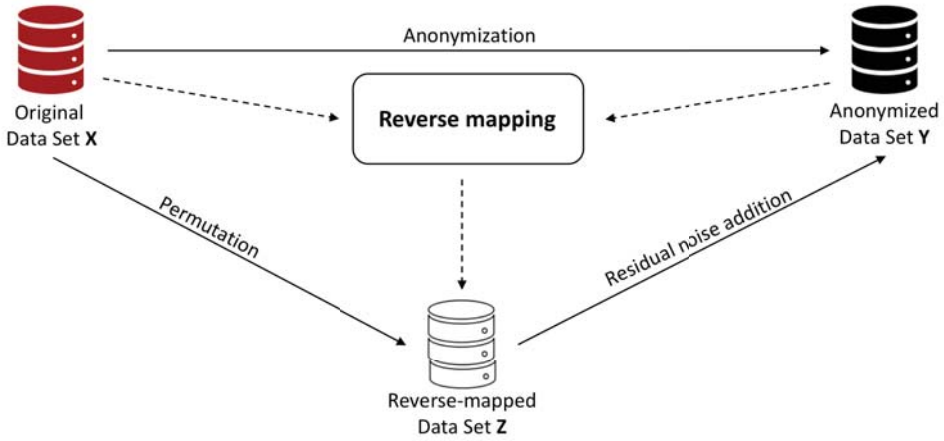


Fig. 1. The permutation model of anonymization

In the case of synthetic data, such a natural mapping between original and anonymized records does not exist, but the permutation model tells us that replacing original by synthetic data can still be viewed as a permutation. A possible approach is for the intruder to sort $\mathbf{X}$ by the j-th original attribute and $\mathbf{Y}$ by the j-th anonymized attribute, for any $1 \leq j \leq m$, and hypothesize that the i-th sorted original record corresponds to the i-th anonymized record. From that hypothesized mapping, the intruder may derive hypothesized permutation matrices.

In the remainder of this paper, $\mathbf{X}$ and $\mathbf{Y}$ will represent the *ranks* of the attributes in the original and anonymized data sets, respectively, rather than their magnitude values. The reason is that our interest lies in the permutation of the ranks.

3.2 Confidentiality and disclosure

The more accurate the intruder's estimation of the permutation matrices, the less confidentiality is left and the more chances for disclosure.

Re-identification disclosure cannot be prevented unless there is a change in ranks, that is, unless $\mathbf{X} \neq \mathbf{Z}$ and $\mathbf{Z}$ is not a trivial permutation of $\mathbf{X}$. If $\mathbf{X} = \mathbf{Z}$ or both data sets are related by a trivial permutation, it is immediate for an intruder to link each anonymized record in $\mathbf{Y}$ to the record in $\mathbf{X}$ that has the same ranks for all attributes. Once the subject's original record has been determined, re-identification becomes possible.

Let us now look at protection against attribute disclosure. If $\mathbf{X} = \mathbf{Z}$ or one data set is a trivial permutation of the other, then protection comes only from noise addition that transforms $\mathbf{Z}$ into $\mathbf{Y}$ but does not change ranks. Unless data are very sparse, the noise has to be necessarily small, which affords little protection against attribute disclosure. Thus, in general, protection against attribute disclosure necessitates also changes in ranks.

4 Bounded confidentiality metrics

In this section we present three confidentiality metrics. To compute the first two metrics, one needs to know the mapping between records in the original data set and records in the anonymized data set. The reason is that they are based on canonical correlations and therefore they require $\mathbf{C}_{XY}$, the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$. The third metric is based on the second metric but it does not need to know the mapping between records in $\mathbf{X}$ and $\mathbf{Y}$. Thus, it is especially suitable for anonymization via synthetic data.

All three metrics use Spearman's rank-based correlation. This is a non-parametric (distribution-free) measure of the strength of the monotonic association between two attributes. It can be used even when attributes are measured in ordinal scales. In certain situations (such as when the relationship between the attributes is not linear and/or their distributions are not normal), Pearson's product-moment correlation —more usual and based on attribute values rather than ranks— can be unreliable. In these situations, Spearman's correlation based on ranks is a better measure than Pearson's [22]. Furthermore, Spearman's correlation has also been shown to be more robust than Pearson's in the presence of outliers [38] and provides higher power for tests of association [19].

Since the proposed metrics are based on the permutation model, they implicitly assume that the number of anonymized records is the same as the number of original records. However, there are SDC methods that may decrease or increase the number of records. This is not problematic as long as the party computing the metrics is the controller who has performed the anonymization — which is the usual situation, because the purpose of the metrics is to guide anonymization. Indeed, if the number of records is reduced due to suppression or sampling, the controller knows which records have been suppressed and may discard them to obtain an original data set with the same number of records as the anonymized data set. If the number of records decreases or increases as

a result of synthetic data generation, then the controller (or in fact anyone) can sample whichever of the original or anonymized data sets is larger so that the sampled data set has the same number of records as the smaller data set.

In the sequel, we assume that the attributes in $\mathbf{X}$ and $\mathbf{Y}$ are numerical or ordinal categorical, so that canonical correlations and covariances can be computed on them. For nominal categorical attributes, an ontology can be used whereby a semantically coherent numerical value such as marginality [9] can be assigned to each nominal category.

4.1 Confidentiality metric from the largest canonical correlation

A first approach is to measure confidentiality based on the *largest* canonical correlation ρ_1 between the original data set $\mathbf{X}$ and the anonymized data set $\mathbf{Y}$.

Since canonical correlations are bounded in $[-1, 1]$, we can define our permutation-based confidentiality metric as

$$CM1(\mathbf{X}, \mathbf{Y}) = 1 - \rho_1^2, \tag{2}$$

where, as mentioned above, $\mathbf{X}$ and $\mathbf{Y}$ contain ranks rather than values.

According to Expression (2), top confidentiality ($CM1(\mathbf{X}, \mathbf{Y}) = 1$) is attained when ranks of attributes in $\mathbf{Y}$ are independent of the ranks of attributes in $\mathbf{X}$, which means that the anonymization can be viewed as a random permutation.

In contrast, zero confidentiality ($CM1(\mathbf{X}, \mathbf{Y}) = 0$) is achieved if the ranks in $\mathbf{X}$ and $\mathbf{Y}$ are the same for *at least* one original attribute X^i and one anonymized attribute Y^i, that is, if the anonymization method leaves all ranks unchanged for at least one original attribute. Ranks can stay unchanged either because the values in X^i and Y^i are the same or because the original values have been perturbed so little that ranks are unaffected. Note that this notion of confidentiality is quite strict: leaving a single attribute unprotected brings the confidentiality metric down to zero. We use the same notion in the next two metrics.

4.2 Confidentiality metric from all canonical correlations

A more refined approach is to take all m canonical correlations into account when measuring confidentiality. In [27], [26], a connection between canonical correlations and mutual information is shown if the collated data sets $\mathbf{T} = (\mathbf{X}, \mathbf{Y})$, where $\mathbf{T}$ has $2m$ attributes and n records, follow an elliptically symmetrical distribution (a generalization of the multivariate Gaussian). The connection is:

$$I(u_i; v_i) = \ln\left(\frac{1}{1-\rho_i^2}\right), \tag{3}$$

where u_i and v_i are the linear combinations yielding ρ_i.

