A Methodology to Compare Anonymization Methods Regarding Their Risk-Utility Trade-Off

Josep Domingo-Ferrer, Sara Ricci, and Jordi Soria-Comas

UNESCO Chair in Data privacy
Department of Computer Science and Mathematics
Universitat Rovira i Virgili
Av. Països Catalans 26
43007 Tarragona, Catalonia
{josep.domingo,sara.ricci,jordi.soria}@urv.cat

Abstract. We present here a methodology to compare statistical disclosure control methods for microdata in terms of how they perform regarding the risk-utility trade-off. Previous comparative studies (e.g. [3]) usually start by selecting some parameter values for a set of SDC methods and evaluate the disclosure risk and the information loss yielded by the methods for those parameterizations. In contrast, here we start by setting a certain risk level (resp. utility preservation level) and then we find which parameter values are needed to attain that risk (resp. utility) under different SDC methods; finally, once we have achieved an equivalent risk (resp. utility) level across methods, we evaluate the utility (resp. the risk) provided by each method, in order to rank methods according to their utility preservation (resp. disclosure protection), given a certain level of risk (resp. utility) and a certain original data set. The novelty of this comparison is not limited to the above-described methodology: we also justify and use general utility and risk measures that differ from those used in previous comparisons. Furthermore, we present experimental results of our methodology when used to compare the utility preservation of several methods given an equivalent level of risk for all of them.

Keywords: Record linkage, disclosure risk, utility preservation, privacy, permutation paradigm.

1 Introduction

With the expansion of information technology, the importance of data analysis (e.g. to support decision making processes) has increased significantly. Although data collection has become easier and more affordable than ever before, releasing data for secondary use (that is, for a purpose other than the one that triggered the data collection) remains very important: in most cases, researchers cannot afford collecting themselves the data they need. However, when the data released for secondary use refer to individuals, households or companies, the privacy of the data subjects must be taken into account.

Statistical disclosure control (SDC) methods aim at releasing data that preserve their statistical validity while protecting the privacy of each data subject. Among the possible types of data releases, this work focuses on microdata (that is, on the release of data about individual subjects).

While there is a great diversity of SDC methods for microdata protection, all of them imply some level of data masking. The greater the amount of masking, the greater are both privacy protection and information loss. Different SDC methods tackle the trade-off between privacy and utility in different ways. For example, in global recoding the level of information loss is set beforehand (the amount of coarsening of the categories of each attribute), whereas the disclosure risk is evaluated afterwards on the protected data set. In contrast, in k-anonymity [9] the risk of disclosure (the risk of record re-identification, in particular) is set beforehand, whereas the actual information loss results from the masking needed to attain the desired level of disclosure risk.

Although some general assertions about specific SDC methods/models can be made, comparing the latter regarding the privacy-utility trade-off is not straightforward. Let us illustrate this point with two well-known privacy models: differential privacy [5] and k-anonymity [9]. In terms of privacy protection, ϵ-differential privacy is regarded as stronger than k-anonymity. On the contrary, k-anonymity is regarded as more utility-preserving than ϵ-differential privacy. The practical value of these general statements is dubious. After all, by increasing ϵ we reduce the protection of differential privacy, and by increasing k we reduce the utility of k-anonymous data. An accurate comparison between SDC methods has to take into consideration both aspects of the privacy-utility trade-off.

Contribution and plan of this paper

Many risk and utility measures have been proposed in the literature, but some of them are designed for use with specific SDC methods. For example, the probability of record re-identification is the natural risk measure in k-anonymity, but it may not be appropriate in SDC methods that are not predicated on protecting privacy by hiding each data subject within a crowd. In this work, we propose a framework based on general empirical measures of utility and risk to compare the risk-utility trade-off of several SDC methods.

Previous comparative studies (e.g. [3]) usually start by selecting some parameter values for a set of SDC methods and evaluate the disclosure risk and the information loss yielded by the methods for those parameterizations. In contrast, here we start by setting a certain risk level (or a certain utility level) and then we find which parameter values are needed to attain that risk (resp. that utility) under different SDC methods; finally, once we have achieved an equivalent risk level (resp. utility level) across methods, we evaluate the utility (resp. the risk) provided by each method, in order to rank methods according to their utility preservation (resp. disclosure protection), given a certain level of risk (resp. utility) and a certain original data set. Furthermore, we present experimental work that illustrates the application of the proposed methodology.

