

GOBIERNO
DE ESPAÑA

MINISTERIO
DE CIENCIA
E INNOVACIÓN



ARES
Advanced Research on Information Security and Privacy
CONSOLIDER INGENIO 2010

DELIVERABLE

PROGRAMME REFERENCE: CSD2007-00004
DELIVERABLE: D-WP4-T1-1
TITLE: Report on database anonymization methods
AUTHORS: Vicenç Torra (CSIC)
DATE: October 2008

Deliverable: Report on database anonymization methods

Vicenç Torra
IIIA-CSIC

October 2008

1 Introduction

The technological shift has caused that retrieval and storage of information is nowadays ubiquitous. This implies that companies and government offices store in huge databases a large amount of information about ourselves, and that misuse of this information can cause a leakage of privacy [18].

Nevertheless, the information stored should be distributed in order that the appropriate agencies and companies analyse them. Privacy preserving data mining [4, 33] and Statistical Disclosure Control [34, 12] are the area of research that is devoted to the study of methods that permit the user to extract information from the data and at the same time ensure the privacy of the respondents.

There exist two approaches to privacy preserving data mining. One of them is the cryptographic approach, based on secure multiparty computation [35]. The approach is based on the scenario in which several parties want to compute a function f of their data but without disclosing their own data to the other parties. Another approach is the perturbative or anonymization approach [21, 9, 34]. In this case, data are released once they have been perturbated. That is, given some original data (typically, a data file), some noise is added before data is released. The rationale is that an appropriate level of perturbation on the data makes the disclosure of sensitive information difficult or impossible. The difficulties of this approach are due to the fact that the more distortion data suffers, the less data utility. As data users expect that their analyses on the distorted data are equivalent to the ones with the original data, large distortion would make the results of the same analysis on the two data sets diverge.

This report describes database anonymization methods, the methods to construct the anonymized database. The structure of the report is as follows. In Section 2 we describe the scenario. In Section 3 we present a classification of the existing anonymization methods. Then, in Section 4 we review both perturbative and non-perturbative methods. Our focus is on rank-swapping and microaggregation, as these are the methods in that achieve a better performance with respect the trade-off between information loss and disclosure risk.

Section 5 reviews a family of synthetic data generators: the family of methods called Information Preserving Statistical Obfuscation (IPSO).

An annex follows this report. It includes some of our recently published papers describing new methods and analysis of the methods with respect to information loss and disclosure risk.

This report is based on [32, 22, 16].

2 Scenario

A data set X can be seen as a matrix with n rows (*records*) and k columns (*attributes*). Each row contains the values of the attributes for an individual. The attributes in a data set can be classified in three non-disjoint categories:

- **Identifiers.** They are attributes which unambiguously identify the individual, for example, the passport number.
- **Quasi-identifiers.** They are attributes which can identify the individual when some of those attributes are combined. For example, age, postal code or job cannot identify an individual, but a particular combination of age, postal code and job can lead to disclosure.
- **Confidential.** They are attributes which contain sensitive information about the individual. For example, salary.

When considering this classification, a data set X is defined as $X = id||X_{nc}||X_c$, where id are the identifiers, X_{nc} are the non-confidential quasi-identifier attributes, and X_c are the confidential attributes. Normally, before releasing a data set X with confidential attributes, a protection method ρ is applied, leading to a protected data set X'. Indeed, we will assume the following typical scenario: (i) identifier attributes in X are either removed or encrypted, therefore we will write $X = X_{nc}||X_c$; (ii) confidential attributes X_c are not modified, and so we have $X'_c = X_c$; (iii) the protection method itself is applied to non-confidential quasi-identifier attributes, in order to preserve the privacy of the individuals whose confidential data is being released. Therefore, we have $X'_{nc} = \rho(X_{nc})$. This scenario allows third parties to have precise information on confidential data without revealing to whom the confidential data belongs to. Figure 1 depicts the process of protection and release of a microdata file, as formerly explained.

In this scenario, as shown in Figure 2, an intruder might try to re-identify individuals by obtaining the non-confidential quasi-identifier data (X_{nc}) together with identifiers (Id) from other data sources. By applying record linkage between the protected attributes (X'_{nc}) and the same attributes obtained from other data sources (X_{nc}), the intruder might be able to re-identify a percentage of the protected individuals together with their confidential data (X_c). This is what protection methods try to prevent. This scenario is similar to the scenario used in [6, 32].

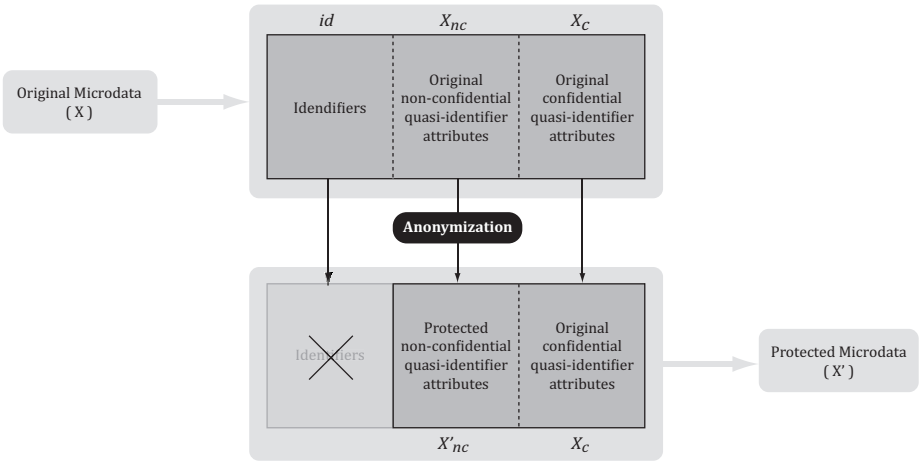


Figure 1: Data set protection and release process.

3 A Classification of Protection Methods

Protection methods can be classified depending on their effect on original data into three different categories:

- **Perturbative.** The data set is distorted adding noise. In this way, in the original data set, the combinations of values which unambiguously identify an individual (or respondent) disappear and then, new combinations appear in the protected data set. This obfuscation makes difficult for an intruder to obtain the values of the original data set. A perturbative protection method has to ensure that the statistical information in the original data set is preserved on the protected one.

- **Non-perturbative.** Non-perturbative methods do not distort the original data set. They do partial suppressions or detail reductions on the original data set. These protection methods convert the combinations of values which unambiguously identify an individual into more general ones. Thus, the re-identification process is more difficult.

- **Synthetic Data Generators.** Synthetic data generators build a data model from the original data set and subsequently, a new (protected) data set is randomly generated constrained by the model computed. This approach is very promising for the statistical disclosure, although recent works as [32] show that it is possible to link synthetic data with the original data set.

Another dimension to classify protection methods is to consider the different type of data that protection methods can be used:

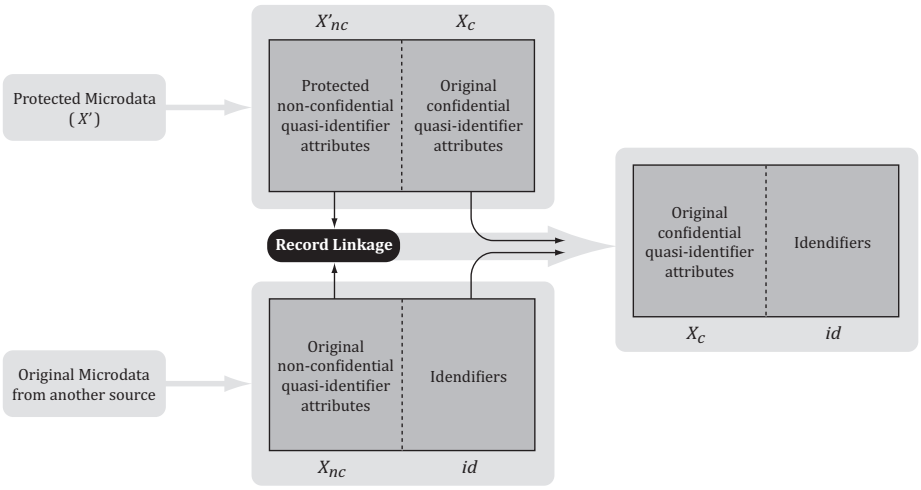


Figure 2: Disclosure Risk Scenario.

- **Numerical.** An attribute is considered numerical if arithmetic operations can be performed with it (*e.g.* age or income). Note that a numerical attribute does not necessarily have an infinite range, as in the case of age. When we are designing methods to protect numerical data, one has the advantage that arithmetic operations are possible, and the drawback that every combination of numerical values in the original data set is likely to be unique, which leads to disclosure if no action is taken.

- **Categorical.** An attribute is considered categorical when it takes values over a finite set and standard arithmetic operations do not make sense. Ordinal and nominal scales can be distinguished among categorical attributes. In ordinal scales the order between values is relevant (*e.g.* academic degree), whereas in nominal scales it is not (*e.g.* hair color). In the former case, max and min operations are meaningful while in the latter case only pairwise comparison is possible. When we are designing methods to protect this kind of data, the inability to perform arithmetic operations is an inconvenient, but the finiteness of the value range is an interesting property that can be successfully exploited.

During the past few years, special efforts have been made to develop a wide range of protection methods. Good surveys about data protection methods can be found in the literature [1, 10, 34].

In the next sections we review some of the protection methods. First we will review perturbative and non-perturbative ones. Then, we will review a family of synthetic data generators.

4 Perturbative and Non-Perturbative Protection Methods

Among all the proposed data protection methods, rank swapping and microaggregation are ones of the most used by the statistical agencies [13]. This wide application is due to they are very simple and have a low computational cost. In this section, we start reviewing these two methods in some detail and, then, some of the other ones.

4.1 Rank Swapping

Rank swapping is a widely used microdata protection method, which was originally described only for ordinal attributes in [20]. However, in the comparisons made in [10], rank swapping was ranked among the best microdata protection methods for numerical attributes.

Rank swapping with parameter p and with respect to an attribute $attr_j$ (*i.e.*, the j-th column of the original data set X) can be defined as follows: firstly, the records of X are sorted in increasing order of the values x_{ij} of the considered attribute $attr_j$. For simplicity, we assume that the records are already sorted, that is $x_{ij} \leq x_{\ell j}$ for all $1 \leq i < \ell \leq n$. Then, each value x_{ij} is swapped with another value $x_{\ell j}$, randomly and uniformly chosen from the limited range $i < \ell \leq i + p$. Finally, the sorting step is undone.

Generaly, rank swapping of a data set consists in running the algorithm explained above for each attribute to be protected, in a sequential way.

The parameter p is used to control the swap range. Normally, p is defined as a percent of the total number of records in X. Therefore, when p increases the difference between x_{ij} and $x_{\ell j}$ may increase accordingly. This fact makes re-identification more difficult, but of course the differences between the original and the protected data set are higher, decreasing in this way its statistical utility.

Example 1 *Let us consider the data set shown in the left side of Table 1. Then we protect this data set using rank swapping with $p = 2$. The protected data set is shown in the right side of the same table. For the sake of simplicity, in the example original and protected files are identically sorted.*

We have proven that this approach to determine the set of possible values that can be swapped with x_{ij} (i.e., the set $B_{ij} = \{x_{lj} | i < l \leq i + p\}$ for a particular p) can help an intruder to determine linkages. Basically, because the intruder can calculate the sets B_{ij} for all i and j and, then, with the intersection of $\cap_j B_{ij}$ might eventually be able to reidentify the ith record. Thus, standard rank swapping can be attacked using a specific record linkage algorithm. In [23] we have described this attack and proposed two variants of rank swapping that avoids it. The annex of this report includes [23].

Original Data Set X				Protected Data Set X'			
$attr_1$	$attr_2$	$attr_3$	$attr_4$	$attr'_1$	$attr'_2$	$attr'_3$	$attr'_4$
8	9	1	3	10	10	3	5
6	7	10	2	5	5	8	1
10	3	4	1	8	4	2	2
7	1	2	6	9	2	4	4
9	4	6	4	7	3	5	6
2	2	8	8	4	1	10	10
1	10	3	9	3	9	1	7
4	8	7	10	2	6	9	8
5	5	5	5	6	7	6	3
3	6	9	7	1	8	7	9

Table 1: Rank swapping example.

4.2 Microaggregation

Recently, microaggregation [7, 8] has emerged as one of the most promising protection methods. For example, [13] shows that microaggregation is used by many statistical agencies for data anonymization. This method was first defined for numerical data and then extended to categorical data [31] (see also [11]).

The basic implementation of microaggregation works as follows: given a data set of a attributes, microaggregation builds small clusters of at least k elements and replaces the original values by the centroid of the cluster to which the record belongs to. A certain level of privacy is ensured because k records have an identical protected value (k-*anonymity* [28, 27, 30, 29]). Note that there are other ways to achieve k-anonymity; in some of them (just as it happens with basic microaggregation), the released data set enjoys k-anonymity as a whole (see [3], for example). In other solutions, the data holder chooses different subsets of attributes, and k-anonymity is ensured, independently, for each of these subsets of attributes (see [14]).

When the number of attributes is large, the basic microaggregation technique suffers from a low statistical utility (see for example [2]), especially if the attributes are not much correlated. This is so because in this case the distances between original records in the data set and the centroids are quite large. Therefore, much information on the original data is lost and is not included in the released (protected) data set.

To solve this drawback, the following natural strategy is applied by statistical agencies: the data set is split into smaller blocks of attributes, and microaggregation is applied separately to each block (see [24] about attribute selection). In this way, the information loss is lower but at the cost of a loss in the achieved level of privacy. Indeed, the property of k-anonymity is not ensured now. For example, the k records which fall in the same cluster for the first block of attributes, can fall in different clusters for all the other blocks of attributes. So, the resulting protected records will not be equal and no k-

anonymity is ensured. The simplest approach for microaggregation is when the size of the attribute blocks is equal to one, in other words, each attribute is protected independently. This corresponds to *Univariate Microaggregation* or *Individual Ranking Microaggregation*.