Since the pairs $\{(u_i, v_i) : i = 1, 2, \ldots, m\}$ are mutually uncorrelated, we can add Expression (3) for all pairs to obtain the mutual information between the original data set $\mathbf{X}$ and the anonymized data set $\mathbf{Y}$:

$$I(\mathbf{X}; \mathbf{Y}) = \sum_{i=1}^{m} I(u_i; v_i)$$

$$= \sum_{i=1}^{m} \ln\left(\frac{1}{1-\rho_i^2}\right) = \ln\left(\frac{1}{\prod_{i=1}^{m}(1-\rho_i^2)}\right). \tag{4}$$

From Expression (4) and by analogy with Expression (2), we can derive the following confidentiality metric that has the advantages of taking all canonical correlations into account and being related to the mutual information between the original and the anonymized data sets.

$$CM2(\mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{m}(1-\rho_i^2) \left[= e^{-I(\mathbf{X};\mathbf{Y})}\right]. \tag{5}$$

The second equality between brackets in Expression (5) can only be guaranteed if the above distributional assumptions hold, in which case Expression (5) can be justified using mutual information.

Regardless of the distributional assumptions, $CM2(\mathbf{X}, \mathbf{Y})$ can be computed from the canonical correlations and the following holds:

- Top confidentiality $CM2(\mathbf{X}, \mathbf{Y}) = 1$ is reached when the anonymized data set and the original data sets tell nothing about each other, which is the same as saying that mutual information between them is $I(\mathbf{X}; \mathbf{Y}) = 0$.
- Zero confidentiality $CM2(\mathbf{X}, \mathbf{Y}) = 0$ occurs if at least one of the canonical correlations is 1. This occurs if at least one original attribute is disclosed when releasing $\mathbf{Y}$. Since ρ_1 is the largest correlation, this means that we have $CM2(\mathbf{X}, \mathbf{Y}) = 0$ if and only if $\rho_1 = 1$, in which case we also have that the metric of Expression (2) is $CM1(\mathbf{X}, \mathbf{Y}) = 0$.

4.3 Mapping-free confidentiality metric

The confidentiality metrics defined in Sections 4.1 and 4.2 implicitly assume a known mapping between records in the original data set $\mathbf{X}$ and records in the anonymized data set $\mathbf{Y}$ to compute canonical correlations (in particular to compute the covariance $\mathbf{C}_{XY}$ between $\mathbf{X}$ and $\mathbf{Y}$). Such a mapping is naturally known to the controller if she obtains each record in $\mathbf{Y}$ by masking a record in $\mathbf{X}$, via noise addition or another SDC method.

However, in the case of synthetic data generation there is no natural mapping between original and anonymized values. Indeed, data synthesis generates a complete data set by using the distributional characteristics of the original data set, rather than the original data themselves. Many synthetic data generation procedures have been proposed in the literature [2], [32], [33], [13]. Since the

individual records in the synthetic data set $\mathbf{Y}$ do not depend on the *individual* records in the original data set $\mathbf{X}$, the correlation between original data and synthetic data can be expected to be zero, subject to the sampling error. Consequently, metrics $CM1$ and $CM2$ will always be close to 1, even if the synthetic data leak the original data (see examples further down in this section and in Section 6).

In this section, we propose a confidentiality metric that does not need to know in advance the mapping between records of $\mathbf{X}$ and $\mathbf{Y}$. It uses the permutation model and more specifically reverse mapping [8], whereby the values of an anonymized attribute can be viewed as a permutation of the values of the corresponding original attribute (plus perhaps a marginal amount of noise). Hence, even if anonymized values look uncorrelated with the original values, a permutation linking anonymized and original values exists. As mentioned in Section 3.1, a maximum-knowledge intruder knowing $\mathbf{X}$ and $\mathbf{Y}$ can try to guess the mapping between records across both data sets by sorting $\mathbf{X}$ and $\mathbf{Y}$ by one attribute and evaluating how similar the values of the rest of attributes are in the sorted data sets.

To reflect the above procedure, we propose the confidentiality metric $CM3$ computed by Algorithm 1.

Algorithm 1:

1) For $j = 1$ to m do:
   a) Sort the original data set by its j-th attribute and let $\mathbf{X}^{-j}$ be the projection of the sorted data set on all attributes except the j-th one.
   b) Sort the anonymized data set by its j-th attribute and let $\mathbf{Y}^{-j}$ be the projection of the sorted data set on all attributes except the j-th one.
   c) Compute $CM2(\mathbf{X}^{-j}, \mathbf{Y}^{-j})$ according to Expression (5).
2) Let

$$CM3(\mathbf{X}, \mathbf{Y}) = \min_{1 \leq j \leq m} CM2(\mathbf{X}^{-j}, \mathbf{Y}^{-j}). \quad (6)$$

The $CM3$ confidentiality metric can be readily applied when $\mathbf{Y}$ is synthetic: a mapping between records in $\mathbf{X}$ and $\mathbf{Y}$ is not needed because one tries all m possible mappings obtained when using each single attribute as a sorting key.

The following are interesting cases of synthetic data sets:

- Let $\mathbf{X}$ be such that attributes X^i and X^j are perfectly correlated. Assume that the synthetic $\mathbf{Y}$ also preserves the relationship between Y^i and Y^j to be the same as the one between X^i and X^j. In other words, the permutations from X^i to Y^i and from X^j to Y^j are exactly the same. Hence, if we sort $\mathbf{X}$ by X^i and $\mathbf{Y}$ by Y^i, attributes X^j in $\mathbf{X}^{-i}$ and Y^j in $\mathbf{Y}^{-i}$ are perfectly correlated. Thus $CM2(\mathbf{X}^{-i}, \mathbf{Y}^{-i}) = 0$ and in consequence $CM3(\mathbf{X}, \mathbf{Y}) = 0$. However directly using $CM2$ on $\mathbf{X}$ and $\mathbf{Y}$ yields in general $CM2(\mathbf{X}, \mathbf{Y}) \neq 0$.
- If the attributes in $\mathbf{X}$ are very highly correlated, any masking method that preserves the correlation structure of $\mathbf{X}$ in $\mathbf{Y}$ cannot permute much. Consequently, it offers less confidentiality than if the correlation structure was not preserved. Equation (6) captures this situation of rank preservation among $\mathbf{X}$ and $\mathbf{Y}$ and gives a low value for $CM3$, even if $CM1$ and $CM2$ may be quite high, as illustrated in an experiment in Section 6.3.

4.4 Summary on confidentiality metrics

$CM1$ or $CM2$ should be applied whenever the mapping between original records and anonymized records is known. If the mapping is not known, such as in synthetic data, then $CM3$ should be applied.