The rest of the paper is organized as follows. In Section 2, we introduce some background relevant to the remaining sections. In Section 3, we describe the proposed framework for comparing methods regarding their risk-utility trade-off. In Section 4, we propose an empirical measure of disclosure risk that is based on record linkage. Experimental results are reported in Section 5. Conclusions are gathered in Section 6.

2 Background

2.1 Permutation paradigm and permutation distance

In [2], a permutation paradigm to model anonymization was proposed. Let $X = \{x_1, \ldots, x_n\}$ be the values taken by attribute X in the original data set. Let $Y = \{y_1, \ldots, y_n\}$ represent the anonymized version of X. Consider the attribute Z obtained using the following reverse-mapping procedure

```
For $i = 1$ to $n$
    Compute $j = rank(y_i)$
    Set $z_i = x_{(j)}$ (where $x_{(j)}$ is the value of $X$ of rank $j$)
Endfor
```

We can now view the anonymization of X into Y as a permutation step to turn X into Z, plus a small noise addition to turn Z into Y. Note the noise addition must be necessarily small, because it cannot alter ranks: by construction the ranks of Y and Z are the same. If we perform the above procedure independently for all attributes of an original data set $\mathbf{X}$ and corresponding attributes of an anonymized data set $\mathbf{Y}$, we can say that anonymization can be decomposed into a permutation step to obtain a data set $\mathbf{Z}$ plus a (small) noise addition to obtain $\mathbf{Y}$ from $\mathbf{Z}$.

The permutation distance measures the dissimilarity between two records in terms of the ranks of the values of their attributes. Assume the original data set $\mathbf{X}$ consists of m attributes $X^1, \ldots, X^m$ and the anonymized data set consists of corresponding attributes $Y^1, \ldots, Y^m$. Let $\mathbf{x} = (x^1, \ldots, x^m)$ be a record in $\mathbf{X}$ and $\mathbf{y} = (y^1, \ldots, y^m)$ be a record in $\mathbf{Y}$. The permutation distance between $\mathbf{x}$ and $\mathbf{y}$ is the maximum of the rank distances of the attributes:

$$d(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq m} |rank(x^i) - rank(y^i)|.$$

The permutation distance between records is used in [2] to conduct a record linkage between the original data set $\mathbf{X}$ and the anonymized data set $\mathbf{Y}$. In particular, records with minimal permutation distance are linked.

2.2 Utility measures

Utility measures are a key component to compare SDC methods. We introduce two utility measures that will be used in the empirical evaluation of the proposed methodology: the propensity scores [12] and the earth mover's distance [8].

Algorithm 1 shows a way to use the propensity scores as a utility measure.

Algorithm 1

1. *Merge the original data set* $\mathbf{X}$ *and the anonymized data set* $\mathbf{Y}$ *and add a binary attribute* T *with value 1 for the anonymized records and 0 for the original records.*
2. *Regress* T *on the rest of attributes of the merged data set and call the adjusted attribute* $\hat{T}$*. Let the* propensity score $\hat{p}_i$ *of record* i *of the merged data set be the value of* $\hat{T}$ *for record* i*.*
3. *The utility can be considered high if the propensity scores of the anonymized and original records are similar. Hence, if the original and the anonymized data sets have the same number* n *of records, the following is a utility measure*

$$\mathcal{U}_{ps}(\mathbf{X},\mathbf{Y}) = \frac{1}{2n}\sum_{i=1}^{2n}[\hat{p}_i - \frac{1}{2}]^2. \tag{1}$$

The value $\mathcal{U}_{ps}$ resulting from Equation (1) is close to zero if the propensity scores computed with the regression model for all records are similar (in which case they will be neither 0 nor 1, but close to 1/2). This situation means that the original and the anonymized records cannot be distinguished by the regression model, and hence the utility of the anonymized data set is high (its records "look" like the original records). In contrast, if the adjusted propensity scores were exactly the original values of T, it would mean that the regression model can exactly tell the original from the anonymized records, so the utility of the latter is low; in this case, we would have n propensity scores 0 and n propensity scores 1, which would yield a large $\mathcal{U}_{ps}$. Obviously, propensity scores as a utility measure are very dependent on the accuracy of the regression model adjusted to the data: the more accurate the model, the more discriminating it is and the less likely are values of $\mathcal{U}_{ps}$ indicating good utility (close to 0).