The goal of microaggregation methods is to minimize the total sum of distances between all the elements to be protected and the centroid of the cluster where an element belongs to, *i.e* minimize the total Sum of Square Errors (SSE). The rational of this process is to make the protected data as similar as possible to the original one. In any case, the methods should provide clusters with at least k elements. The optimal multivariate microaggregation has been proven as NP-Hard [26]. For this reason, heuristic methods can be found in the literature. On the other hand, several polynomial approaches for the optimal univariate microaggregation as [15] can be found in the literature.

In this section we will explain several different algorithms that have been proposed (in more or less detail) for microaggregation. Firstly we will explain a deterministic and optimal algorithm for univariate microaggregation; it will also be used later on when we will explain two methods for projection based multivariate microaggregation: PCP microaggregation and Zscores microaggregation. Finally, we will describe one of the most used methods for heuristic microaggregation (specially for the multivariate case, although it can be applied to the univariate case as well): the MDAV (Maximum Distance to Average Vector) algorithm.

4.2.1 Optimal Univariate Microaggregation

In [15] optimal univariate microaggregation is defined as the univariate microaggregation which minimizes the Sum of Square Errors (SSE):

$$SSE = \sum_{i=1}^{C} \sum_{x_{ij} \in c_i} (x_{ij} - \bar{x}_i)^T (x_{ij} - \bar{x}_i) \tag{1}$$

where C is the total number of clusters, c_i is the i-th cluster and $\bar{x}_i$ is the centroid of c_i. The restriction is $|c_i| \geq k$, for all $i = 1, \ldots, C$.

In [8], the authors present two results for the optimal univariate microaggregation:

- **Result 1.** When elements are sorted according to an attribute, for any optimal partition, elements in each cluster are contiguous (non overlapping clusters exist)
- **Result 2.** All clusters of any optimal partition have between k and $2k-1$ elements.

This method for optimal univariate microaggregation is as follows:

Let $A = (a_1 \ldots a_n)$ be a vector of size n containing all the values for the attribute being protected. The values are sorted in ascending order so that if $i < j$ then $a_i \leq a_j$. Obviously, a_1 is the smallest element in A and a_n is the

largest element in A. Let k be an integer such that $1 \leq k < n$ (k is directly obtained from the microaggregation configuration).

Given A and k, a graph $G_{k,n}$ is defined as follows. Firstly, we define the nodes of G as the elements a_i in A plus one additional node g_0 (this node is later needed to apply the Dijkstra algorithm). Then, for each node g_i, we add to the graph the directed edges (g_i, g_j) for all j such that $i+k \leq j < i+2k$. The edge (g_i, g_j) means that the values $(a_i, \ldots, a_j)$ might define one of the possible clusters. Then, the cost of the edge (g_i, g_j) is defined as the within-group sum of squared error for such cluster. That is, $SSE = \Sigma_{l=i}^{j}(a_l - \bar{a})^2$, where $\bar{a}$ is the average record of the cluster.

Given this graph, the optimal univariate microaggregation is defined by the shortest path algorithm between the nodes g_0 and g_n. This shortest path can be computed using the Dijkstra algorithm.

Original Data Set X		Protected Data Set X'	
$attr_1$	$attr_2$	$attr'_1$	$attr'_2$
1	4	2	5
2	15	2	15.5
3	5	2	5
6	17	6.5	17.5
7	6	6.5	5
8	18	8.5	17.5
9	16	8.5	15.5

Table 2: Optimal univariate microaggregation example.

Example 2 *Let us consider the data set shown in the left side of Table 2. Then, we protect this data set using the optimal microaggregation protection method with $k = 2$. The protected data set is shown in the right side of the same table. In this example the SSE value is equal to 0.14. Note that here k-anonymity is only preserved in records 1 and 3.*

4.2.2 Projection Based Microaggregation

The basic idea of two of the microaggregation methods that we analyze herein is to project $a > 1$ attributes (corresponding to some attributes of the records) into a single one. In this way we reduce the multivariate microaggregation problem into the univariate one. The employed projection should maintain as much as possible the global statistical properties of the initial (non-projected) values. With this goal in mind, two projection methods seem particularly appealing; we explain them now.

4.2.3 Principal Component Projection

Formally Principal Component Projection (PCP in short) works as follows: let us assume that values of a attributes for n individuals are stored in a matrix X of dimension $n \times a$, where columns contain attributes and rows contain individuals. For the sake of simplicity, we will assume here that data is standardized (*i.e.*, the data has $\mu = 0$ and $\sigma = 1$, and so the covariance matrix is $S = 1/nX^T X$).

The first principal component is defined as the linear combination of the attributes which has the maximum variance. Therefore, this first principal component will be represented using a vector $z_1 = Xa_1$, for some vector a_1 with a components, to be found. Since the original values have $\mu = 0$, we have that z_1 also has $\mu = 0$, and its variance is

$$\frac{1}{n} z_1^T z_1 = \frac{1}{n} a_1^T X^T X a_1 = a_1^T S a_1 \tag{2}$$

Since S is positive-definite, the variance increases when the module of the vector a_1 does. For this reason, to find a concrete solution for the maximization of Expression (2), some constraint on the module of a_1 is needed; in this case, the search is limited to vectors a_1 with module 1 (*i.e.* $a_1^T a_1 = 1$). This is equivalent to maximize the following expression, where a Lagrange multiplier has been added to the variance:

$$M = a^T S a_1 - \lambda(a_1^T a_1 - 1) \tag{3}$$

To maximize Expression (3), the derivative with respect to the a_1 components must be made equal to 0.

$$\frac{\partial M}{\partial a_1} = 2Sa_1 - 2\lambda a_1 = 0 \tag{4}$$

The solution for such equation is $Sa_1 = \lambda a_1$, which implies that a_1 is an eigenvector of the matrix S, and λ is its corresponding eigenvalue. To determine which eigenvalue of S is the right solution, Equation (4) is left-multiplied with a_1^T, leading to

$$a_1^T S a_1 = \lambda a_1^T a_1 = \lambda.$$

Summing up, λ is the variance of z_1. Since the goal is to maximize the variance, λ is the largest eigenvalue of the matrix S, and its associate eigenvector a_1 defines the coefficients of the projection (PCP). Therefore, the final projected value is

$$PCP = \sum_{i=1}^{a} a_i x_i.$$

4.2.4 Zscores Projection

As in the previous section, we assume that values of the a attributes for the n individuals are stored in a matrix X of dimension $n \times a$. Given a record (a row) $(x_1, x_2, \ldots, x_a)$ in X, the sum of Zscores Projection is defined as the single element

$$Z = \sum_{i=1}^{a} \frac{x_i - \mu_i}{\sigma_i}$$

where μ_i is the average and σ_i is the variance of the i-th attribute, computed by taking into consideration all the records in X.

4.2.5 Algorithm for Projected Microaggregation

The main problem when one tries to extend the optimal univariate solution to the case of multivariate microaggregation is how to sort multivariate data. One possibility, as we will see later, is to order the points with respect to their distance to the global centroid of the data. MDAV is an heuristic microaggregation method that takes this information into account.

A different possibility is to reduce the dimensionality of the problem, from more than one attribute to 1 attribute, by applying some projection method. In more detail, projected multivariate microaggregation is described in Algorithm 1, when applied to a data set X with n records and A attributes.

Depending on the projection method which is applied to the attributes, we will obtain different methods of multivariate microaggregation. Due to the fact that they should preserve as much statistical properties of the data as possible (desirable in the scenario of statistical data protection), the PCP and Zscores projection methods seem to be the best choice. We call the resulting microaggregation algorithms *PCP microaggregation* and *Zscores microaggregation*.

Example 3 *Let us consider the same data set used in Example 2, the original data set shown in the left side of Table 3. Then, we protect this data set using the PCP and Zscores microaggregation with $k = 2$. The protected data sets are shown in the middle and right side of the same table. In this example, SSE value is equal to 1.02 and 0.65 for PCP and Zscores microaggregation respectively.*

4.2.6 MDAV Microaggregation

The MDAV (Maximum Distance to Average Vector) algorithm [8, 19] is an heuristic algorithm for clustering records in a data set X so that each cluster is constrained to have at least k records. This algorithm can be used for univariate microaggregation and multivariate microaggregation. The MDAV algorithm is described in Algotihm 2.

MDAV generic algorithm can be instantiated for different data types, using appropriate definitions for distance and average. Normally, *the most distant*

Algorithm 1: Projected Microaggregation

Data: X: original data set, k: integer
Result: X': protected data set

begin
  Split the data set X into r sub-data sets $\{X_i\}_{1 \leq i \leq r}$, each one with a_i attributes of the n records, such that $\sum_{i=1}^{r} a_i = A$
  foreach $(X_i \in X)$ **do**
    Apply a projection algorithm to the attributes in X_i, which results in an univariate vector z_i with n components (one for each record)
    Sort the components of z_i in increasing order
    Apply to the sorted vector z_i the following variant of the univariate optimal microaggregation method explained in Section 4.2.1: use the algorithm defining the cost of the edges $\langle z_{i,s}, z_{i,t} \rangle$, with $s < t$, as the within-group sum of square error for the a_i-dimensional cluster in X_i which contains the original attributes of the records whose projected values are in the set $\{z_{i,s}, z_{i,s+1}, \ldots, z_{i,t}\}$
    For each cluster resulting from the previous step, compute the v_i-dimensional centroid and replace all the records in the cluster by the centroid
end

record and the *closest records* are computed using the Euclidean distance, and *the average record* is defined as the arithmetic mean of the records. This same mean record is used to replace the original records when building the protected data set.

[11] describes a variation of MDAV for categorical data.

Example 4 *Let us consider the same data set used in Example 2 and 3, the original data set is shown in the left side of Table 4. Then, we protect this data set using the MDAV microaggregation protection method with* $k = 2$. *The protected data set is shown in the right side of the same table. In this example, SSE value is equal to 0.49.*

4.3 Other Methods

In this section we briefly review additive noise, data distortion by probability distribution, resampling, and lossy compression.

Additive noise. Data is perturbated adding noise into the original file. In

Original Data Set X		PCP Prot. Data Set X'		Zscores Prot. Data Set X'	
$attr_1$	$attr_2$	$attr'_1$	$attr'_2$	$attr'_1$	$attr'_2$
1	4	3.67	5.0	2.0	4.5
2	15	4.0	16.0	4.5	10.5
3	5	3.67	5.0	2.0	4.5
6	17	4.0	16.0	7.67	17.0
7	6	3.67	5.0	4.5	10.5
8	18	8.5	17.0	7.67	17.0
9	16	8.5	17.0	7.67	17.0

Table 3: Projection based microaggregation example.

order to have a low information loss, the random noise has the same correlation structure as the original data.

Data distortion by probability distribution. This masking method is defined by three steps. The steps are applied to each variable separately. First, we need to identify an underlying density function and determine the parameters. Second, generate distorted series using the estimated density function. Third, the distorted series are put in place of the original ones.

Resampling. This method, that is applied to each variable separately, consists on applying three steps. Their description assumes that V is the variable and that the database contains n records. The steps are as follows. First, obtain t independent samples for variable V obtained from the original file with replacement. Let $X_1, \ldots, X_t$ be such samples. Second, order the samples (using the same ordering criteria in all cases). Let X'_i be the samples once ordered. Third, define the value for variable V of the new j-th record (for $j = 1, \ldots n$) as the average of the j-th values in samples X'_i (*i.e.,* $x'_{j,V} = \sum_{i=1}^{t} X'_{ij}/t$ for $j = 1, \ldots, n$).

Lossy compression. The approach consists on taking a numerical data file as an image. Then, a lossy compression method is applied to the *image*, obtaining a *compressed image*. Such *compressed image* corresponds to the masked file. JPEG has been used for such purpose.

5 Synthetic data generators

In recent years, a new research trend has emerged in the perturbative approach. It consists of publishing synthetic data instead of the original one. The rationale is that synthetic data cannot compromise the privacy of the respondents as the data is not *"real"*. Synthetic data generation methods consist first of constructing a model of the data, and then generating artificial data from this model.

Algorithm 2: MDAV

```
Data: X: original data set, k: integer
Result: X': protected data set
begin
    while (|X| > k) do
        Compute the average record x̄ of all records in X
        Consider the most distant record x_r to the average record x̄
        Form a cluster around x_r. The cluster contains x_r together
        with the k − 1 closest records to x_r
        Remove these records from data set X
        if (|X| > k) then
            Find the most distant record x_s from record x_r
            Form a cluster around x_s. The cluster contains x_s together
            with the k − 1 closest records to x_s
            Remove these records from data set X
        Form a cluster with the remaining records
end
```

Original Data Set X		Protected Data Set X'	
$attr_1$	$attr_2$	$attr'_1$	$attr'_2$
1	4	1.5	4.5
2	15	1.5	12.33
3	5	5.33	4.5
6	17	5.33	17.5
7	6	5.33	12.33
8	18	8.5	17.5
9	16	8.5	12.33

Table 4: MDAV microaggregation example.

The difficulties of the approach are twofold. On the one hand, we have proven recently [32] that although data is synthetic, some reidentification is possible. This represents a threat to respondents' privacy. Therefore, it is still meaningful to consider the analysis of privacy for this kind of data. On the other, as the synthetic data is generated from a particular data model built from the original data, all those aspects that are not explicitly included in the model, are often not later found in the data[1].

Several methods for synthetic data generation have been developed in the last few years. We will review here the family of methods called Information Preserving Statistical Obfuscation (IPSO). This is a family of methods, pro-

[1]In a recent paper under submission we have proven that for some type of applications (e.g. clustering) the quality of data is similar to the one of perturbative masking methods as microaggregation. Nevertheless, the same situation applies with respect to risk.

posed recently in [5], with the property that when their complexity is increased, the generated data resembles more to the original one. Here, resemblance is evaluated in terms of the properties (statistics) the data satisfy and not on the exact records the data contain.