The following holds regarding $CM1$ and $CM2$:

- $CM2$ is a product of terms not greater than 1 whose first term is $CM1$. Hence $CM2$ is not greater than $CM1$.
- $CM2 = 0$ if and only if $CM1 = 0$, because ρ_1 is the largest correlation.
- Since $CM2$ takes all canonical correlations into account, it is a better metric than $CM1$, although $CM1$ is easier to compute.

It is difficult to compare $CM3$ with $CM1$ or $CM2$. The former is intended for use with synthetic microdata in which there is no linkage between the records in the original and masked data. When applied to non-synthetic masking methods, since $CM3$ is based on an arbitrary linkage and $CM2$ is based on the true linkage, we normally have $CM3 > CM2$. The exception is the case of trivial permutation (swapping entire records), in which $CM3 = 0$. Indeed, under trivial permutation sorting $\mathbf{X}$ by any attribute X^i and $\mathbf{Y}$ by the corresponding Y^i yields identical sorted data sets, and thus $CM3 = 0$.

Therefore, $CM3$ has the advantage of detecting trivial permutation, which $CM1$ and $CM2$ do not detect.

$CM3$ can also be viewed as the confidentiality metric from the intruder's perspective. Unlike the data controller, the intruder does not know the true linkage between original and masked records and may evaluate confidentiality using $CM3$. Interestingly, when synthetic microdata are released neither the data controller nor the intruder know the "true" linkage and $CM3$ is a natural confidentiality metric for both.

5 A COVARIANCE-BASED BOUNDED UTILITY METRIC

In statistical disclosure control, a confidentiality metric needs to have a companion utility metric to allow for the necessary trade-off evaluation between utility and confidentiality. Since in Expressions (2), (5) and (6) we have proposed confidentiality metrics that are bounded

between 0 and 1, we need companion utility metrics that are also bounded.

After measuring confidentiality in terms of covariance matrices, it is natural to examine whether covariances can also conveniently characterize utility. Although some utility metrics focus on the mean error between original and anonymized data, preserving the covariance structure seems the most relevant utility feature for all those analyses aimed at discovering relationships between attributes.

If the attributes in the original and the anonymized data set are Gaussian (resp. near-Gaussian), then they are fully (resp. almost fully) described by their second-order statistics. Hence, in this case the covariance matrix is a sufficient measure of utility. The more the distribution of the attributes departs from Gaussian, the more likely higher-order relationships are that stay uncaptured by the covariance matrix, which nonetheless remains a meaningful utility measure.

Let a data set $\mathbf{X}$ be masked as $\mathbf{Y}$. As said above, we will consider the ranks of values in both data sets, rather than the values themselves. If all attributes are numerical and sparse, one might choose to work on values rather than ranks in order to capture utility more closely.

In terms of covariances, maximum utility occurs when $\mathbf{C}_{XX} = \mathbf{C}_{YY}$, in which case the (second-order) relationships between attributes in the original data set are exactly preserved in the masked data set. To compare how similar $\mathbf{C}_{XX}$ and $\mathbf{C}_{YY}$ are, a rough procedure is to compare their respective eigenvalues. Let the eigenvalues of $\mathbf{C}_{XX}$ be λ_1^X, ..., λ_m^X, and the eigenvalues of $\mathbf{C}_{YY}$ be λ_1^Y, ..., λ_m^Y. In the case of a covariance matrix, the first eigenvalue represents the magnitude of the maximum spread of the data, the second eigenvalue is the magnitude of the second largest data spread in a direction orthogonal to the maximum spread direction, etc. Thus, eigenvalues appear in non-increasing order. The case of all m eigenvalues of a covariance matrix being equal would reflect a set of records having equal spread in all directions of the m-dimensional space, a sort of m-dimensional sphere; this occurs when all attributes are uncorrelated.

Unfortunately, just comparing eigenvalues is not sufficient to assess utility, because eigenvalues capture only the magnitude of the maximum spreads on orthogonal directions, but not the directions themselves. Thus, two data sets can share the same set of eigenvalues while being different: in particular, if $\mathbf{Y}$ is a rotation of $\mathbf{X}$, both data sets have the same eigenvalues.

For a given spread magnitude (eigenvalue), the direction of spread is described by the corresponding eigenvector. We loosely adapt a procedure proposed in [20] for comparing covariance matrices. Let λ_1^X, ..., λ_m^X, resp. λ_1^Y, ..., λ_m^Y be the eigenvalues of $\mathbf{C}_{XX}$, resp. $\mathbf{C}_{YY}$ in non-increasing order. Let $\mathbf{v}_1^X$, ..., $\mathbf{v}_m^X$, resp. $\mathbf{v}_1^Y$, ..., $\mathbf{v}_m^Y$ be the corresponding eigenvectors of $\mathbf{C}_{XX}$, resp. $\mathbf{C}_{YY}$. Then it holds that

$$\lambda_j^X = (\mathbf{v}_j^X)^T \mathbf{C}_{XX} \mathbf{v}_j^X, \;\; j = 1, \ldots, m.$$

Now consider

$$\lambda_j^{Y|X} = (\mathbf{v}_j^X)^T \mathbf{C}_{YY} \mathbf{v}_j^X, \;\; j = 1, \ldots, m.$$

Just as each eigenvalue λ_j^X can be viewed as the proportion of the variance of the attributes in $\mathbf{X}$ explained by the corresponding eigenvector $\mathbf{v}_j^X$, we can view $\lambda_j^{Y|X}$ as the proportion of the variance of the attributes in $\mathbf{Y}$ explained by $\mathbf{v}_j^X$. Note that the values $\lambda_j^{Y|X}$, for $j = 1, \ldots, m$, are not necessarily non-increasing.

The highest level of utility occurs when $\lambda_j^X = \lambda_j^{Y|X}$ for $j = 1, \ldots, m$, which occurs when $\mathbf{C}_{XX} = \mathbf{C}_{YY}$.

Covariance matrices are positive semi-definite, which means that their eigenvalues are all non-negative. If $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of a covariance matrix, let $\hat{\lambda}_1, \ldots, \hat{\lambda}_m$ be their scaled versions so that they add to 1. Then the extent to which $\mathbf{C}_{XX}$ and $\mathbf{C}_{YY}$ differ can be expressed as

$$\sum_{j=1}^{m} (\hat{\lambda}_j^X - \hat{\lambda}_j^{Y|X})^2. \tag{7}$$

Proposition 1: Expression (7) is bounded between 0 and 2. The maximum occurs when all the variance of $\mathbf{X}$ occurs in a single direction and all the variance of $\mathbf{Y}$ also occurs in a single direction that is orthogonal to the previous one.