Earth mover's distance (EMD) is a natural extension of the notion of distance between single elements to distance between sets, or distributions, of elements. Given two distributions, one can be seen as a mass of earth in the space and the other as a collection of holes in that same space. Then, the EMD measures the least amount of work needed to fill the holes with earth, i.e. the minimal cost needed to transform one distribution into another by moving distribution mass. Thus, the EMD distance can be used to evaluate the similarity between the distribution of the original data set and the distribution of the anonymized data set. Note here that measuring similarity amounts to measuring utility, because, the more similar the distribution of the anonymized data to the distribution of the original data, the more useful are the anonymized data.

Formally, we can group records in clusters and represent each cluster j by its mean and the fraction ω_j of records that belong to that cluster. Let the original data set $\mathbf{X}$ be clustered as $\{(t_1, \omega_{t_1}), \ldots, (t_h, \omega_{t_h})\}$, and the anonymized data set $\mathbf{Y}$ as $\{(q_1, \omega_{q_1}), \ldots, (q_k, \omega_{q_k})\}$. Let $D = (d_{ij})$ be the matrix of the distance between the h clusters of $\mathbf{X}$ and the k clusters of $\mathbf{Y}$, i.e. $d_{ij} = t_i - q_j$ (in the multivariate case, we take the Euclidean distance between cluster means). The problem is to find a flow $F = (f_{ij})$, with f_{ij} being the flow between t_i and q_j, that minimizes the overall cost under some constraints (see [8] for more details). Once the optimal flow F is found, the earth mover's distance is defined as the resulting work normalized by the total flow:

$$\mathcal{U}_{emd}(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{h} \sum_{j=1}^{k} d_{ij} f_{ij}}{\sum_{i=1}^{h} \sum_{j=1}^{k} f_{ij}} \tag{2}$$

The greater $\mathcal{U}_{emd}$ is, the more different are the distributions of $\mathbf{X}$ and $\mathbf{Y}$ and hence the more utility has been lost in the anonymization process.

3 A methodology for comparing the risk-utility trade-off in SDC

In this section we describe a methodology for comparing SDC methods. Looking only at either the disclosure risk or the utility of an SDC method would be a flawed comparison. We need to analyze the privacy-utility trade-off, as explained in the introduction. Even if this principle may seem evident, very often it is not followed.

To make the proposed methodology as general as possible, we will employ empirical measures of risk and utility. That is, we will choose risk and utility measures that depend on the original and the anonymized data sets, rather than being prior conditions. To select specific measures, we need to define the aspects of risk and utility that we consider relevant for our comparison. In turn, the choice of measures will shape the outcome of the evaluation.

Let us illustrate the difference between empirical measures and prior conditions by taking differential privacy as an example. As a privacy model, differential privacy states some privacy guarantees but does not tell how they ought to be attained. Let us assume that A_1 and A_2 are ϵ-differentially private algorithms that output a data set. Let us also assume that A_2 is a refined version of A_1 that manages to attain ϵ-differential privacy while adding less noise than A_1. If we use the level ϵ of differential privacy as our risk measure, both A_1 and A_2 are equally good (they are both ϵ-differentially private). However, the fact that A_1 adds more noise to the original records may indicate that the data set output by A_1 entails less disclosure risk than the data set generated by A_2, even if differential privacy is unable to capture the difference. Alternative measures of disclosure risk (e.g. risk measures based on record linkage) should be able to capture the difference in risk between A_1 and A_2. In this work, we do not deny the value of any measure of disclosure risk, but, due to their broader applicability, we will employ empirical risk measures based on record linkage.

Let us assume that we are given functions

$$\mathcal{U} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$$

$$\mathcal{R} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$$

such that, for any given original data set $\mathbf{X}$ and anonymized data set $\mathbf{Y}$,

- $\mathcal{U}(\mathbf{X}, \mathbf{Y})$ measures the utility of $\mathbf{Y}$ as a replacement for $\mathbf{X}$.
- $\mathcal{R}(\mathbf{X}, \mathbf{Y})$ measures the disclosure risk of $\mathbf{Y}$ as a replacement for $\mathbf{X}$.