These methods are known as IPSO-A, IPSO-B and IPSO-C. IPSO-A is the simplest method and IPSO-C is the most elaborated one. Thus, the data generated by IPSO-C is more similar to the original data than the data generated by IPSO-A.

All three methods presume that the variables of the original data file can be divided into two sets of variables X and Y, where X contains the confidential outcome variables and Y the quasi-identifier variables. Here, we understand by confidential variables the ones that should not be disclosed to an intruder, and by quasi-identifier variables the ones that permit to identify a respondent. For example, the age, town of residence and job might be considered as quasi-identifiers (e.g. there is only one person born in 1954, that is dentist in Smallville) and salary a confidential one.

The IPSO-A proceeds as described below. Take X as independent and Y as dependent variables, and then fit a multiple regression model of Y on X. Compute Y'_A with this model. IPSO-A releases the variables X and Y'_A in place of X and Y.

The following can be stated with respect to IPSO-A (see e.g. [5, 32]). In the above setting, conditional on the specific confidential attributes x_i, the quasi-identifier attributes Y_i are assumed to follow a multivariate normal distribution with covariance matrix $\Sigma = \{\sigma_{jk}\}$ and a mean vector $x_i B$, where B is the matrix of regression coefficients. Let $\hat{B}$ and $\hat{\Sigma}$ be the maximum likelihood estimates of B and Σ derived from the complete data set (y, x). If a user fits a multiple regression model to (y'_A, x), she will get estimates $\hat{B}_A$ and $\hat{\Sigma}_A$ which, in general, are different from the estimates $\hat{B}$ and $\hat{\Sigma}$ obtained when fitting the model to the original data (y, x).

The goal of IPSO-B is to modify y'_A into y'_B in such a way that the estimate $\hat{B}_B$ obtained by multiple linear regression from (y'_B, x) satisfies $\hat{B}_B = \hat{B}$.

Finally, the goal of IPSO-C is to obtain a matrix y'_C such that when a multivariate multiple regression model is fitted to (y'_C, x), *both* sufficient statistics $\hat{B}$ and $\hat{\Sigma}$ obtained on the original data (y, x) are preserved.

6 Conclusions

In this report we have reviewed major database anonymization methods. An annex follows this report with some of the papers we have prepared this year about methods, and their performance with respect to information loss and disclosure risk.

References

[1] Adam, N. R., Wortmann, J. C.,(1989), Security-control for statistical databases: a comparative study, ACM Computing Surveys, Volume: 21, 515-556.

[2] Aggarwal, C. (2005) On k-anonymity and the curse of dimensionality, Proceedings of the 31st International Conference on Very Large Databases, pages 901-909.

[3] Aggarwal, G., Feder T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., Zhu, A., (2006) Achieving anonymity via clustering, Proceedings of the 25th ACM Symposium on Principles of Databases Systems, pages 153-162.

[4] Agrawal, R., Srikant, R. (2000) Privacy Preserving Data Mining, Proc. of the ACM SIGMOD Conference on Management of Data, 439-450.

[5] Burridge, J. (2003) Information preserving statistical obfuscation, Statistics and Computing, 13:321–327.

[6] Wang, D.W., Liau, C.J., Hsu, T.S., (2007), An epistemic framework for privacy protection in database linking, Data & Knowledge Engineering, Elsevier, Volume: 61, 176-205

[7] Defays, D., Nanopoulos, Ph., (1993), The Small Aggregates Method. Proceedings of the 92 Symposium on Design and Analysis of Longitudinal Surveys. Ottawa, Statistics Canada.

[8] Domingo-Ferrer, J., Mateo-Sanz, J.M., (2002), Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering, Volume: 14, 189-201.

[9] Domingo-Ferrer, J., Torra, V. (2001) Disclosure Control Methods and Information Loss for Microdata, Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, Pages 91-110, 2001.

[10] Domingo-Ferrer, J., Torra, V. (2001) A Quantitative Comparison of Disclosure Control Methods for Microdata, Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, Pages 111-133, 2001.

[11] Domingo-Ferrer, J., Torra, V. (2005) Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation, Data Mining and Knowledge Discovery 11:2 195-212.

[12] Doyle, P., Lane, J., Theeuwes, J., Zayatz, L., eds. (2001), Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies, Elsevier Science.

[13] Felsö, F., Theeuwes, J., Wagner, G., (2001) Disclosure Limitation in Use: Results of a Survey, Pages 17-42 of [12].

[14] Fung, B., Wang, K., Yu, P., (2005) Top-down specialization for information and privacy preservation, Proceedings of the 21st IEEE International Conference on Data Engineering, pages 205-216.

[15] Hansen, S., Mukherjee, S. (2003) A Polynomial Algorithm for Optimal Univariate Microaggregation. Trans. on KDE, 15:4, 1043-1044.

[16] Ladra, S., Torra, V. Information loss for synthetic data through fuzzy clustering, submitted.

[17] Ladra, S., Torra, V. (2008) On the comparison of generic information loss measures and cluster-specific ones, Int. J. of Unc., Fuzz. and Knowledge Based Systems 16(Supplement-1) 107-120.

[18] Lane, J., Heus, P., Mulcahy, T. (2008) Data Access in a Cyber World: Making Use of Cyberinfrastructure, Transactions on Data Privacy 1:1 2-16.

[19] Mateo-Sanz, J.M., Domingo-Ferrer, J., (1999), A method for data-oriented multivariate microaggregation, Statistical Data Protection for Official Publications of the European Communities, pages 89-99.

[20] Moore, R., (1996), Controlled data wwapping techniques for masking public use microdata sets, U. S. Bureau of the Census (Unpublished manuscript).

[21] Mukherjee, S., Chen, Z., Gangopadhyay, A. (2006) A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms, The VLDB Journal 15 293-315.

[22] Nin, J. (2008) Contributions to record linkage for disclosure risk assessment, PhD Dissertation, Universitat Autònoma de Barcelona.

[23] Nin, J., Herranz, J., Torra, V. (2007) Rethinking Rank Swapping to Decrease Disclosure Risk, Data and Knowledge Engineering, 64:1 346-364.

[24] Nin, J., Herranz, J., Torra, V. (2008) How to Group Attributes in Multivariate Microaggregation, Int. J. of Unc., Fuzz. and Knowledge Based Systems 16(Supplement-1) 121-138.

[25] Nin, J., Herranz, J., Torra, V. (2008) On the Disclosure Risk of Multivariate Microaggregation, Data and Knowledge Engineering, in Press.

[26] Oganian, A., Domingo-Ferrer, J. (2000) On the Complexity of Optimal Microaggregation for Statistical Disclosure Control, Statistical J. United Nations Economic Commission for Europe, 18, 4, 345-354.

[27] Samarati, P. (2001) Protecting Respondents' Identities in Microdata Release, IEEE Trans. on Knowledge and Data Engineering, 13:6 1010-1027.

[28] Samarati, P., Sweeney, L. (1998) Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, SRI Intl. Tech. Rep.

[29] Sweeney, L. (2002) Achieving k-anonymity privacy protection using generalization and suppression, Int. J. of Unc., Fuzz. and Knowledge Based Systems 10:5 571-588.

[30] Sweeney, L. (2002) k-anonymity: a model for protecting privacy, Int. J. of Unc., Fuzz. and Knowledge Based Systems 10:5 557-570.

[31] Torra, V. (2004) Microaggregation for categorical variables: a median based approach, Proc. Privacy in Statistical Databases (PSD 2004), Lecture Notes in Computer Science 3050 162-174.

[32] Torra, V., Abowd, J. M., Domingo-Ferrer, J. (2006), Using Mahalanobis distance-based record linkage for disclosure risk assessment, Privacy in Statistical Databases 2006, Lecture Notes in Computer Science, Springer-Verlag, volume: 4302, 233-242.

[33] Verykios, V. S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., Theodoridis, Y. (2004) State-of-the-art in Privacy Preserving Data Mining, SIGMOD Record 33:1 50-57.

[34] Willenborg, L., de Waal, T. (2001) *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, Springer-Verlag.

[35] Yao, A. C. (1982) Protocols for Secure Computations, Proc. of 23rd IEEE Symposium on Foundations of Computer Science, Chicago, Illinois, 160-164.

ANNEX

Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 64 (2008) 346–364
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Rethinking rank swapping to decrease disclosure risk

Jordi Nin *, Javier Herranz, Vicenç Torra

IIIA, Artificial Intelligence Research Institute, CSIC, Spanish National Research Council, Campus UAB s/n, 08193 Bellaterra, Catalonia, Spain

Received 12 April 2007; accepted 31 July 2007
Available online 6 September 2007

Abstract

Nowadays, the need for privacy motivates the use of methods that allow to protect a microdata file both minimizing the disclosure risk and preserving the data utility. A very popular microdata protection method is rank swapping.

Record linkage is the standard mechanism used to measure the disclosure risk of a microdata protection method. In this paper we present a new record linkage method, specific for rank swapping, which obtains more links than standard ones. The consequence is that rank swapping has a higher disclosure risk than believed up to now.

Motivated by this, we present two new variants of the rank swapping method, which make the new record linkage technique unsuitable. Therefore, the real disclosure risk of these new methods is lower than the standard rank swapping.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Privacy in statistical databases; Statistical disclosure risk; Record linkage; Rank swapping

1. Introduction

In many situations, it is necessary to release confidential data providing useful information like the data statistics, but without revealing confidential information of the involved individuals. For example, recent surveys of web users show that a majority of them are unwilling to provide sensitive data to a web site unless they know that privacy protection measures are in place [1].

During the past few years, special efforts have been made to develop a wide range of protection methods. Good surveys about data protection methods can be found in the literature [2,10]. Among all the proposed data protection methods, rank swapping is one of the most used, because it is very simple and has a low computational cost. The original proposal [7], which worked better for tabular data, was extended to the scenario of ordinal microdata in [21]. The basic idea of the method is swapping values of a database inside a predefined (ordered) rank. A good survey on rank swapping techniques can be found in [18]. We explain the method for microdata in more detail in Section 2.3, along with a simple example.

* Corresponding author.
E-mail addresses: jnin@iiia.csic.es (J. Nin), jherranz@iiia.csic.es (J. Herranz), vtorra@iiia.csic.es (V. Torra).

0169-023X/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2007.07.006

The main goal of rank swapping, and in general of all protection methods, is to minimize both the *disclosure risk* (DR) and the *information loss* (IL) of the protected released microdata. Disclosure risk measures the capacity of an intruder to obtain some information about the original dataset from the protected one, and information loss measures the reduction of the statistical utility of the protected microdata with respect to the original one.

However, there is a trade off between these two parameters: when one of them decreases the other one increases. Finding the optimal combination of these two measures becomes a difficult and challenging task. Moreover, in some situations, an organization could be interested in releasing the data by fixing a desirable level for one of the parameters. For these two reasons it becomes necessary to compute both measures in a very accurate manner to release the protected dataset ensuring an enough protection level and statistical utility.

Some approaches are used to calculate the information loss. In [9] the authors calculate the average difference between some statistics computed on both the original and the protected microdata. A probabilistic variation of this measure was presented in [20] to ensure that the information loss value is always inside the interval $[0, 1]$. A different approach was presented in [3], where some measures (*accuracy*, *completeness* and *consistency*) are calculated over the protected data to evaluate the information loss.

In order to compute the disclosure risk, many works like [9,23,30] consider *record linkage* methods [24,28,29]. Such methods try to model the situation where an intruder sees the protected dataset, whereas he has access to some records of the original dataset obtained from other sources. The goal of the methods (i.e., of the intruder) is to link the original records with the corresponding records in the protected dataset. Of course, the more records can be linked by means of these record linkage methods, the more disclosure risk has the employed protection method. Some examples of standard record linkage methods are distance based ones and probabilistic ones.

Other methods to evaluate the disclosure risk can be considered. In the recent work [27], the authors define a framework for privacy protection where the intruder can only query the database by using propositional sentences. Therefore, if the database answers these questions with enough level of generalization, it is really hard for the intruder to infer any confidential information of a specific individual. The measure of disclosure risk in this scenario is the percentage of individuals for which an intruder is able to discover the value of a confidential attribute.

Normally, information loss and disclosure risk are combined to obtain a global value about a specific protection method, this value weights the relationship between IL and DR. One approach was presented in [9], where the authors combine both information loss and disclosure risk in a *Score* using the arithmetic mean. A different approach is the R-U (risk-utility) maps [13–15], this measure shows in a graphic way the relationship between any numerical measure of statistical disclosure risk (R) and any numerical measure of data utility (U). Both measures, R and U, can be general or specific for a certain protection method.

Recently, some authors have been working on possible improvements to record linkage methods, in order to give better estimates of disclosure risk for specific protection methods. For example, [26] studies distance based record linkage for the IPSO protection methods [5], a synthetic data generator based on multiple regression models. The authors show that the disclosure risk increase when the Euclidean distance is replaced by the Mahalanobis one. A similar result is obtained for microaggregation [11], a protection method which builds small clusters of at least k records and replaces the original values of each record with the centroid of the cluster where this record belongs. Ref. [25] shows that the disclosure risk of microaggregation increases when the Euclidean distance is replaced by a specific one that takes into account the clusters. These two works, as well as the results that we present here for rank swapping, show that the design of protection methods should avoid ad hoc attacks.

In this paper, we describe a new record linkage method, specific for rank swapping. This method exploits the fact that only a few values of the dataset are eligible when doing rank swapping. Using this information, we can limit the records where the record linkage method must be applied, decreasing in this way the probability of finding incorrect links. The result is an increase on the number of correct links, and therefore an increment in the disclosure risk of standard rank swapping.