Proof: The minimum is clearly 0. To compute the maximum, take into account that $\sum_{j=1}^m \hat{\lambda}_j^X = \sum_{j=1}^m \hat{\lambda}_j^{Y|X} = 1$. Thus, the maximum occurs when two of the squares added in Expression (7) are 1, and a square can be 1 if it is either $(1-0)^2$ or $(0-1)^2$. This situation occurs when: i) all the variance of $\mathbf{X}$ is explained by the first eigenvector $\mathbf{v}_j^X$, in which case we have $\hat{\lambda}_1^X = 1$ and $\hat{\lambda}_l^X = 0$ for all $l = 2, \ldots, m$; and ii) all the variance of $\mathbf{Y}$ is explained by one eigenvector $\mathbf{v}_{j'}^X$ with $j' \neq 1$ and hence orthogonal to $\mathbf{v}_j^X$, in which case $\hat{\lambda}_{j'}^{Y|X} = 1$ and the rest of $\hat{\lambda}_j^{Y|X}$ are zero. □

According to Proposition 1, the maximum difference between two covariance matrices $\mathbf{C}_{XX}$ and $\mathbf{C}_{YY}$ can be quantified as 2. However, value 2 is reached when $\mathbf{C}_{YY}$ has a very specific shape with respect to $\mathbf{C}_{XX}$. Rather, we are interested in finding a measure of utility, that is, to see how much the covariances in $\mathbf{X}$ are preserved in $\mathbf{Y}$. In this sense, the intuition is that the maximum utility loss occurs when the covariances of $\mathbf{X}$ are completely lost in $\mathbf{Y}$, or equivalently when any of the m eigenvectors of $\mathbf{C}_{XX}$ explains a fraction $1/m$ of the variance of $\mathbf{Y}$. In this case, $\hat{\lambda}_j^{Y|X} = 1/m$ for $j = 1, \ldots, m$, and Expression (7) becomes:

$$\sum_{j=1}^{m} (\hat{\lambda}_j^X - 1/m)^2. \tag{8}$$

Proposition 2: Expression (8) is bounded between 0 and $(m-1)/m$.

Proof: The minimum is clearly 0. The maximum is reached when all the variance of $\mathbf{X}$ occurs in a single direction. In that situation:

$$\sum_{j=1}^{m}(\hat{\lambda}_j^X - 1/m)^2 = (1-1/m)^2 + (0-1/m)^2 + \ldots + (0-1/m)^2$$

$$= (m-1)/m.$$

□

We are now in a position to define the following utility measure based on Expressions (7) and (8):

$$UM(\mathbf{X},\mathbf{Y}) = \begin{cases} 1 & \text{if } \hat{\lambda}_j^X = \hat{\lambda}_j^{Y|X} = 1/m \text{ for } j = 1,\ldots,m; \\ 1 - \min\left(1, \frac{\sum_{j=1}^{m}(\hat{\lambda}_j^X - \hat{\lambda}_j^{Y|X})^2}{\sum_{j=1}^{m}(\hat{\lambda}_j^X - 1/m)^2}\right) & \text{otherwise.} \end{cases} \quad (9)$$

The first case in Expression (9) covers the (very exceptional) situation in which both the original data set and the anonymized data set are perfectly uncorrelated, which means there is no utility loss. Regarding the second case, from Propositions 1 and 2, the ratio within the argument of the minimum function can be greater than 1. By using the minimum, we make sure $UM(\mathbf{X},\mathbf{Y})$ is bounded between 0 and 1. Thus we have:

- Top utility $UM(\mathbf{X},\mathbf{Y}) = 1$ is reached when information loss is zero, which occurs when $\hat{\lambda}_j^X = \hat{\lambda}_j^{Y|X}$ for $j = 1,\ldots,m$.
- Zero utility $UM(\mathbf{X},\mathbf{Y}) = 0$ occurs if $\hat{\lambda}_j^X$ and $\hat{\lambda}_j^{Y|X}$ differ at least as much as $\hat{\lambda}_j^X$ and the eigenvalues of an uncorrelated data set.

6 Empirical work

The purpose of the experiments reported in this section is to highlight that the confidentiality and utility metrics presented above can be applied to a variety of privacy-first and utility-first anonymization approaches. Specifically, we consider privacy models (k-anonymity [35] and differential privacy [16]) and SDC methods (additive noise, multiplicative noise and synthetic data).

6.1 Privacy models

To test k-anonymity and ϵ-differential privacy, we took as original data set the "Census" data set, which contains 1,080 records with numerical attributes [1]. This data set was used in the European project CASC and in [7], [5], [40], [29], [12], [10], [6], [36]. Like in [6], [36], we took attributes FICA (Social security retirement payroll deduction), FEDTAX (Federal income tax liability), INTVAL (Amount of interest income) and POTHVAL (Total other persons income). We considered all four attributes as quasi-identifiers in all of our tests. The resulting records were all different from each other. Since all attributes represent non-negative amounts of money, we took as boundaries for the domain of each attribute 0 and 1.5 times the maximum value of the attribute in the data set.

We then took three versions of the "Census" data set: one with all 4 attributes, one with 3 attributes (FICA, FEDTAX and INTVAL) and one with 2 attributes (FICA and FEDTAX). We separately anonymized the three versions as follows:

- Achieving k-anonymity for $k = 2, 3, \ldots, 100$ and $k = 200, 300, 400, 500$ using the MDAV microaggregation algorithm [12].
- Achieving ϵ-differential privacy via Laplace noise addition to unagreggated attribute data for ϵ= 0.01, 0.1, 1, 10, 25, 50, 100, which covers the usual range of differential privacy levels observed in the literature [17], [3], [4], [30] plus some very large ϵ values. For each ϵ value, five differentially private data sets were generated and utility and confidentiality metrics were averaged over the five data sets.

Figure 2 shows the utility metric and the confidentiality metrics $CM1$ and $CM2$ for the k-anonymized data as a function of k and the number of attributes. As expected, as k increases, utility decreases and confidentiality increases. Also, as anticipated in Section 2.1, for fixed k a decrease of utility and an increase of confidentiality is observed when the number of attributes grows.

Figure 3 shows the utility metric and the confidentiality metrics $CM1$ and $CM2$ for the ϵ-differentially private data as a function of ϵ and the number of attributes. As expected, as ϵ increases, utility increases and confidentiality decreases. Also, consistently with Section 2.1, for fixed ϵ a decrease of utility and an increase of confidentiality is observed when the number of attributes grows[1].

Note that, since k-anonymity is achieved via microaggregation and ϵ-differential privacy via noise addition, for both privacy models the controller knows the correspondence between each anonymized record and the original record it derives from. Therefore, it does not make sense for the controller to use the mapping-free confidentiality metric $CM3$.

By superposing Figures 2 and 3 (or rather the numbers behind them), one can compare the utility-confidentiality trade-offs achieved by k-anonymity and ϵ-differential privacy. For most parameters we tried on the three

1. The particular data set that we used in this study consists of skewed economic data. The range and hence the global sensitivity of attributes in the data set are large. As a result, the variance of the noise added is also relatively large even when $\epsilon = 100$. In addition, with multiple attributes, the ϵ budget must be split among the attributes, effectively reducing the budget for each attribute. For four attributes, the variance of the Laplace noise added is four times the variance of the Laplace noise added for two attributes. Consequently, the correlation between the original and masked data in the four-attribute case is substantially lower than in the two-attribute case. Furthermore, canonical correlation evaluates correlation among *all* original and masked attributes simultaneously. Not only is the variance added in the case of four attributes four times larger than in the case of two attributes, but in the former case four masked attributes are compared against four original attributes, which also contributes to a greater utility loss than comparing two masked attributes against two original attributes as in the latter case. If utility is an important consideration, the data controller may wish to consider an alternative mechanism for implementing differential privacy when there are many attributes.

versions of the "Census" data set, k-anonymity yields substantially more utility (above 0.9 for all k) but substantially less confidentiality than differential privacy.