We have described some utility measures in Section 2.2. In Section 4 we will describe several risk measures based on record linkage.

SDC methods usually accept some parameters that can be adjusted to select the desired level of disclosure risk/utility. Let $M_\alpha(\mathbf{X})$ be the anonymized data output by SDC method M with parameter α when applied to data set $\mathbf{X}$.

Given an original data set $\mathbf{X}$ and two anonymization algorithms M^1 and M^2, we say that M^1 is more utility-preserving than M^2 at risk level r if

$$\mathcal{U}(\mathbf{X}, M^1_\alpha(\mathbf{X})) \geq \mathcal{U}(\mathbf{X}, M^2_\beta(\mathbf{X})),$$

for α and β such that $\mathcal{R}(\mathbf{X}, M^1_\alpha(\mathbf{X})) = \mathcal{R}(\mathbf{X}, M^2_\beta(\mathbf{X})) = r$.

In a similar fashion, we can compare the risk associated to a given level of data utility. We say that M^1 is less disclosive than M^2 at utility level u if

$$\mathcal{R}(\mathbf{X}, M^1_\alpha(\mathbf{X})) \leq \mathcal{R}(\mathbf{X}, M^2_\beta(\mathbf{X})),$$

for α and β such that $\mathcal{U}(\mathbf{X}, M^1_\alpha(\mathbf{X})) = \mathcal{U}(\mathbf{X}, M^2_\beta(\mathbf{X})) = u$.

The results of the previous utility (resp. risk) comparison depend not only on the SDC method, but also on the original data set, the risk and the utility measures selected, and the target level of risk (resp. utility). Actually, this comparison methodology is designed for use by a data controller who must decide which among several SDC methods is best suited to anonymize a given data set with a given target level of disclosure risk or utility. In other words, the aim is not to make general statements about the relative goodness of several SDC methods. Although such statements may make sense in some cases, our results can only be taken as empirical clues of such underlying truths.

4 Empirical measures of disclosure risk

To compare the risk-utility trade-off between SDC methods, we need adequate measures of disclosure risk. For the methodology described in Section 3 to be broadly applicable, the risk measure should be as general as possible (rather than based on specific characteristics of an SDC method).

We propose a risk measure based on record linkage [11], which is a technique that seeks to match original records that correspond to the same individual. Among its several uses, record linkage has a direct application to disclosure risk assessment [10]. Such an application bears some resemblance to the way an intruder having access to the anonymized data and to some side knowledge would proceed. Let $\mathbf{E}$ be a data set that represents the non-anonymous side information available to the intruder. By linking records in $\mathbf{E}$ to records in $\mathbf{Y}$, the intruder associates identities to the records in $\mathbf{Y}$.

The number (or the proportion) of correct re-identifications is a common record linkage-based measure of disclosure risk. However, this measure has some limitations that we next discuss. It is certainly appropriate when SDC is achieved by masking the quasi-identifier attributes, whereas the sensitive attributes are left unmodified (or are only slightly modified). However, if the sensitive attributes have been significantly altered, a correct linkage may not be equivalent to disclosure. Furthermore, if we use SDC methods that are not based on masking the original records, we may not even be able to tell what a correct linkage is. Generating a synthetic data set by repeatedly sampling from a statistical model adjusted on the original data is an example of an SDC method not based on masking; and indeed, it is not possible to say what is the correct mapping between the original records and the synthetic records.

In the spirit of [2], rather than measuring the disclosure risk as the proportion of correct re-identifications, we will measure the risk of disclosure associated to a record in the original data set $\mathbf{X}$ by means of a distance to its linked record in $\mathbf{Y}$. Such an approach has two important advantages with respect to counting the number of correct re-identifications:

- It is more broadly applicable. The linkage between records in $\mathbf{X}$ and $\mathbf{Y}$ can be performed independently of the SDC methodology used, even when the correct mapping between original and anonymized records cannot be established.
- The distance between a record in $\mathbf{X}$ and its linked record in $\mathbf{Y}$ provides more detailed information about the risk associated to a record in $\mathbf{X}$ than a mere binary outcome (right/wrong linkage):
  - On the one hand, the binary nature of correct linkages could lead to understating the risk of disclosure when, in spite of failing to find the correct linkage, the intruder links to a record that is similar to the correct one.
  - On the other hand, if all the attributes in $\mathbf{Y}$ have been thoroughly altered by the SDC method, a correct linkage may not disclose any useful information to the intruder; in this case, the proportion of correct linkages would overstate the risk of disclosure.