Then, we propose two new protection methods, obtained as variants of rank swapping, which are called *rank swapping p-distribution* and *rank swapping p-buckets* and which are in some way immune to the effect

of the new record linkage method. The main idea of these methods is that each attribute value can be potentially (maybe with very low probability, but never equal to 0) swapped with any other value. For this reason, an intruder looking for record linkages will not be able to limit the swap interval with total confidence; this will lead to a higher number of incorrect linkages. Our empirical results show that the two new proposed protection methods obtain a better *score* (the average of information loss and disclosure risk) than standard rank swapping, once the new record linkage method is taken into account.

The rest of the paper is organized as follows. In Section 2 we explain some preliminaries needed later on. We present in Section 3 a description of the new record linkage method, specific for rank swapping. In Section 4 we propose the two new variants of rank swapping. We present some experimental comparisons among rank swapping methods in Section 5. In Section 6 we discuss about the real disclosure risk of the new variants of rank swapping. Finally, in Section 7, we draw some conclusions and present some future work.

2. Preliminaries

In this section, we explain some basics about microdata protection methods, including the standard way of computing the score of such a method, by combining its information loss and its disclosure risk. Then, we give a description of the rank swapping protection method, which is illustrated with a toy example.

2.1. Microdata protection methods

A dataset X can be thought as a matrix with n rows (*records*) and k columns (*attributes*). Each row contains the values of the attributes for some individual. Each attribute can be classified as *identifier*, when they unambiguously identify the individual (for example, the passport number), or *quasi-identifier* otherwise (for example, the age). Note however that some combinations of quasi-identifier attributes can identify an individual, depending on the scenario. For example, the combination (age = 16, city = NY, job = photographer in the 'Daily Bugle') univocally identifies Peter Parker, Spiderman. Quasi-identifier attributes can be divided in two classes: *confidential* and *non-confidential*, depending on whether they contain private information. For example, the zip code is non-confidential, whereas the salary can be a confidential attribute.

Normally, before releasing a dataset with confidential attributes, a protection method ρ is applied, leading to a protected dataset X'. Indeed, we will assume the following typical scenario: (i) identifier attributes in X are either removed or encrypted, therefore we will write $X = X_{nc} \| X_c$; (ii) confidential quasi-identifier attributes X_c are not modified, and so we have $X'_c = X_c$; (iii) the protection method itself is applied to non-confidential quasi-identifier attributes, in order to preserve the privacy of the individuals whose confidential data is being released. Therefore, we have $X'_{nc} = \rho(X_{nc})$. This scenario, which was used first in [10] to compare several protection methods, has also been adopted in other works [26,27].

2.1.1. Record linkage

Once the modified dataset X' is released, everybody can see its content: $X' = X'_{nc} \| X_c$. We assume now that an intruder has access to some other original dataset $Y = Y_{id} \| Y_{nc}$ which includes the identifier and the non-confidential quasi-identifier attributes of some of the individuals whose data is in X. If this intruder is able to correctly link a record $y_{nc} \in Y_{nc}$ (which is also in X_{nc}) with the corresponding protected record $x'_{nc} \in X'_{nc}$, then he will know that the matching (not modified) confidential information x_c belongs to the individual with identifier y_{id}, breaking therefore the privacy of this individual.

For this reason, the disclosure risk of a protection method must be related to the difficulty of finding correct linkages between original and protected data. Two record linkage (also known as *re-identification*) methods are usually considered in the literature:

- *Distance based record linkage* (*DB-RL*). It can be applied when a distance between records in the original and protected datasets can be defined. If this is the case, every protected record is linked to the closest original record. When the data are numerical, one usually employs the Euclidean distance, after normalizing the whole dataset. This record linkage method was proposed in [22].

- *Probabilistic record linkage* (*P-RL*). In this case, the matching algorithm uses the linear sum assignment model to choose the original and protected records that must be matched, in order to compute this model, EM (Expectation–Maximization) algorithm is normally used. Formally, for each pair of records (a,b) where a is an original record of the original dataset X and b is a protected record of the protected dataset X', we define a coincidence vector $\gamma(a,b) = (\gamma_1(a,b) \dots \gamma_n(a,b))$, where $\gamma_i(a,b)$ is defined as 1 if $V_i(a) = V_i(b)$ and as 0 if $V_i(a) \neq V_i(b)$. Then an index is computed over this coincidence vector. Afterwards, using such index, pairs are classified as either a linked pair (LP) or a non-linked pair (NP). This method was introduced in [19].

2.2. Score: Information loss and disclosure risk

Among the different measures explained in the Introduction to evaluate the quality of a data protection method, we will use the *score*. This approach, first presented in [9] and used in many other works [23,30], is a simple and natural way to evaluate the trade-off between the information loss and the disclosure risk. Besides, it allows us to compare easily our new rank swapping methods with all protection methods evaluated in [10]. In particular, standard rank swapping.

As stated, the score takes into account two different contradictory aspects, the *disclosure risk* (DR), that is the risk that an intruder obtains any correct relation between the protected and the original data, and the *information loss* (IL) produced in the protected microdata. It is important to note that IL and DR are inversely related parameters: if one of them increases, the other one necessarily decreases. The two extreme cases are the following ones: (i) if the original microdata are released, then information loss is zero, but the disclosure risk is maximum; (ii) if the original microdata are encrypted and then released, the disclosure risk is (almost) zero, but the information loss is maximum.

Formally, the score for a given method is defined by

$$\text{score} = \frac{(IL + DR)}{2}.$$

That is, the average of the information loss and the disclosure risk.

There are different possible ways to measure IL and DR. We have mainly used the expressions in [10] for the sake of comparability with previous results. Nevertheless, it is worth mention that such measures are generic in the sense that they are not specific to a particular protection method or a particular data use. Specific ones might also be considered. We will consider this case in Section 3 in what concerns disclosure risk.

Formally, IL and DR are computed as follows.

- *Information loss.* The overall IL is computed as $IL = 100(0.2IL_1 + 0.2IL_2 + 0.2IL_3 + 0.2IL_4 + 0.2IL_5)$, where
  (i) IL_1: It is the mean absolute error of the original microdata X vs. the protected data X'. It is defined as the average of the absolute values of differences between the corresponding components (records) in both matrices. We understand ''corresponding components'' as the map of each record in X with the nearest record in X' using a p-dimensional Euclidean distance. Then, we define the mean variation of X vs X' as

$$IL_1 = X - X' = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n}\frac{|x_{ij}-x'_{ij}|}{|x_{ij}|}}{nk}$$

  (ii) IL_2: Let $\overline{X}$ and $\overline{X}\prime$ be the vectors of averages of attributes (rows) in X and X', we define the mean variation of this two vectors as

$$IL_2 = \overline{X} - \overline{X}' = \frac{\sum_{j=1}^{k}\frac{|\bar{x}_j-\bar{x}'_j|}{|\bar{x}_j|}}{k}$$

(iii) IL_3: Let V and V' be the covariance matrices of X and X', we compute the mean variation of these two matrices as

$$IL_3 = V - V' = \frac{\sum_{j=1}^{k}\sum_{1\leqslant i\leqslant j}\frac{|v_{ij}-v'_{ij}|}{|v_{ij}|}}{\frac{(k+1)k}{2}}$$

(iv) IL_4: Let S and S' be the vectors of variances of attributes (rows) in X and X', these vectors are the diagonal of V and V' respectively, we compute the mean variation of these two vectors as

$$IL_4 = S - S' = \frac{\sum_{j=1}^{k}\frac{|v_{jj}-\bar{x}'_{jj}|}{|\bar{x}_{jj}|}}{k}$$

(v) IL_5: Let R and R' be the correlation matrices of X and X', we compute the mean variation of these two matrices as

$$IL_5 = R - R' = \frac{\sum_{j=1}^{k}\sum_{1\leqslant i\leqslant j}|r_{ij} - r'_{ij}|}{\frac{(k-1)k}{2}}$$

- *Disclosure risk* (DR). Three alternatives are considered for measuring the disclosure risk. They are two variations of record linkage and the interval disclosure. Again, as in the case of information loss, other approaches might be considered here. Record linkage presents the advantage that it is a unified approach suitable for all types of data (it can even be applied to synthetic data [26]) and also applicable in the case of considering the risk with multiple databases. So, it permits to model different scenarios. Interval disclosure is an alternative approach where disclosure does not mean 'finding the correct link' but 'obtaining an approximation of the value of a sensitive attribute'. So, in the sense of risk, it is a looser measure.Formally, we define the measure of the risk giving half weight to record linkage and half weight to interval disclosure. Then, for record linkage we consider the average of different alternatives. Formally, this corresponds to $DR = 0.25 \cdot DLD + 0.25 \cdot PLD + 0.5 \cdot ID$, where
  (i) DLD, the *Distance Linkage Disclosure risk*, is the average percentage of correctly linked records using distance based record linkage,
  (ii) PLD, the *Probabilistic Linkage Disclosure risk*, is the average percentage of correctly linked records using probabilistic record linkage, and
  (iii) ID, the *Interval Disclosure risk*, is the average percentage of original values falling into the intervals around their corresponding masked values. These three values are computed over the number of attributes that the intruder is assumed to know (from one to half).

2.3. Rank swapping

Rank swapping is a widely used microdata protection method, which was originally described only for ordinal attributes in [21]. However, in the comparisons made in [10], rank swapping was ranked among the best microdata protection methods for numerical attributes.

Rank swapping with parameter p and with respect to an attribute $attr_j$ (i.e., the jth column of the original dataset X) can be defined as follows: first, the records of X are sorted in increasing order of the values x_{ij} of the considered attribute $attr_j$. For simplicity, assume that the records are already sorted, that is $x_{ij} \leqslant x_{\ell j}$ for all $1 \leqslant i < \ell \leqslant n$. Then, each value x_{ij} is swapped with another value $x_{\ell j}$, randomly and uniformly chosen in the set of still unswapped values in the limited range $i < \ell \leqslant i + p$. Finally, the sorting step is undone.

Usually, when rank swapping is applied to a dataset, the algorithm explained above is run for each attribute to be protected, in a sequential way. The parameter p is used to control the swap range. Normally, p is defined as a percent of the total number n of records in X. Therefore, when p increases the difference between x_{ij} and $x_{\ell j}$ may increase. This fact makes record linkage more difficult, but of course the information loss of the protected data is higher, decreasing in this way its statistical utility.

Table 1
Rank swapping example

Original file X				Protected file X'			
$attr_1$	$attr_2$	$attr_3$	$attr_4$	$attr'_1$	$attr'_2$	$attr'_3$	$attr'_4$
8	9	1	3	10	10	3	5
6	7	10	2	5	5	8	1
10	3	4	1	8	4	2	2
7	1	2	6	9	2	4	4
9	4	6	4	7	3	5	6
2	2	8	8	4	1	10	10
1	10	3	9	3	9	1	7
4	8	7	10	2	6	9	8
5	5	5	5	6	7	6	3
3	6	9	7	1	8	7	9

Table 2
Score calculation for the rank swapping example

IL	DLD	PLD	ID	Score
39.22	17.5	0.0	44.81	32.99

Example 1. Let us consider the microdata file shown in the left side of Table 1. Then we protect this file using rank swapping with $p = 2$. The protected microdata is shown in the right side of the same table. For the sake of simplicity, in the example original and protected files are identically sorted. The score obtained by applying to these datasets the measures described in Section 2.2 is shown in Table 2.

3. A new record linkage method for rank swapping

As we stated in Section 2, only standard generic measures, like distance based or probabilistic record linkage, are considered when the disclosure risk of a protection method is computed. The use of these generic measures causes an underestimation of the resulting disclosure risk.

We consider here a specific method for rank swapping. We will call this method *rank swapping record linkage* (RS-RL for short). The idea is quite easy to understand: standard rank swapping swaps one original value with one of the p following values in the sorted table. Therefore, if the protected values of the attribute are known, as in our scenario, it is possible to limit the protected records into which a specific original record can have been transformed. One must compare the original record x_i that he wants to link only with $2p$ records in the protected microdata file (note that a protected value can be either the source or the destination in the swap process). In other words, for every original value x_{ij}, there is an efficiently computable set $B(x_{ij})$ of $2p$ protected records which may be the result of transforming the original record x_i.

Obviously, if more than one attribute is known, it is possible to repeat the process for each attribute. If the original record $x_i = (x_{i1}, \ldots, x_{ic})$ to be linked has c attributes $attr_1, \ldots, attr_c$, then the matching protected record x'_ℓ will necessarily satisfy the condition

$$x'_\ell \in \cap_{1 \leqslant j \leqslant c} B(x_{ij}).$$

That is, the search of the linkage is reduced to the intersection of the sets of possible protected records. Of course, the more attributes are considered, the less records will be in this intersection, and therefore the probability of finding the correct record linkage will increase. This effect is illustrated in Fig. 1. In particular, if some intersection (for some combination of the protected attributes) gives a unique possible record, we can be sure that this is the protected record which matches with the considered original record, because this linkage method does not introduce error probabilities. This is in contrast to the standard record linkage methods, where an original record is compared to all the protected records, possibly leading to incorrect linkages.

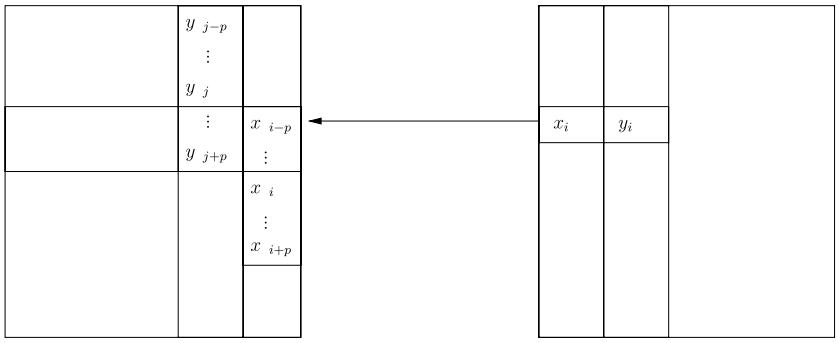


Fig. 1. Graphic representation of disclosure risk.