With privacy models, the data controller is blind regarding any issue other than privacy. In addition, it is also very difficult to compare across privacy models. One of the objectives of this paper is to propose measures that inform the controller on other aspects of the masking procedure, namely utility and confidentiality. This allows the controller to compare the trade-offs offered by different privacy models and to evaluate whether a lower/higher level of privacy may be warranted for the data set. The decision regarding the right levels of anonymity, confidentiality and utility must be made by the data controller. Our measures give him information to make that decision.

6.2 Noise-based SDC methods

We further tried our metrics on two typical ways of utilizing noise for statistical disclosure control: noise addition and noise multiplication [25]. The results are shown in Figures 4 and 5. In the former, an anonymized attribute Y is obtained as $Y = X + E_X$, where X is the corresponding original attribute and E_X is a noise random variable distributed as $N(0, \alpha\sigma_X)$, with $0 < \alpha \leq 1$ and σ_X the standard deviation of X. In multiplicative noise, Y is obtained as $Y = XE_X$, where X is the corresponding original attribute and E_X is a noise random variable generated from $Uniform(1-\beta, 1+\beta)$, with $0 \leq \beta < 1$. As it could be expected, it can be seen that, as the noise standard deviation increases, the utility metric decreases and the confidentiality metrics increase. For additive noise, these effects are more pronounced: the reason is that for multiplicative noise the changes in the ranks are smaller than for additive noise. Like above, we did not use $CM3$ because in noise addition and multiplication the controller knows the mapping between original and anonymized records.

6.3 Synthetic data

Finally, we have tried $CM3$ on synthetic data generated using the IPSO method [2]. IPSO generates a synthetic data set with *exactly* the same means and covariances as the original data set.

The first row of Table 1 shows the confidentiality metrics $CM1$, $CM2$ and $CM3$ as well as the utility metric UM achieved by IPSO when run on "Census" as the original data set (average of 100 replications). The results tell that the synthetic data have high utility, which was to be expected because IPSO preserves covariances. Regarding confidentiality, both $CM1$ and $CM2$ give very high values, but $CM3$ is substantially lower because it explores all possible mappings. This confirms what we said above: for synthetic data, $CM1$ and $CM2$ should not be used, as they need a mapping and no true mapping is known.

TABLE 1
Confidentiality and utility metrics for synthetic data generated with the IPSO method on two original data sets: "Census" (first row) and a simulated data set with highly correlated attributes (second row)

Data set	UM	$CM1$	$CM2$	$CM3$
Census	0.9638	0.9904	0.9849	0.6673
Simulated	1.0000	0.9914	0.9913	0.0277

The second row of Table 1 shows the metrics when IPSO is run on a simulated data set also with 1,080 records and four attributes, but with very high correlation (0.99) between the attributes. Like for "Census", the results are the average of 100 replications of IPSO. The synthetic data provide very high utility UM, as expected, but the confidentiality according to $CM3$ is very low. Note that $CM1$ and $CM2$ are high because they are misled by the lack of a natural mapping in synthetic data. Thus, $CM3$ is the only confidentiality metric that detects how high is the risk of disclosure in IPSO (or in any other synthetic data generation method) in the case of highly correlated original attributes.

6.4 Summary of experimental results

The confidentiality and utility measures proposed in this paper are influenced by the characteristics of the data set. Hence, the results of the above experiments are not generalizable to data sets, SDC methods, or parameter choices other than those considered. However, the above empirical work does illustrate two general facts:

- The application of the proposed confidentiality and utility metrics to substantially different privacy models and SDC methods shows that the metrics are consistent. If the confidentiality parameters of the models and/or methods are set for higher confidentiality (higher k for k-anonymity, lower ϵ for differential privacy, higher noise standard deviation for noise methods), our metrics detect more confidentiality and less utility. And conversely if parameters are set for lower confidentiality.
- Our metrics have been shown helpful to compare not only how different parameter values affect the confidentiality-utility trade-off of a certain privacy model or SDC method, but even more interestingly, to compare the trade-offs across different privacy models and/or SDC methods.

7 Related work

Given that the specific analyses data users will perform on anonymized data are seldom known by the data protector at the time of anonymization, there has been a sustained interest in the literature on generic utility metrics. On the other hand, there has also been substantial activity to design confidentiality metrics that