Any record $\mathbf{x}$ in the original data set $\mathbf{X}$ is linked to the record $\mathbf{y}_\mathbf{x} \in \mathbf{Y}$ at the smallest distance, that is, such that

$$d(\mathbf{x}, \mathbf{y}_\mathbf{x}) = d(\mathbf{x}, \mathbf{Y}) = \min_{\mathbf{y} \in \mathbf{Y}} d(\mathbf{x}, \mathbf{y}).$$

The distance $d(\mathbf{x}, \mathbf{Y})$ is an indicator of the disclosure risk associated to $\mathbf{x}$. If the distance is small, there is a record in $\mathbf{Y}$ that is quite similar to $\mathbf{x}$ and the risk of disclosure is high.

The choice of the distance $d(\mathbf{x}, \mathbf{Y})$ is an important step in determining the disclosure risk. Along the lines of the permutation paradigm (see Section 2), our proposal is based on ranks, but it differs from [2] in the way attributes are aggregated. Let $\mathbf{x} = (x^1, \ldots, x^m)$ be a record from an original data set $\mathbf{X}$ with attributes $X^1, \ldots, X^m$ and $\mathbf{y} = (y^1, \ldots, y^m)$ be a record from an anonymized data set $\mathbf{Y}$ with attributes $Y^1, \ldots, Y^m$. Take the distance between $\mathbf{x}$ and $\mathbf{y}$ to be the Euclidean distance between ranks, that is,

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{m} [rank_{X^i}(x^i) - rank_{Y^i}(y^i)]^2},$$

where the subscript of the rank function denotes the attribute within which the rank of the value in the argument is computed.

The overall risk of disclosure is an aggregation of the distances $d(\mathbf{x}, \mathbf{Y})$ for all $\mathbf{x} \in \mathbf{X}$. Many different aggregations are possible. In this work we focus on the average risk of disclosure by computing the mean of the record distances.

$$\mathcal{R}(\mathbf{X}, \mathbf{Y}) = \frac{1}{n} \log \sum_{\mathbf{x} \in \mathbf{X}} d(\mathbf{x}, \mathbf{Y}). \tag{3}$$

The smaller $\mathcal{R}(X, Y)$, the greater the risk of disclosure. The logarithm accounts for the fact that in disclosure risk the focus is on small distances. Without the logarithm, a large distance for a single record $\mathbf{x} \in \mathbf{X}$ could reduce in a significant manner the perception of risk for the overall data set; the logarithm reduces the influence of large distances.

5 Experimental results

In this section we apply the methodology described in Section 3 to analyze the relative goodness of several anonymizations. Experiments are conducted by taking as original data the "Census" and "EIA" data sets [1], which are usual test sets in the SDC literature. The "Census" contains 13 numerical attributes and 1080 records, and "EIA" contains 11 numerical attributes and 4092 records.

The anonymized data sets have been generated by applying the following methods:

- *Correlated noise addition.* Multivariate normally distributed noise is added to the records in the collected data set, that is

  $$\mathbf{Y} = \mathbf{X} + N(\mathbf{0}, \gamma\mathbf{\Sigma}),$$

  where $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$ and γ is an input parameter. Note that the covariance matrix of $\mathbf{Y}$ is proportional to the covariance matrix of $\mathbf{X}$.
- *Multiplicative noise.* We have used Höhne's variant ([6] and Ch. 3 of [7]). In a first step, each attribute value $x_j^i \in \mathbf{X}$ is multiplied by $1 \pm N(0, s)$, where s is an input parameter. Then, a transformation is applied to preserve the first and second-order moments.
- *Multivariate microaggregation.* We have used the MDAV heuristic [4]. In microaggregation, we partition the records of $\mathbf{X}$ in groups of k or more records, where records in a group are as similar as possible, and we replace each record by the corresponding centroid.