Example 2. Let us illustrate this fact throughout the example showed in Table 1. Consider the standard distance based record linkage method (with the Euclidean distance) applied to the original record $(6,7,10,2)$. The distances between this record and all the protected records are computed, the closest protected record is $(6,7,6,3)$, which is not the matching one. Therefore, this method leads to an incorrect linkage $(6,7,10,2) \leftrightarrow (6,7,6,3)$. On the other hand, consider the new specific technique applied to the same record $(6,7,10,2)$. The set of possible protected values consistent with a 6 in the first original attribute is $B(attr_1 = 6) = \{(4,1,10,10), (5,5,8,1), (6,7,6,3), (7,3,5,6), (8,4,2,2)\}$. Analogously for the other three attributes, we obtain $B(attr_2 = 7) = \{(5,5,8,1), (2,6,9,8), (6,7,6,3), (1,8,7,9), (3,9,1,7)\}$, $B(attr_3 = 10) = \{(5,5,8,1), (2,6,9,8), (4,1,10,10)\}$ and $B(attr_4 = 2) = \{(5,5,8,1), (8,4,2,2), (6,7,6,3), (9,2,4,4)\}$. Therefore, as the intersection of the four sets is just the protected record $(5,5,8,1)$, in this case we obtain the correct linkage. Note that if there had been more than one record in the intersection, the closest one to the considered original record would have been chosen.

Note also that this new method has been defined assuming that the value of the parameter p is known. In situations where this value is kept secret by the owner of the original dataset, then the method can be applied by first fixing an upper bound for the value of p (for example, the 20% of the number of entries of the database). Of course, the results of the method are optimal when the exact value of p is used. This assumption is quite realistic, for example, all available microdata files in the EUROSTAT web page [17] include a full description of the anonymization criteria that has been applied.

In Table 3, we show the number of correctly linked records with the three considered record linkage methods, when different collections of attributes are known. We also show the average disclosure risk. Note that the record linkage process uses only the known attributes to compute the nearest record. From the comparison between the average disclosure risk using the new method and the DB-RL and P-RL showed in Table 2, it is clear that the real disclosure risk is much larger than the standard one.

If we compute the score by using only the new record linkage method for the disclosure risk measure, then the new score in the example is equal to 41.44. If the disclosure risk is computed by combining the standard and the new record linkage methods, then the score is equal to 33.31.

As a final note we want to remark that although the method introduced here seems to be quite intuitive, it has not been described, to the best of our knowledge, in the literature. Literature is focused on the use of standard record linkage methods and no attention is given to specific ones. Nevertheless, as an intruder might

Table 3
Correct links and average disclosure risk of Example 1, computed with rank swapping (RS-RL)/distance based (DBRL)/probabilistic (PRL) record linkage

	$attr_1$	$attr_{1-2}$	$attr_{1-3}$	$attr_{1-4}$	Average DR
RS-RL	0	2	7	8	42.5
DB-RL	0	2	4	1	17.5
P-RL	0	0	0	0	0.00

exploit any weakness in a protection method, the development and study of these methods is of interest to make methods as effective as possible.

4. New rank swapping methods

In this section, we present two new variants of rank swapping which resist the new record linkage method presented in Section 3. The main idea in both variants is the same: to eliminate the fact that the swap interval is closed.

4.1. Rank swapping p-distribution

As we have explained in Section 2 the swap interval in the rank swapping is defined by the parameter p, this fact is exploited by the record linkage technique defined in Section 3 to increase the number of linked records. For this reason, rank swapping p-distribution defines the swap interval using a normal probability distribution, defined by $\mu = \sigma = 0.5 \cdot p$.

This modification makes possible that any value in the dataset could be selected, obviously very distinct values have lower probability to be elected than similar values, but never equals to zero. Therefore some values x_{ij} are swapped with values out of the standard interval $\ell \in [i+1, i+p]$, we can observe this in Fig. 2 where the swap interval is defined by a normal probability distribution, in this case it is clear that the swap interval is an open interval. When this effect is propagated to all protected attributes, the RS-RL method becomes unsuitable. In the experiments presented in the next section we will show that the number of correct links obtained by an intruder decreases when more attributes are known.

Rank swapping p-distribution applied to an attribute $attr_j$ of an original microdata file X can be defined as follows: first, the table (microdata file) is sorted in increasing order of the values x_{ij} of the considered attribute $attr_j$. For simplicity, assume that the records are already sorted, that is $x_{ij} \leqslant x_{\ell j}$ for all $1 \leqslant i < \ell \leqslant n$. Then, for each value x_{ij}, a random value r is computed by using the $N(0.5 \cdot p, 0.5 \cdot p)$ normal distribution, and the values x_{ij} and $x_{\ell j}$ are swapped, where $\ell = i + r$. Finally, the sorting step is undone.

The negative effect produced by the swapping between two far values is that the information loss of the protected files increases. We will show in Section 5 that the increment of information loss is compensated by a reduction of the disclosure risk, and therefore the scores obtained by this method are lower than the standard rank swapping.

In Section 6 we present a possible specific record linkage for this new rank swapping p-distribution method. The experiments show that the performance of this specific record linkage is similar to distance based record linkage.

In the rest of this paper, we will use *rs α-d* to denote the application of rank swapping p-distribution with $p = \alpha$.

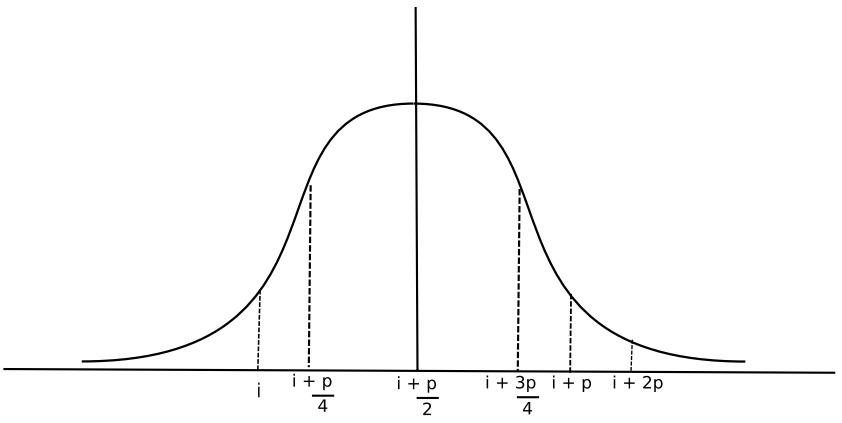


Fig. 2. Graphic representation of the p-distribution swap interval.

4.2. Rank swapping p-buckets

The rank swapping p-buckets method pursues the same goal than rank swapping p-distribution. For this reason it also substitutes the close swap interval of the rank swapping with an unlimited interval, but using now a different technique.

The idea of this method is to split the sorted original values of an attribute in several buckets. First, a probability function is used to choose a bucket for each value. Once the bucket is selected, the method works identically to the standard rank swapping, by using this bucket as the closed swap interval. Again, every original record will have some non-zero probability of being the correct link of every protected record. For this reason, the RS-RL method will be less effective here than when it is applied to the standard rank swapping.

Rank swapping p-bucket applied to an attribute $attr_j$ of an original microdata file X can be defined as follows: first, the table is sorted in increasing order of the values x_{ij} of the considered attribute $attr_j$. For simplicity, assume that the records are already sorted, that is $x_{ij} \leqslant x_{\ell j}$ for all $1 \leqslant i < \ell \leqslant n$. The sorted values $\{x_{ij} | 1 \leqslant i \leqslant n\}$ are split into p buckets $B_1, \ldots, B_p$. For each value x_{ij}, which belongs to some bucket B_r, a bucket B_s is chosen, according to the probability distribution

$$\Pr[B_s \text{ is chosen}] = \frac{1}{2^{s-r+1}}.$$

Then, a value $x_{\ell j}$ is randomly and uniformly chosen in the selected bucket B_s, and the values x_{ij} and $x_{\ell j}$ are swapped. Note that, if the same bucket $B_s = B_r$ is chosen, the condition $\ell > i$ must be imposed.

Intuitively, closer buckets to the original value are selected with higher probability than the far-off buckets and, therefore, the information loss of the protected microdata file is under control. Note that many other probability distributions could be used to define similar variants of rank swapping.

In the rest of the paper, *rs* α-*b* will be used to denote a rank swapping p-bucket with $p = \alpha$.

5. Experiments

As stated above, we have introduced the two variants of rank swapping to mitigate the effect of the specific record linkage method RS-RL. Of course, this is at the cost of increasing information loss, because some values may be swapped with values which are quite far. Our feeling was that this increment of the information loss had to be less significant than the saving in disclosure risk, and, therefore, the new methods would obtain better general scores than standard rank swapping. Experiments described in this section confirm our feelings.

In short, the section describes the analysis of disclosure risk using our specific record linkage method and the comparison of the new rank swapping methods against standard ones. We start reviewing the data files used in the experiments.

5.1. Data

In our experiments, we have considered the two reference microdata files proposed in the CASC project [4]. The first one, called Census, has been extracted using the Data Extraction System (DES) from the US Census Bureau [8]. This dataset contains 1080 records consisting of 13 attributes (i.e., 14 040 values are to be protected). The second one, called EIA, was extracted from the U.S. Energy Information Authority [16]. It contains 4092 records consisting of 10 attributes (i.e., 40 920 values to be protected).

5.2. Disclosure risk analysis for rank swapping

In order to evaluate the specific record linkage method (RS-RL) introduced in Section 3, we have protected the Census and EIA datasets by using different parameterizations of standard rank swapping ($p = 2 \ldots 20$), rank swapping p-distribution ($p = 2 \ldots 20$) and rank swapping p-buckets ($p = 75$, 50, 35, 30, 25, 20, 15,10, 5). For each protected dataset, we have computed its disclosure risk using RS-RL. At this point, in order to study the worst case scenario, the parameter p is assumed to be known.

It is clear that the rank swapping p-distribution and the rank swapping p-buckets have an advantage with respect to standard rank swapping when RS-RL is used. For this reason, disclosure risk measures have been computed using a larger parameter. In particular, we used $2 \cdot p$ for the rank swapping p-distribution, and $p = 2 \cdot \textit{Bucket Size}$ for the rank swapping p-buckets. For standard rank swapping, the parameter p was used.

Fig. 3 shows, in a graphic way, the number of correct links that RS-RL obtains, when applied to standard rank swapping, for both datasets, when different numbers of attributes are assumed to be known by the intruder (from one to all). Detailed results are presented in Tables A.1 and A.2 in Appendix A. It is easy to observe that the more attributes are known by the intruder, the more records are linked. Figures also show that for the five less protected datasets from Census, an intruder links more than 70% of the records when only half of the attributes are known. Another interesting result with the Census dataset is that the intruder is always able to link more than 50% of the records if he knows all the attributes. These results confirm with real data that when data providers are interested in releasing a dataset with low disclosure risk, standard rank swapping is not the most suitable method.

Similar results are obtained for the EIA dataset. For the three less protected datasets, the intruder is able to link more than 50% of records when all the attributes are known. In EIA, the results of our specific record linkage are not so good because (i) the EIA dataset has four times more records than Census dataset, (ii) the EIA dataset has less attributes than Census.

Fig. 4 presents the results of RS-RL applied to the files protected using rank swapping p-distribution. Detailed results are given in Appendix A. The chart lines show that an intruder can make no use of knowing all the attributes when data is protected using rank swapping p-distribution. Therefore, the rank swapping record linkage becomes unsuitable now. This is so, because when the intruder knows many attributes, he is

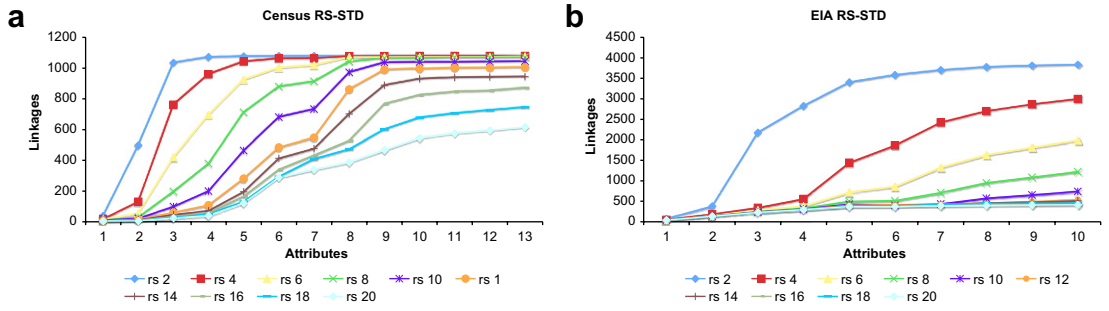


Fig. 3. Graphic representation of the results obtained by the rank swapping disclosure risk measure applied to the Census dataset (a) and EIA dataset (b), protected with standard rank swapping.