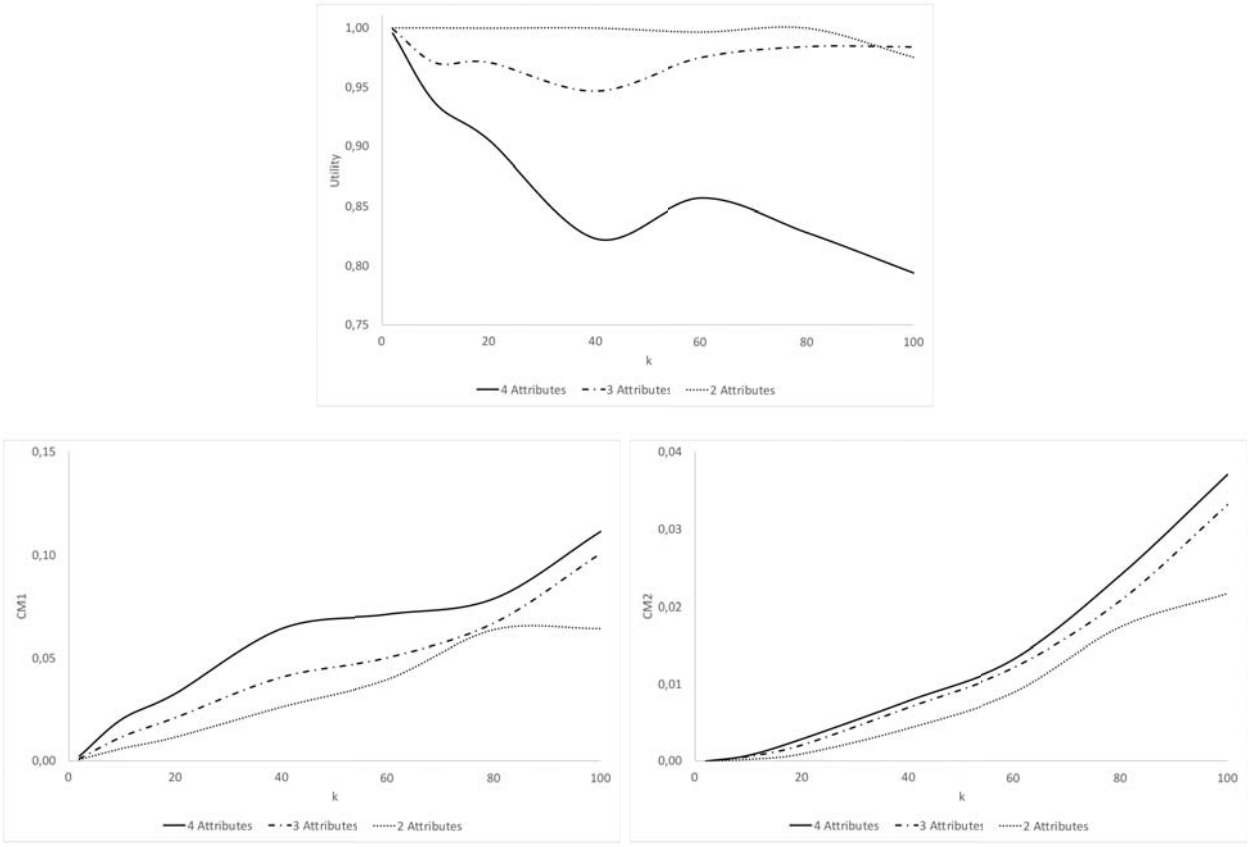


Fig. 2. Utility metric (top) and confidentiality metrics $CM1$ (bottom left) and $CM2$ (bottom right) for k-anonymized data as a function of k and the number of attributes

could circumvent the costly empirical approach based on record linkage.

In [11], a score was proposed that combines utility loss and disclosure risk (confidentiality loss) metrics. The approach to disclosure risk assessment in that paper relies on record linkage experiments. On the other hand, utility loss is measured by comparing records and some statistics in the original data set and the anonymized data set. Specifically, the mean square error, the mean absolute error and the mean variation are used as comparison criteria. The resulting utility loss measures are unbounded and thus hard to compare with disclosure risk.

In [31], a bounded utility loss metric based on probabilities is presented. The metric is the probability that the absolute value of the discrepancy between a sample statistic $\hat{\Theta}$ and the corresponding population parameter θ is less than or equal to the discrepancy $|\hat{\theta}-\theta|$ measured in the anonymized data set. The intuition is that, the more different from θ is the value $\hat{\theta}$ of the sample statistic in the anonymized data set, the more utility is lost when publishing the anonymized data set. Being bounded, this metric can be readily compared with the risk of disclosure, that cannot be above 100%. However, it has the drawbacks of being only applicable to continuous microdata and not benefiting from the generality offered by the permutation model.

In [39], another generic utility loss metric is proposed that relies on propensity scores. The original microdata and the anonymized microdata are merged and a binary attribute T is added that takes value 0 for the original records and value 1 for the anonymized records. Then T is regressed on the rest of attributes. Let $\hat{T}$ be the adjusted attribute and let the propensity score $\hat{p}_i$ of record i of the merged data be the value of $\hat{T}$ for record i. Then utility is high if the propensity scores of the anonymized and the original records are similar. This metric is attractive because it focuses on the actual microdata rather than on preselected statistics. However, it has the drawbacks of being unbounded and being dependent on the specific regression model chosen.

In [34], power means were used to obtain confidentiality and utility metrics based on the permutation model. The idea is to aggregate the absolute permutation distances $p_1, \ldots, p_n$ resulting from anonymizing the values of an attribute in the n records of a data set:

$$J((p_1,\ldots,p_n),\alpha) = \begin{cases} \left(\frac{1}{n}\sum_{i=1}^{n} p_i^{\alpha}\right)^{\frac{1}{\alpha}} & \text{for } \alpha \neq 0; \\ \Pi_{i=1}^{n} p_i^{\frac{1}{n}} & \text{for } \alpha = 0, \end{cases} \quad (10)$$

where $\alpha < 1$ turns the above expression into a disclo-

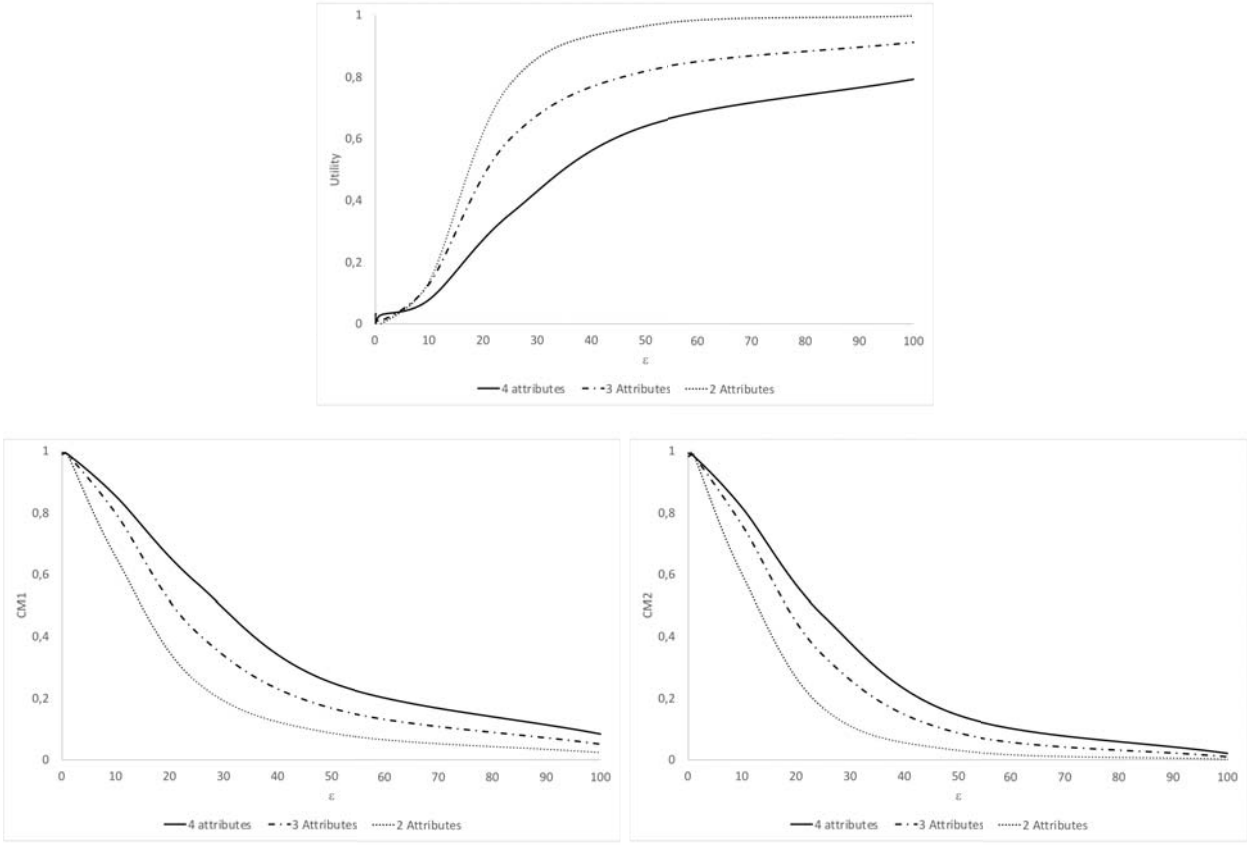


Fig. 3. Utility metric (top) and confidentiality metrics $CM1$ (bottom left) and $CM2$ (bottom right) for ϵ-differentially private data as a function of ϵ and the number of attributes