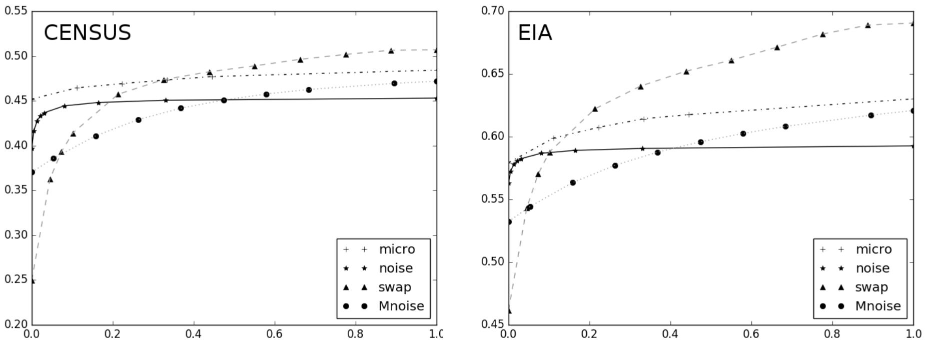

Fig. 1. Disclosure risk computed according to Eq. (3) for the anonymization methods under test and several input parameters. The x-axis shows the input parameter of the anonymization method (k, γ, p and s, respectively), so its scale should be disregarded. The y-axis shows the disclosure risk value. Left, "CENSUS" data set. Right, "EIA" data set.

– *Rank swapping.* Independently for each attribute, this method swaps the attribute's values within a restricted range: the ranks of two swapped values cannot differ by more than $p\%$ of the total number of records, where p is an input parameter.

More details about these methods can be found in [7].

5.1 Disclosure risk assessment

Recall that the comparison of anonymized data sets in Section 3 was performed on data sets that had either the same level of risk or the same level of utility. In this experimental work, we aim at determining which among the previous anonymization approaches gives better utility at a given level of disclosure risk. Thus, the first step is to find appropriate parameters for the previous anonymization algorithms that result in a given level of disclosure risk.

Figure 1 shows the disclosure risk computed according to Equation (3) for the anonymization methods under test:

1. The curve labeled "micro" shows the risk of multivariate microaggregation for values of $k \in \{5, 10, 15, 20, 25, 50\}$.
2. The curve labeled "noise" shows the risk of correlated noise addition when $\gamma \in \{0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 1, 3\}$.
3. The curve labeled "swap" shows the risk of rank swapping when $p \in \{0.01,$ $0.05,\ 0.075,\ 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.
4. The curve labeled "Mnoise" shows the risk of multiplicative noise when $s \in$ $\{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9, 1\}$.

Table 1. Utility loss measured using propensity scores (Equation (1)) and the earth mover's distance (Equation (2)) for the anonymization methods under test and for input parameters that were found to yield the same level of disclosure risk.

Methods	CENSUS		EIA	
	Propensity	EMD	Propensity	EMD
Microaggregation	4.28×10^{-4}	0.16	2.17×10^{-5}	0.040
Correlated noise addition	3.83×10^{-2}	0.38	4.22×10^{-5}	0.065
Rank swapping	3.51×10^{-3}	0.28	9.01×10^{-4}	0.091
Multiplicative noise	6.3×10^{-3}	0.29	9.85×10^{-5}	0.066

For the "Census" data set, a possible match between methods occurs at $\mathcal{R}(X, Y) = 0.45$ and is given by:

1. multivariate microaggregation with $k = 5$,
2. correlated noise addition with $\gamma = 1$,
3. rank swapping with $p = 0.2$, and
4. multiplicative noise with $s = 0.5$.

The microaggregation cluster size $k = 5$ may seem small compared to the parameter values that we get for the other methods. However, such a difference in magnitude can be explained by the fact that multivariate microaggregation is known to yield poorly homogeneous clusters when the number of dimensions is large, even if the cluster size k is small.

For the "EIA" data set, a possible match between methods occurs at $\mathcal{R}(X, Y) = 0.58$ and is given by:

1. multivariate microaggregation with $k = 5$,
2. correlated noise addition with $\gamma = 0.05$,
3. rank swapping with $p = 0.08$, and
4. multiplicative noise with $s = 0.3$.