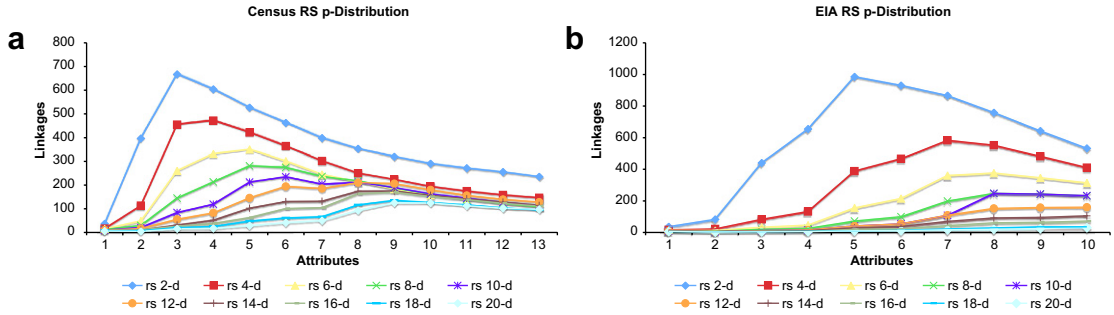


Fig. 4. Graphic representation of the results obtained by the rank swapping disclosure risk measure applied to the Census dataset (a) and EIA dataset (b), protected with rank swapping p-distribution.

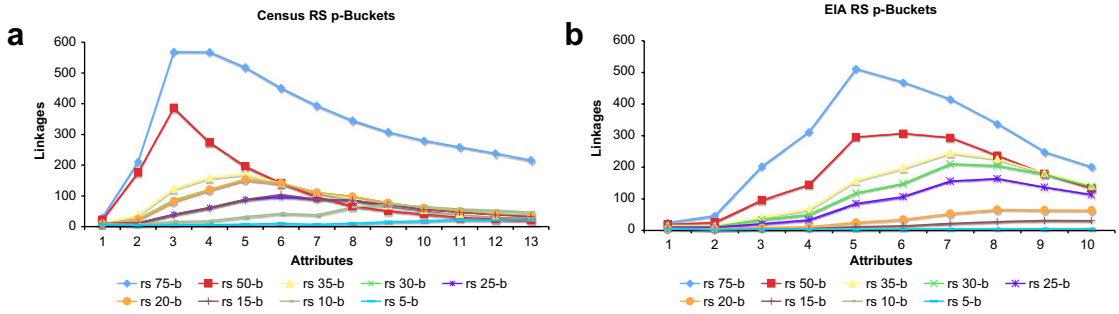


Fig. 5. Graphic representation of the results obtained by the rank swapping disclosure risk measure applied to the Census dataset (a) and EIA dataset (b), protected with rank swapping *p*-buckets.

forced to consider all the records, and not a small subset, to take advantage of his knowledge. For the Census dataset, only in the less protected parameterization ($p = 2$) the intruder is able to link more than 50% of the records, and this happens when he knows three or four attributes: the knowledge of more attributes is not useful in this case. The rest of the cases are protected enough to avoid a large number of correct linkages for both the Census and EIA datasets.

The results obtained for rank swapping *p*-buckets, which are shown in Fig. 5 and detailed in Appendix A, are similar to the results of rank swapping *p*-distribution. Only in the less protected case of Census dataset the intruder is able to link more than 50% of the records, the remainder configurations are protected enough to avoid a large number of correct linkages. As it happens with the rank swapping *p*-distribution, the knowledge of all attributes is not useful for an intruder when using rank swapping record linkage.

5.3. New rank swapping methods vs. standard rank swapping

In this section, we compare the overall behavior of the new methods with the standard one. We study the effects of the introduced modifications in information loss and standard disclosure risk measures.

The comparison is based on the score defined in Section 2.2. We have modified the score so that the disclosure risk measure takes into account the new rank swapping record linkage. Formally, we use the following disclosure risk measure $DR = 0.166 \cdot RSLD + 0.166 \cdot DLD + 0.166 \cdot PLD + 0.5 \cdot ID$ instead of $DR = 0.25 \cdot DLD + 0.25 \cdot PLD + 0.5 \cdot ID$. Here, *RSLD* stands for Rank Swapping Linkage Disclosure, the average percentage of correctly linked records using rank swapping record linkage (RS-RL). Information loss measures are not changed, and, thus, they are computed using the standard measures presented in Section 2.2.

Table 4 presents the scores as well as the original values before their aggregation, for both Census and EIA datasets protected using standard rank swapping. We can observe that the largest disclosure risk measure in

Table 4
Score calculation for standard rank swapping

	Census						EIA					
	IL	*RSLD*	*DLD*	*PLD*	*ID*	Score	*IL*	*RSLD*	*DLD*	*PLD*	*ID*	Score
rs 2	3.89	77.73	73.52	71.28	93.98	42.63	4.24	43.27	21.71	16.85	93.10	28.06
rs 4	6.54	66.65	58.40	42.92	83.09	36.67	9.67	12.54	10.61	4.79	82.09	21.89
rs 6	10.57	54.65	43.76	22.49	72.12	31.93	14.63	7.69	7.40	2.03	72.21	21.42
rs 8	16.54	41.28	32.13	11.74	62.11	29.16	18.71	6.12	5.98	1.12	63.90	21.61
rs 10	20.18	29.21	23.64	6.03	53.28	26.31	22.87	5.60	5.19	0.69	57.09	22.37
rs 12	23.46	19.87	18.96	3.46	47.17	24.77	26.60	5.39	4.87	0.51	51.64	23.25
rs 14	28.93	16.14	15.63	2.06	43.39	25.86	29.42	5.28	4.55	0.32	47.49	23.91
rs 16	35.16	13.81	13.59	1.29	40.78	27.97	32.38	5.19	4.54	0.23	44.19	24.82
rs 18	32.52	12.21	11.50	0.83	38.90	25.81	34.22	5.20	4.54	0.22	41.42	25.28
rs 20	35.12	10.88	10.87	0.59	37.33	26.55	36.27	5.15	4.36	0.18	38.97	25.87

IL stands for Information Loss, *RSLD* stands for Rank Swapping Linkage Disclosure, *DLD* stands for Distance Linkage Disclosure, *PLD* stands for Probability Linkage Disclosure and *ID* stands for Interval Disclosure.

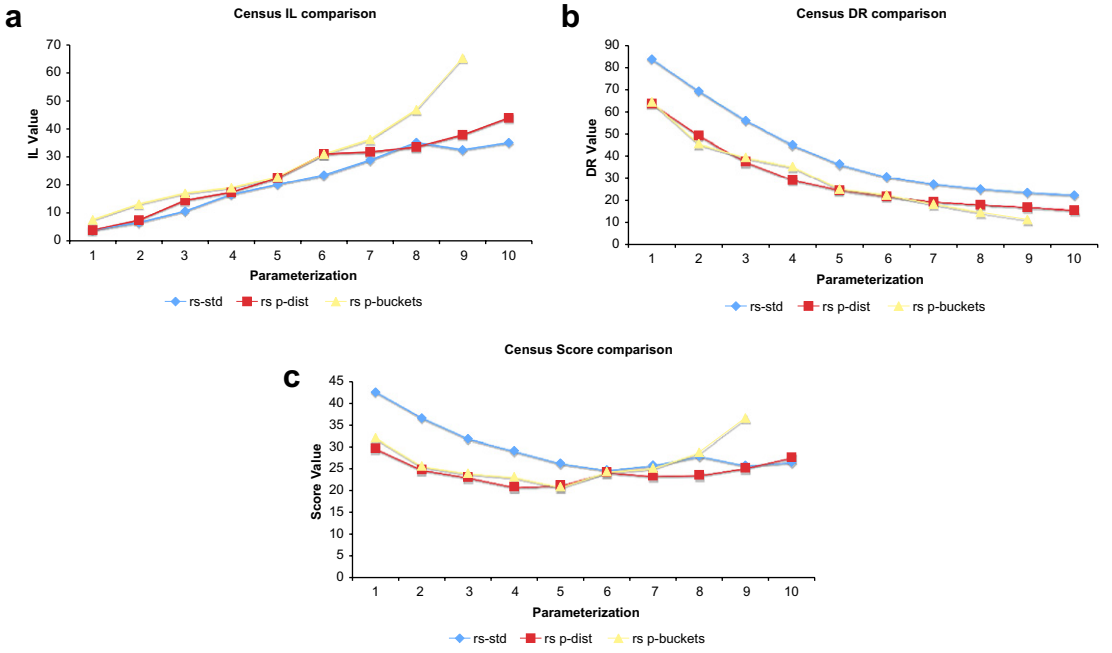


Fig. 6. Graphic representation of the Information loss (a), Disclosure Risk (b) and Score (c) values for the three rank swapping methods when Census dataset is protected.

all cases is *RSLD*. Therefore, it is clear that the new method increases the risk with respect to standard ones for the standard rank swapping. Another interesting result is that *PLD* is always lower than *RSLD* and *DLD*. This is because all the values in all the attributes are swapped, and therefore the coincidence vectors for the correct links are always equal to zero, unless the swapped positions have the same value (which is possible only if the attributes have repeated values). It is possible to observe in Figs. 6b and 7b that the standard rank

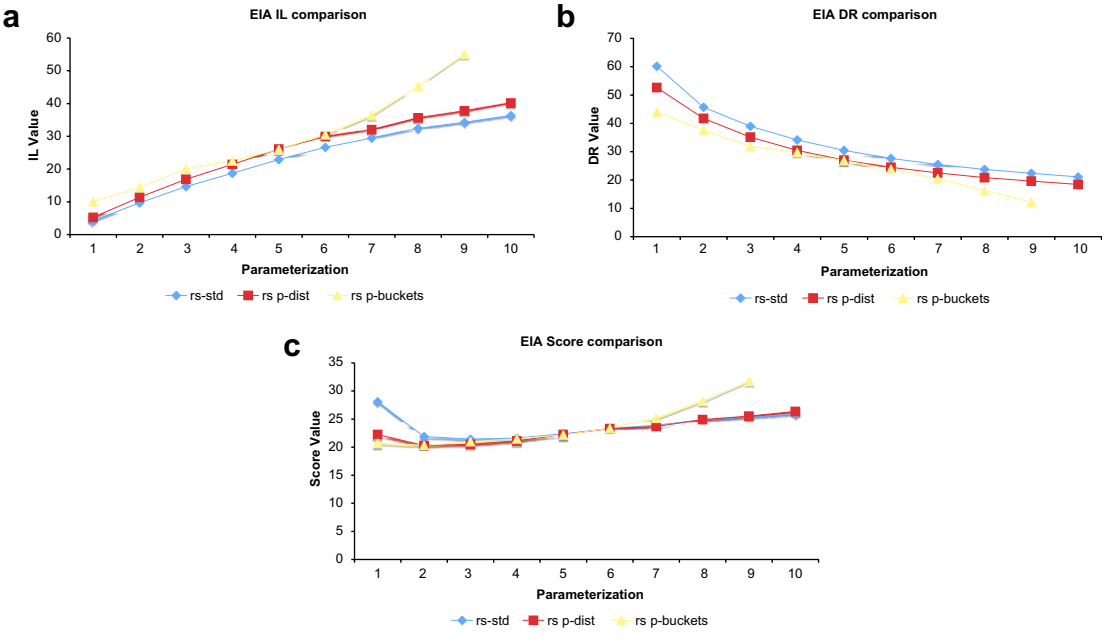


Fig. 7. Graphic representation of the Information loss (a), Disclosure Risk (b) and Score (c) values for the three rank swapping methods when EIA dataset is protected.

Table 5
Score calculation for rank swapping *p*-distribution

	Census						EIA					
	IL	*RSLD*	*DLD*	*PLD*	*ID*	Score	*IL*	*RSLD*	*DLD*	*PLD*	*ID*	Score
rs 2-d	3.80	40.90	71.63	1.38	89.81	29.70	5.24	10.74	18.78	13.94	90.76	22.26
rs 4-d	7.42	28.34	53.10	0.68	71.48	24.83	11.33	3.13	9.12	3.90	78.07	20.28
rs 6-d	14.44	20.51	37.57	0.62	55.58	23.08	16.83	1.27	6.41	1.61	67.20	20.50
rs 8-d	17.31	15.74	25.98	0.54	44.44	20.90	21.40	0.66	5.40	0.93	58.55	21.09
rs 10-d	22.49	11.69	19.42	0.46	38.60	21.28	26.11	0.39	4.81	0.55	52.05	22.24
rs 12-d	31.04	9.10	16.17	0.41	35.05	24.26	29.89	0.39	4.62	0.42	47.11	23.28
rs 14-d	31.80	6.12	13.16	0.38	32.19	23.41	32.01	0.30	4.51	0.29	43.30	23.67
rs 16-d	33.53	4.63	11.77	0.32	30.42	23.61	35.59	0.20	4.60	0.22	39.95	24.89
rs 18-d	37.89	3.10	11.12	0.35	28.98	25.25	37.69	0.10	4.69	0.18	37.52	25.52
rs 20-d	43.92	2.15	9.59	0.31	27.05	27.65	40.12	0.10	4.57	0.11	35.25	26.34

IL stands for Information Loss, *RSLD* stands for Rank Swapping Linkage Disclosure, *DLD* stands for Distance Linkage Disclosure, *PLD* stands for Probability Linkage Disclosure and *ID* stands for Interval Disclosure.

Table 6
Score calculation for rank swapping *p*-buckets

	Census						EIA					
	IL	*RSLD*	*DLD*	*PLD*	*ID*	Score	*IL*	*RSLD*	*DLD*	*PLD*	*ID*	Score
rs 75-b	7.51	36.24	58.08	44.50	83.02	32.18	10.03	5.35	10.90	6.50	80.37	20.75
rs 50-b	13.14	17.09	42.71	27.21	62.60	25.68	14.35	2.83	7.53	3.17	70.70	20.32
rs 35-b	17.03	9.96	29.93	13.36	61.29	24.00	20.03	1.46	5.53	1.56	60.80	20.98
rs 30-b	18.96	8.56	24.86	9.81	56.37	23.19	22.49	1.11	5.42	1.16	56.92	21.47
rs 25-b	22.79	5.26	20.18	0.40	41.60	20.92	25.98	0.78	4.90	0.79	51.61	22.19
rs 20-b	30.97	8.51	15.86	0.42	36.61	24.36	30.46	0.28	4.65	0.54	45.88	23.35
rs 15-b	36.29	5.36	11.59	0.44	30.82	25.18	36.32	0.14	4.82	0.33	39.12	25.13
rs 10-b	46.91	2.19	8.53	0.40	25.40	28.80	45.17	0.09	4.74	0.17	31.04	28.18
rs 5-b	65.29	0.63	6.54	0.31	20.18	36.68	54.96	0.08	4.55	0.08	22.87	31.69

IL stands for Information Loss, *RSLD* stands for Rank Swapping Linkage Disclosure, *DLD* stands for Distance Linkage Disclosure, *PLD* stands for Probability Linkage Disclosure and *ID* stands for Interval Disclosure.

swapping has the largest disclosure risk when the different parameterizations of the three methods are compared. This effect is much more clearer in Fig. 6b (corresponding to the Census dataset) than in Fig. 7b (corresponding to the EIA dataset).