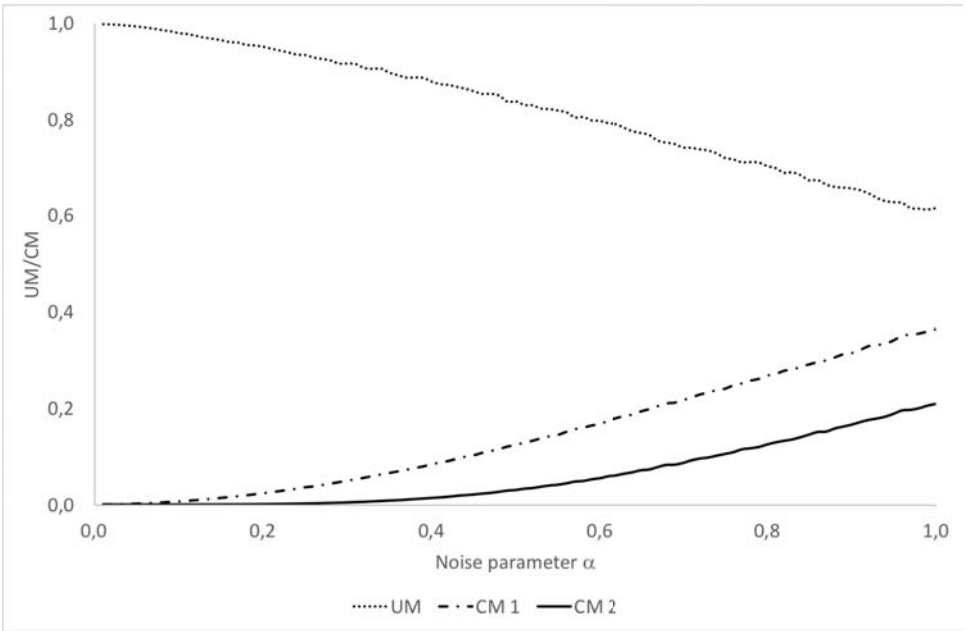


Fig. 4. Confidentiality and utility metrics for additive noise as a function of the noise parameter α (the larger α, the larger the noise standard deviation)

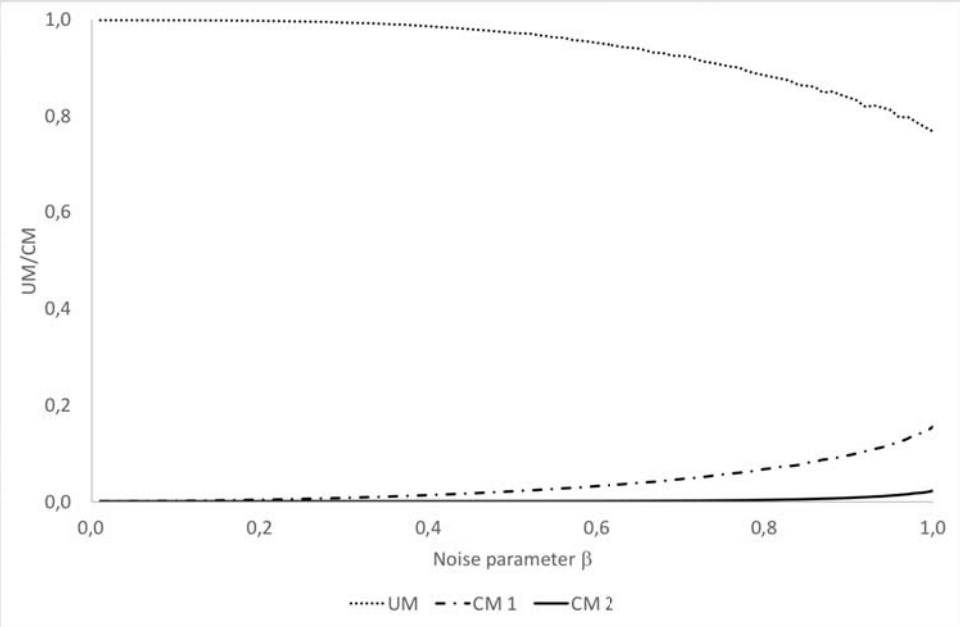


Fig. 5. Confidentiality and utility metrics for multiplicative noise as a function of the noise parameter β (the larger β, the larger the noise standard deviation)

sure risk metric and $\alpha > 1$ into a utility loss metric. Indeed, the more α approaches $-\infty$, the greater is the weight of smaller permutation distances in Expression (10); since disclosure occurs when permutation distances for some values are too small, we have a disclosure risk metric when α is small. On the other hand, the more α approaches $+\infty$, the greater is the weight of larger permutation distances in Expression (10); since large permutation distances are the ones that most deteriorate utility, we have a utility loss metric when α is large. Thus, for $\alpha < 1$, the greater the value of $J((p_1, \ldots, p_n), \alpha)$, the more disclosure risk, whereas, for $\alpha > 1$, the greater the value of $J((p_1, \ldots, p_n), \alpha)$, the more utility loss.

These power-means metrics may be used to compare the disclosure protection and the information loss achieved by two different anonymization methods M and M' (or by the same method M with different parameters $parms$ and $parms'$). However, they have the shortcomings of being intrinsically univariate (they operate independently for each attribute) and unbounded. In contrast, in this paper, we have proposed bounded metrics that take all attributes of the data set into account.

8 Conclusions and Future Work

The permutation model is useful to capture the underlying nature of microdata anonymization, which turns out to be essentially permutation (altering ranks) plus some residual noise (to alter values and make them different from the original ones). It seems natural to leverage this general model to derive general metrics for utility loss and disclosure risk. This is what we have done in this paper, with the additional feature of providing bounded metrics that allow easily evaluating the trade-off between utility loss and disclosure risk for any anonymization method.

We have presented experimental work that shows that our metrics provide results that are consistent with the intuition for many anonymization approaches in the literature, including privacy models as well as SDC methods based on noise and synthetic data. In particular, we have been able to compare the utility-confidentiality trade-offs achieved by these widely heterogeneous methods, which would not be possible without the confidentiality and utility methods developed in this study.

Future research lines may include comparing the results of our metrics with those obtained with the alternative metrics in the literature mentioned in Section 7. Also, it may be interesting to compare the confidentiality metrics with the risk estimated via record linkage and the utility loss metric with the utility for specific data uses.