5.2 Utility assessment

We evaluate the utility of the anonymization methods for the parameters above that were found to yield the same level of disclosure risk. The utility is evaluated using the measures based on propensity scores and EMD, that were described in Section 2.2.

We found in Section 5.1 that, for the "Census" data set, the SDC methods being compared with parameters $k = 5$, $\gamma = 1$, $p = 0.2$ and $s = 0.5$, respectively, yielded the same risk of disclosure. By comparing the utility measures for these methods, we can determine which among them is preferable in this case. Table 1 shows the results for the propensity scores and EMD measures. Both utility measures are consistent and tell us that microaggregation has the best utility, followed by rank swapping, multiplicative noise and, finally, correlated noise addition.

For the "EIA" data set, the SDC methods being compared with parameters $k = 5$, $\gamma = 0.05$, $p = 0.08$ and $s = 0.3$, respectively, yielded the same risk of disclosure. The utility results for the propensity scores and EMD measures for this data set are shown in Table 1. Like in the other data set, methods are consistently ranked by the both measures, but the ranking is different: multivariate microaggregation has the best utility, followed by correlated noise addition, multiplicative noise, and, finally, rank swapping.

The results have shown that the SDC methods under comparison perform differently in different situations. Multivariate microaggregation always had the best utility (at the given level of disclosure risk), but the relative utility performance of the other methods changed between "Census" and "EIA". This shows that, unless there are good reasons for using a given anonymization method, it is usually better to make several anonymizations at the desired level of disclosure risk and select the one that has the greatest utility.

6 Conclusions

We have described a methodology to compare different anonymizations in terms of the risk-utility trade-off they attain. It is not enough to compare methods based on the level of risk or the utility they provide, because that gives only a partial picture.

We have proposed a disclosure risk measure based on record linkage and in the spirit of the permutation paradigm (which tells that disclosure risk control comes essentially from rank permutation)

We have contributed an experimental analysis for two well-known data sets and four well-known anonymization methods. The results differ between data sets. As a conclusion from the experimental analysis, the best strategy seems to be to make several anonymizations at the desired level of disclosure risk and select the one that has the greatest utility.

Acknowledgments and disclaimer

The following funding sources are gratefully acknowledged: European Commission (projects H2020 644024 "CLARUS" and H2020 700540 "CANVAS"), Government of Catalonia (ICREA Acadèmia Prize to J. Domingo- Ferrer and grant 2014 SGR 537), Spanish Government (projects TIN2011-27076-C03-01 "CO-PRIVACY", TIN2014-57364-C2-R "SmartGlacis" and TIN2016-80250-R, "Sec-MCloud"). The authors are with the UNESCO Chair in Data Privacy, but the views in this paper are their own and do not necessarily reflect those of UNESCO.

References

1. Brand, R., Domingo-Ferrer, J., Mateo-Sanz,J.M.: Reference data sets to test and compare SDC methods for protection of numerical microdata. European Project IST-2000-25069 CASC (2002).

2. Domingo-Ferrer, J., Muralidhar, K.: New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. Information Sciences, 337:11-24 (2016).
3. Domingo-Ferrer, J., Torra, V.: 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).
4. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2):195-212 (2005).
5. Dwork, C.: Differential privacy. Automata, Languages and Programming-ICALP 2006, LNCS 4052, pp. 1–12. Springer (2006).
6. Höhne, J.: Varianten von Zufallsüberlagerung (in German). Working paper of the project "Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" (2004).
7. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., De Wolf, P.P.: Statistical Disclosure Control. John Wiley & Sons (2012).
8. Rubner, Y., Tomasi, C. and Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40(2):99-121 (2000).
9. Samarati, P., and Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International (1998).
10. Torra, V., Domingo-Ferrer, J.: Record linkage methods for multidatabase data mining. In: Information Fusion in Data Mining, pp. 101-132. Springer (2003).
11. Winkler, W.E.: Matching and Record Linkage. John Wiley & Sons, Inc. (1995).
12. Woo, M.J., Reiter, J.P., Oganian, A., Karr, A.F.: Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality 1(1):7 (2009).