Tables 5 and 6 show the score values for rank swapping *p*-distribution and rank swapping *p*-buckets respectively. In both cases the largest disclosure risk measure is *DLD*. Therefore, when an intruder is interested in linking the original dataset with the one protected using any of the two methods, the best way is just considering all possible links.

In Section 6 we discuss a possible record linkage method specifically designed for rank swapping *p*-distribution and *p*-buckets. Since the obtained results are essentially the same as with distance based record linkage we have not considered this method to compute the disclosure risk and the score.

Interval disclosure for standard rank swapping is higher than the one for the two new rank swapping methods. This is so because the rank swapping *p*-distribution and the *p*-buckets may do swaps between two far values avoiding the interval disclosure. This is not the case in standard rank swapping.

In general, when we compare the same parameterizations for the three different rank swapping methods (*e.g.* rs 2, rs 2-d and rs 75-b), information loss is higher for the rank swapping *p*-distribution and the rank swapping *p*-buckets. Nevertheless, disclosure risk is higher for standard rank swapping (in some cases more than 15%) for these parameterizations. In Figs. 6a and 7a we can observe that some cases of standard rank swapping have lower information loss than the other rank swapping versions. However, these differences are rather small in most of the cases. See, for example, the IL in the EIA dataset for the fourth

parameterization of the three rank swapping methods (rs 8, rs 8-d and rs 30-b). It is clear that the values for IL are rather similar.

In relation to the overall scores, the best scores obtained for the standard rank swapping (see Table 4) are 24.77 for Census and 21.42 for EIA. In contrast to that, the best scores obtained for the rank swapping p-distribution (see Table 5) are between 20.90 and 21.30 for the Census dataset and between 20.20 and 21.00 for the EIA dataset. The best score presented in Table 6 for the rank swapping p-buckets is 20.92 for the Census dataset and 20.32 for the EIA dataset. So, the new rank swapping methods lead to better scores and, thus, the trade-off between information loss and disclosure risk benefits the rank swapping methods introduced in this paper. Even though the new methods have a small increment in the information loss, they gain an important reduction in disclosure risk. This effect is illustrated in Figs. 6c and 7c where in most of the cases the standard rank swapping has the largest score.

6. Specific record linkage methods for the new variations of rank swapping

In Section 2.3 we have shown that standard rank swapping can be attacked using a specific record linkage method. These results have motivated the introduction of two variances of rank swapping, variations that cannot be attacked using the specific methods. Nevertheless, it is worth considering at this point whether other ad hoc attacks might be developed for these new methods. In this section we present another specific record linkage method specially designed for rank swapping p-distribution and p-buckets. We show that this new method, when applied to the same datasets used in Section 5 leads to similar results to the ones obtained with standard distance based record linkage method.

The difficulty of applying record linkage to data protected by rank swapping p-distribution and p-bucket is that the intruder cannot limit the swap interval. In other words, it is always possible that the correct linkage is not included in the intersection set $B(x_i) = \cap_{1 \leqslant j \leqslant c} B(x_{ij})$. Therefore, the use of $B(x_i)$ is sometimes useless.

To avoid this, but reducing at the same time the complexity of record linkage, we can consider proper subsets $B(x_{ij})$, and then compute their union. Naturally, with a large probability, the correct link will be in $B(x_{ij})$ for at least one of the attributes j. In fact, the record will belong to most of the $B(x_{ij})$. Due to this, the record will be also in the union of the $B(x_{ij})$. That is, in $B(x_i) = \cup_{1 \leqslant j \leqslant c} B(x_{ij})$.

We will compute an *annotated* union that takes into account the number of sets $B(x_{ij})$ where each record is stored. Then, we assume that the intruder only compares the original record with the protected records which have the maximum cardinality in the annotated union set (i.e., records stored in the maximum number of $B(x_{ij})$ sets). In this approach the intruder is minimizing the far swaps in the protected attributes. So, the intruder is exploiting his knowledge of the rank swapping p-distribution and p-buckets because he knows that the probability of a far swap is near to 0.

As we did in Section 3, we suppose here that the parameter p is known. We have applied this variation of the rank swapping record linkage method to the rank swapping p-distribution. In order to apply the method in the worse case scenario, we also suppose that the intruder has the maximum number of attributes, i.e., the intruder knows the half of the protected attributes (this assumption is the same than in Section 2.2). At this point, the intruder needs to decide the size of the intervals $B(x_{ij})$. From the cumulative normal distribution, the intruder can deduce the confidence of the selected swap intervals. For example, if the swap intervals are $[i - 0.5p, i + 0.5p]$ the probability that the correct linkage is inside one of these intervals is equal to 68%. Instead, if the swap intervals are $[i - p, i + p]$; then, the same probability increases to 96%. Note that when the union is computed the probability to find the correct link inside increases with respect to the number of attributes known.

Table 7 shows the results for the distance based record linkage and for the rank swapping record linkage with two different swap intervals. The table also includes the results obtained by the DBRL for the same files and number of attributes.

The results show that, in general, when we use the interval $[i - p, i + p]$ the new method lead to results similar to the ones by DBRL. There are only two cases were we can find a significant improvement, they correspond to the application of the method when $p = 2$ and $p = 4$ and using the EIA database. RS-RL finds 1443

Table 7
Average linkage values for rank swapping p-distribution

	Census			EIA		
	DBRL	RS-RL	RS-RL	DBRL	RS-RL	RS-RL
Interval		$\left[i-\frac{p}{2}, i+\frac{p}{2}\right]$	$[x_i-p, x_i+p]$		$\left[i-\frac{p}{2}, x_i+\frac{p}{2}\right]$	$[i-p, i+p]$
rs 2-d	1048.2	752.8	1054.4	1267.2	771.5	1443.8
rs 4-d	938.9	461.0	907.4	330.1	256.0	419.2
rs 6-d	723.6	285.3	686.6	130.4	118.4	159.4
rs 8-d	502.5	190.0	487.1	60.4	58.9	72.9
rs 10-d	335.1	136.2	347.2	36.0	37.2	41.1
rs 12-d	219.5	96.4	241.7	22.4	23.9	28.4
rs 14-d	145.2	70.4	161.7	16.8	17.0	19.9
rs 16-d	95.7	48.9	109.7	11.5	11.3	13.1
rs 18-d	69.6	42.5	76.5	8.9	11.2	9.9
rs 20-d	47.8	33.7	54.3	7.0	8.3	9.4

DBRL stands for Distance Based Record Linkage, RS-RL stands for Rank Swapping Record Linkage.

records while DBRL finds 1267, and RS-RL finds 419 where DBRL finds 330. Nevertheless, these cases are not a real threat to the protection method because they already correspond to the cases with a lower protection and high risk (note that in the case of the Census database, that is a smaller dataset, almost all records are re-identified by DBRL). Taking this into account, we have that the influence of the results shown here in the score would not be significant.

The table also shows the results for the interval $[i-p/2, i+p/2]$. This smaller interval reduces more the search than the other one. However, the results obtained by the new record linkage are much worse than the ones obtained by the DBRL.

Summing up, we have that the ad hoc record linkage method has a similar behavior than DBRL, except for a few cases that cannot be considered a great threat to the protection mechanism.

7. Conclusions and future work

In this paper, we have presented a new type of record linkage specially designed to measure the disclosure risk of the rank swapping protection method, and we have shown that this new method outperforms the standard record linkage measures. We have also presented two modifications of rank swapping, called rank swapping p-distribution and rank swapping p-buckets, which make unsuitable the new rank swapping record linkage measure. We have shown that these two new rank swapping methods have lower disclosure risk and better scores than the standard one. Also, we have presented a possible specific attack for the new methods. The analysis performed show that the new protection methods are resistant to our specific ad hoc attacks.

As future work, we plan to further study protection methods where ad hoc disclosure risk measures cannot be applied. Another interesting research line is to study modifications of the rank swapping record linkage (RS-RL) to improve linkage results in very large databases.

Acknowledgements

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) is acknowledged. Jordi Nin wants to thank the Spanish Council for Scientific Research (CSIC) for his I3P grant.

Appendix A. A Linkage results

See Table A.1–A.6.

Table A.1
Number of correctly linked records when RS-RL is applied to Census dataset, protected with standard rank swapping

	rs 2	rs 4	rs 6	rs 8	rs 10	rs 12	rs 14	rs 16	rs 18	rs 20
1	38.4	18.0	16.8	10.8	10.8	6.4	6.8	5.2	4.0	4.0
2	497.0	130.2	54.2	29.2	21.8	15.4	13.0	10.4	7.0	6.0
3	1034.2	761.2	420.6	197.2	99.0	60.2	45.2	32.4	28.8	24.2
4	1071.8	959.6	694.2	378.6	199.4	107.2	71.6	58.0	49.4	39.6
5	1076.8	1042.0	925.2	711.6	463.2	281.4	195.0	165.6	131.6	121.2
6	1079.0	1063.2	1001.6	879.2	681.0	484.2	413.0	340.4	293.6	287.4
7	1079.2	1064.2	1018.6	913.8	733.0	547.4	475.4	432.4	408.6	339.2
8	1079.2	1077.6	1071.8	1042.6	972.0	861.6	701.6	528.6	472.4	386.4
9	1079.6	1077.6	1071.4	1065.6	1036.6	988.8	888.0	766.6	602.0	466.0
10	1079.6	1078.2	1072.1	1066.6	1039.2	996.6	930.8	824.8	677.4	544.0
11	1079.6	1079.1	1073.4	1069.2	1039.4	1001.0	939.4	846.8	706.2	574.0
12	1079.6	1079.1	1073.4	1069.6	1041.2	1002.0	942.0	853.4	726.4	593.8
13	1079.6	1079.1	1076.7	1070.3	1044.8	1004.4	944.2	871.0	745.6	615.4

The first columns shows the number of known attributes.

Table A.2
Number of correctly linked records when RS-RL is applied to EIA dataset, protected with standard rank swapping

	rs 2	rs 4	rs 6	rs 8	rs 10	rs 12	rs 14	rs 16	rs 18	rs 20
1	70.3	46.1	43.6	42.4	40.0	37.6	37.6	35.1	37.5	36.3
2	378.4	183.4	145.1	139.8	135.7	135.4	134.6	133.2	133.0	132.3
3	2174.8	338.8	284.1	246.8	236.1	229.1	226.1	224.7	224.3	223.5
4	2827.1	557.0	380.5	327.6	310.2	301.7	298.3	296.0	294.7	294.9
5	3402.9	1441.3	720.9	496.1	423.7	398.7	384.1	373.0	374.0	367.7
6	3582.9	1859.4	856.3	512.4	400.3	415.9	397.0	380.2	378.8	371.3
7	3699.8	2420.6	1325.3	709.6	431.4	423.9	410.1	393.1	424.5	391.6
8	3778.6	2699.0	1631.7	947.3	572.8	448.8	458.3	411.5	445.6	401.3
9	3810.1	2862.2	1808.5	1081.8	654.6	492.2	479.4	420.9	451.9	409.3
10	3831.5	2996.8	1986.3	1221.5	741.5	539.2	507.9	432.6	455.4	411.8

The first columns shows the number of known attributes.

Table A.3
Number of correctly linked records when RS-RL is applied to Census dataset, protected with rank swapping *p*-distribution

	rs 2-d	rs 4-d	rs 6-d	rs 8-d	rs 10-d	rs 12-d	rs 14-d	rs 16-d	rs 18-d	rs 20-d
1	37.0	16.4	14.8	12.4	9.6	9.2	7.6	6.8	8.0	5.6
2	396.4	112.8	48.6	29.4	20.6	16.2	9.8	11.0	7.2	4.4
3	667.4	454.6	260.4	144.8	84.0	55.2	30.6	26.8	19.8	15.6
4	604.0	471.4	331.4	213.4	119.2	82.6	51.4	37.4	25.0	17.2
5	527.0	421.4	350.4	280.4	212.6	145.2	102.0	62.8	47.4	28.6
6	461.8	364.6	300.4	274.2	234.4	194.2	130.2	100.8	60.8	41.4
7	398.2	301.0	244.8	235.6	203.4	185.4	131.4	104.2	65.8	50.0
8	353.0	249.8	213.4	217.8	211.4	210.0	173.6	163.0	116.0	92.6
9	319.6	223.0	183.2	186.6	189.0	205.8	174.4	169.4	135.4	124.6
10	290.4	194.2	157.2	161.2	162.4	179.2	156.8	151.2	125.6	124.2
11	271.0	174.2	141.0	139.8	143.8	156.0	141.6	138.0	116.4	112.6
12	254.8	157.6	126.2	121.8	126.8	138.6	127.6	123.0	103.6	102.2
13	235.4	146.0	111.6	107.4	113.2	127.8	113.8	110.8	95.0	97.0

The first columns shows the number of known attributes.