Acknowledgments

Thanks go to Sergio Martínez for generating the k-anonymous and the differentially private data sets used in the empirical section. Partial support for this work has been received from the Government of Catalonia (ICREA Acadèmia Award to J. Domingo-Ferrer and grant 2017 SGR 705), the European Commission (project H2020-871042 "SoBigData++") and the Spanish Government (project RTI2018-095094-B-C21 "CONSENT"). J. Domingo-Ferrer holds the UNESCO Chair in Data

Privacy, but the opinions in this paper do not commit UNESCO.

REFERENCES

[1] R. Brand, J. Domingo-Ferrer and J. M. Mateo-Sanz. Reference data sets to test and compare SDC methods for protection of numerical microdata. European Project IST-2000-25069 CASC, 2002. http://neon.vb.cbs.nl/casc

[2] J. Burridge. Information preserving statistical obfuscation. *Statistics and Computing*, 13(4):321-327, 2003.

[3] A.-S. Charest. How can we analyze differentially-private synthetic data sets? *Journal of Privacy and Confidentiality*, 2(2):21-33, 2010.

[4] A.-S. Charest. Empirical evaluation of statistical inference from differentially-private contingency tables. In *Privacy in Statistical Databases-PSD 2012*, pp. 257-272. Springer, 2012.

[5] R. Dandekar, J. Domingo-Ferrer and F. Sebé. LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In *Inference Control in Statistical Databases*, pp. 153-162. Springer, 2002.

[6] J. Domingo-Ferrer and Ú. González-Nicolás. Hybrid microdata using microaggregation. *Information Sciences*, 180(15):2834-2844, 2010.

[7] J. Domingo-Ferrer, J. M. Mateo-Sanz and V. Torra. Comparing SDC methods for microdata on the basis of information loss and disclosure risk. In *Pre-proceedings of ETK-NTTS'2001* (vol. 2), pp. 807-826. Eurostat, 2001.

[8] J. Domingo-Ferrer and K. Muralidhar. New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *Information Sciences*, 337-338:11-24, 2016.

[9] J. Domingo-Ferrer, D. Sánchez and G. Rufian-Torrell. Anonymization of nominal data based on semantic marginality. *Information Sciences*, 242:35-48, 2013.

[10] J. Domingo-Ferrer, F. Sebé and A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Computers & Mathematics with Applications*, 55(4):714-732, 2008.

[11] J. Domingo-Ferrer and V. Torra. A quantitative comparison of disclosure control methods for microdata. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pp. 111-134. North-Holland, 2001.

[12] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195-212, 2005.

[13] J. Drechsler. *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Springer, 2011.

[14] C. Duhigg. How companies learn your secrets. *New York Times Magazine*, Feb. 16, 2012.

[15] G. T. Duncan and R. W. Pearson. Enhancing access to microdata while protecting confidentiality: prospects for the future. *Statistical Science*, 6(3):219-232, 1991.

[16] C. Dwork. Differential privacy. In *ICALP'06*, pp. 1-12. Springer, 2006.

[17] C. Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86-95, 2011.

[18] E.A.H. Elamir and C.J. Skinner. Record level measures of disclosure risk for survey microdata. *Journal of Official Statistics*, 22(3):525-539, 2006.

[19] R. L. Fowler. Power and robustness in product-moment correlation. *Applied Psychological Measurement*, 11(4):419-428, 1987.

[20] C. Garcia. A simple procedure for the comparison of covariance matrices. *BMC Evolutionary Biology*, 12:222, 2012.

[21] *General Data Protection Regulation*. Regulation (EU) 2016/679. https://gdpr-info.eu

[22] J. Hauke and T. Kossowsk. Comparison of values of Pearson's and Spearman's correlations coefficients on the same sets of data. *Quaestiones Geographicae*, 30(2):87-93, 2011.

[23] W. Härdle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer, 2007.

[24] H. Hotelling. Relations between two sets of variables. *Biometrika*, 28(3-4):321-377, 1936.

[25] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer and P.-P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.

[26] J. W. Kay. *A Note on Mutual Information and Correlation*. Research report, 1991.

[27] J. W. Kay. Feature discovery under contextual supervision using mutual information. In *International Joint Conference on Neural Networks-IJCNN 1992*, vol. 4, pp. 79-84. IEEE, 1992.

[28] D. Lambert. Measures of disclosure risk and harm. *Journal of Official Statistics*, 9(3):313-331, 1993.

[29] M. Laszlo and S. Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):902-911, 2005.

[30] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke and L. Vilhuber. Privacy: theory meets practice on the map. In *IEEE International Conference on Data Engineering-ICDE 2008*, pp. 277-286, 2008.

[31] 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):181-193, 2005.

[32] K. Muralidhar and R. Sarathy. Data shuffling — A new masking approach for numerical data. *Management Science*, 52(5):658-670, 2006.

[33] J. P. Reiter. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *Journal of the Royal Statistical Society: Series A*, 168(1):185-205.

[34] N. Ruiz. On some consequences of the permutation paradigm for data anonymization: centrality of permutation matrices, universal measures of disclosure risk and information loss, evaluation by dominance. *Information Sciences*, 430-431:620-633, 2018.

[35] P. Samarati and L. Sweeney. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. Tech. rep., SRI International, 1998.

[36] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez and S. Martínez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. *VLDB Journal*, 23(5): 771-794, 2014.

[37] V. Torra and J. Domingo-Ferrer. Record linkage methods for multidatabase data mining. In *Information Fusion in Data Mining*, pp. 99-130. Springer, 2003.

[38] J. C. F. de Winter, S. D. Gosling and J. Potter. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: a tutorial using simulations and empirical data. *Psychological Methods*, 21(3):273-290, 2016.

[39] M.-J. Woo, J. P. Reiter, A. Oganian and A. F. Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1), 2009.

[40] W. E. Yancey, W. E. Winkler and R. H. Creecy. Disclosure risk assessment in perturbative microdata protection. In *Inference Control in Statistical Databases*, pp. 135-152. Springer, 2002.



Josep Domingo-Ferrer (Fellow, IEEE) is a Distinguished Professor of Computer Science and an ICREA-Acadèmia Researcher at Universitat Rovira i Virgili, Tarragona, Catalonia, where he holds the UNESCO Chair in Data Privacy and leads CYBERCAT. He received the MSc and PhD degrees in Computer Science from the Autonomous University of Barcelona in 1988 and 1991, respectively. He also holds an MSc degree in Mathematics. His research interests are in data privacy, data security and cryptographic protocols. More information on him can be found at http://crises-deim.urv.cat/jdomingo



Krishnamurty Muralidhar is the Baldwin Chair in Business and the Research Director for the Center for Business of Healthcare at the University of Oklahoma. He received a BSc from the University of Madras, India, an MBA from Sam Houston State University and a PhD from Texas A&M University. His main research interests are in data privacy.



Maria Bras-Amorós is an Associate Professor of Applied Mathematics at Universitat Rovira i Virgili. She received her MSc (1998) and PhD (2003) in Mathematics from the Technical University of Catalonia (UPC). Her main research interests are in data privacy, coding theory and combinatorics.