Table A.4
Number of correctly linked records when RS-RL is applied to EIA dataset, protected with rank swapping *p*-distribution

	rs 2-d	rs 4-d	rs 6-d	rs 8-d	rs 10-d	rs 12-d	rs 14-d	rs 16-d	rs 18-d	rs 20-d
1	36.4	16.0	13.6	10.6	8.6	8.6	8.1	6.6	5.9	5.1
2	82.5	22.3	11.9	6.9	4.0	4.0	5.3	3.3	2.9	2.2
3	439.3	82.8	33.2	20.3	11.9	11.4	8.4	7.7	4.8	3.4
4	654.0	132.2	46.3	25.7	15.5	15.5	11.1	6.2	6.3	4.4
5	985.1	388.1	155.6	72.0	40.4	40.8	28.3	19.9	9.9	9.4
6	929.5	465.8	214.6	99.7	54.5	54.5	37.8	23.6	13.3	11.5
7	865.8	583.0	360.3	198.1	108.4	109.6	69.8	42.6	22.2	16.7
8	757.9	552.3	376.3	246.9	246.9	151.6	91.7	60.3	29.2	20.2
9	642.0	481.7	346.4	242.6	242.6	157.3	94.7	64.2	34.8	24.4
10	532.4	409.6	314.6	231.9	232.2	159.9	104.7	71.1	35.2	29.5

The first columns shows the number of known attributes.

Table A.5
Number of correctly linked records when RS-RL is applied to Census dataset, protected with rank swapping *p*-buckets

	rs 75-b	rs 50-b	rs 35-b	rs 30-b	rs 25-b	rs 20-b	rs 15-b	rs 10-b	rs 5-b
1	30.0	22.1	8.8	9.6	10.0	9.6	10.0	9.2	6.4
2	210.8	176.2	37.8	24.6	11.8	24.6	11.8	7.4	4.0
3	568.2	386.2	123.6	86.0	39.4	82.6	39.4	16.2	5.4
4	567.8	273.4	159.8	120.8	61.6	120.8	61.6	18.6	6.0
5	518.2	196.0	171.8	153.8	88.2	153.8	88.2	32.4	8.0
6	450.8	141.4	141.2	141.2	96.8	141.2	104.4	43.0	10.4
7	394.0	97.0	109.8	110.8	89.6	110.8	89.6	38.8	7.2
8	345.8	67.0	91.2	98.0	85.8	98.0	85.8	63.2	11.0
9	307.4	51.8	69.4	77.6	70.4	77.6	70.4	72.0	16.0
10	279.8	40.6	54.6	61.4	55.0	61.4	55.0	65.2	18.6
11	259.0	30.0	42.2	49.6	46.4	49.6	46.4	57.2	22.6
12	238.8	24.6	34.2	40.6	38.2	40.6	38.2	53.4	21.0
13	216.4	20.2	31.4	35.8	32.4	35.8	32.8	46.8	21.6

The first columns shows the number of known attributes.

Table A.6
Number of correctly linked records when RS-RL is applied to EIA dataset, protected with rank swapping *p*-buckets

	rs 75-b	rs 50-b	rs 35-b	rs 30-b	rs 25-b	rs 20-b	rs 15-b	rs 10-b	rs 5-b
1	24.7	18.5	15.9	12.8	10.6	6.7	6.4	4.8	4.1
2	46.5	25.4	14.9	11.9	9.7	5.5	2.1	2.3	1.9
3	202.7	95.7	42.0	33.9	20.7	8.8	4.4	3.8	3.2
4	310.8	144.4	67.0	50.1	32.8	11.6	5.4	3.8	3.2
5	510.7	294.9	157.9	118.3	84.8	25.2	11.3	3.7	3.0
6	467.5	306.0	197.8	148.1	106.8	34.9	14.3	4.7	3.5
7	414.9	292.6	246.0	210.2	155.9	53.2	21.9	5.7	3.8
8	336.5	236.6	226.2	205.2	163.7	66.3	27.7	5.9	3.9
9	247.8	179.0	181.0	176.6	136.9	64.7	31.6	6.1	3.9
10	201.2	132.9	141.1	138.6	112.9	63.9	31.1	6.3	4.1

The first columns shows the number of known attributes.

References

[1] Mark S. Ackerman, Lorrie Faith Cranor, Joseph Reagle, Privacy in e-commerce: examining user scenarios and privacy preferences, in: EC'99: Proceedings of the 1st ACM Conference on Electronic Commerce, ACM Press, 1999, ISBN 1-58113-176-3, pp. 1–8.

[2] N.R. Adam, J.C. Wortmann, Security-control for statistical databases: a comparative study, ACM Computing Surveys 21 (1989) 515–556.

[3] E. Bertino, I. Nai Fovino, L. Parasiliti Provenza, A rramework for evaluating privacy preserving data mining algorithms, Data Mining and Knowledge Discovery 11 (2) (2005) 121–154.

[4] R. Brand, J. Domingo-Ferrer, J.M. Mateo-Sanz, Reference datasets to test and compare SDC methods for protection of numerical microdata, Manuscript for [6], 2002.
[5] J. Burridge, Information preserving statistical obfuscation, Statistics and Computing 13 (2003) 321–327.
[6] CASC: Computational Aspects of Statistical Confidentiality, European Project IST-2000-25069. <http://neon.vb.cbs.nl/casc>.
[7] T. Dalenius, S.P. Reiss, Data-swapping: a technique for disclosure control, Journal of Statistical Planning and Inference 6 (1982) 7385.
[8] Data Extraction System, US Census Bureau. <http://www.census.gov/>.
[9] J. Domingo-Ferrer, V. Torra, Disclosure control methods and information loss for microdata, 2001, pp. 91–110 of [12].
[10] J. Domingo-Ferrer, V. Torra, A quantitative comparison of disclosure control methods for microdata, 2001, pp. 111–133 of [12].
[11] J. Domingo-Ferrer, J.M. Mateo-Sanz, Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering 14 (2002) 189–201.
[12] in P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, 2001.
[13] G. Duncan, S. Fienberg, R. Krishnam, R. Padman, S. Roehrig, Disclosure limitation methods and information loss for tabular data, 2001, pp. 135–166 of [12].
[14] G. Duncan, S. Keller-McNulty, S. Stokes, Disclosure risk vs. data utility: the R-U confidentiality map, Technical Report Number 121 of National Institute of Statistical Sciences, 2001. <www.niss.org>.
[15] G. Duncan, S. Keller-McNulty, S. Stokes, Database security and confidentiality: examining disclosure risk vs. data utility through the R-U confidentiality map, Technical Report Number 142 of National Institute of Statistical Sciences, 2004. <www.niss.org>.
[16] US Energy Information Authority. <http://www.eia.doe.gov/cneaf/electricity/page/eia826.html>.
[17] Statistical Office of the European Communities (EUROSTAT), <http://epp.eurostat.ec.europa.eu>.
[18] S.E. Fienberg, S.E. McIntyre, Data swapping: variations on a theme by Dalenius and Reiss, in: Privacy in Statistical Databases 2004, Lecture Notes in Computer Science 3050 (2004) 14–29.
[19] M.A. Jaro, Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Association 84 (1989) 414–420.
[20] J.M. Mateo-Sanz, J. Domingo-Ferrer, F. Sebé, Probabilistic information loss measures in confidentiality protection of continuous microdata, Data Mining and Knowledge Discovery 11 (2) (2005) 181–193.
[21] R. Moore, Controlled data mapping techniques for masking public use microdata sets, US Bureau of the Census, 1996 (unpublished manuscript).
[22] D. Pagliuca, G. Seri, Some results of individual ranking method on the system of enterprise accounts annual survey, Esprit SDC Project, Delivrable MI-3/D2, 1999.
[23] F. Sebé, J. Domingo-Ferrer, J.M. Mateo-Sanz, V. Torra, Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets, in: Inference Control in Statistical Databases, Lecture Notes in Computer Science 2316 (2002) 187–196.
[24] V. Torra, J. Domingo-Ferrer, Record linkage methods for multidatabase data mining, in: Information Fusion in Data Mining, Springer, 2003, pp. 101–132.
[25] V. Torra, S. Miyamoto, Evaluating fuzzy clustering algorithms for microdata protection, in: Privacy in Statistical Databases 2004, Lecture Notes in Computer Science 3050 (2004) 175–186.
[26] V. Torra, J.M. Abowd, J. Domingo-Ferrer, Using Mahalanobis distance-based record linkage for disclosure risk assessment, in: Privacy in Statistical Databases 2006, Lecture Notes in Computer Science 4302 (2006) 233–242.
[27] D.W. Wang, C.J. Liau, T.S. Hsu, An epistemic framework for privacy protection in database linking, Data & Knowledge Engineering 61 (2007) 176–205.
[28] W.E. Winkler, Data cleaning methods, in: Proc. SIGKDD 2003, Washington, 2003.
[29] W.E. Winkler, Re-identification methods for masked microdata, in: Privacy in Statistical Databases 2004, Lecture Notes in Computer Science 3050 (2004) 216–230.
[30] W.E. Yancey, W.E. Winkler, R.H. Creecy, Disclosure risk assessment in perturbative microdata protection, in: Inference Control in Statistical Databases: From Theory to Practice, Lecture Notes in Computer Science 2316 (2002) 135–152.



Jordi Nin (Barcelona, Catalonia, 1979; BSc 2004, MSc 2007 both in Computer Science) is a Ph.D. student under the supervision of Dr. Vicenç Torra at the Artificial Intelligence Research Institute (IIIA-CSIC) near Barcelona, Catalonia, Spain. His fields of interests are privacy technologies, machine learning and soft computing tools. He has been involved in several research projects funded by the Catalan and Spanish governments and the European Community. The research done during his Ph.D. had been published in specialized journals and major conferences (around 15 papers).



Javier Herranz obtained his Ph.D. in Applied Mathematics in 2005, in the Technical University of Catalonia (UPC, Barcelona, Spain). After that, he spent 9 months in the École Polytechnique (France) and 9 months in the Centrum voor Wiskunde en Informatica (CWI, The Netherlands), as a post-doctoral researcher, granted with an ERCIM fellowship. From January 2007, he works as a post-doctoral researcher at IIIA-CSIC (Bellaterra, Spain). His research interests include the design and analysis of cryptographic protocols and the study of privacy preserving operations involving databases.



Vicenç Torra (Barcelona, Catalonia, 1968; BSc 1991, MSc 1992, Ph.D. 1994 all in Computer Science) is an Associate Research Professor at the Artificial Intelligence Research Institute (IIIA-CSIC) near Barcelona, Catalonia, Spain. His fields of interests are privacy technologies, decision making, information fusion and soft computing tools. He has had led several research projects funded by Catalan, Spanish, European and US sources. Research has been published in specialized journals (more than 60 articles in ISI JCR journals and LNCS) and major conferences. He has written a book on Artificial Intelligence (in Catalan) and another, with Y. Narukawa, entitled "Modeling Decisions: Information Fusion and Aggregation Operators" (Springer). He is a member of the editorial board of Fuzzy Sets and Systems, Journal of Privacy Technology, J. of Advanced Computational Intelligence and Intel. Informatics, and Mathware and Soft Computing. In 2004, he was co-founder with Y. Narukawa of the conference series Modeling Decisions for Artificial Intelligence (MDAI), whose successive editions he has co-chaired ever since. He was also a Co-chairman of the Privacy in Statistical Databases conference (PSD 2004).

ARTICLE IN PRESS

Data & Knowledge Engineering xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

On the disclosure risk of multivariate microaggregation

Jordi Nin *, Javier Herranz, Vicenç Torra

IIIA, Artificial Intelligence Research Institute, CSIC, Spanish National Research Council, Campus UAB s/n, 08193 Bellaterra, Catalonia, Spain

ARTICLE INFO

Article history:
Received 29 March 2008
Received in revised form 19 June 2008
Accepted 28 June 2008
Available online xxxx

Keywords:
Privacy in statistical databases
Disclosure risk
Record linkage
Microaggregation
Data projection

ABSTRACT

The aim of data protection methods is to protect a microdata file both minimizing the disclosure risk and preserving the data utility. Microaggregation is one of the most popular such methods among statistical agencies.

Record linkage is the standard mechanism used to measure the disclosure risk of a microdata protection method. However, only standard, and quite generic, record linkage methods are usually considered, whereas more specific record linkage techniques can be more appropriate to evaluate the disclosure risk of some protection methods.

In this paper we present a new record linkage technique, specific for microaggregation, which obtains more correct links than standard techniques. We have tested the new technique with MDAV microaggregation and two other microaggregation methods, based on projections, that we propose here for the first time. The direct consequence is that these microaggregation methods have a higher disclosure risk than believed up to now.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

In many situations, statistical agencies have to disseminate confidential data to decision makers, politicians, researchers or for any other public use. Nevertheless, this dissemination should be in accordance with laws and regulations which ensure the privacy of the respondents. In such scenarios, therefore, statistical agencies must take privacy into account very seriously.

One approach for ensuring privacy is the application of perturbative protection methods over the confidential data when it is released to third parties. At present, a large number of perturbative protection methods exist (see [1,6] for two good surveys on data protection methods).

Microaggregation has emerged as one of the most promising protection methods; for example, [10] shows that microaggregation is used by many statistical agencies for data anonymization. Microaggregation works as follows: given a dataset with V variables, it builds small clusters of at least k elements and replaces the original values by the centroid of the cluster to which the element belongs. A certain level of privacy is ensured because k elements have an identical protected value.

Microaggregation techniques can be classified as *univariate microaggregation* (when the number V of variables is 1) or *multivariate microaggregation* (where $V > 1$). Then, there are two possibilities for multivariate microaggregation. The first one is to apply a multivariate microaggregation technique to the whole set, resulting in protected clusters with at least k elements (or records) with exactly the same values in the V variables. Therefore, this method provides k-anonymity [22], a property which has been studied and analyzed in many other works, in different scenarios [2,23,24]. The second possibility is to split the set of variables in smaller blocks of v_i variables [18], where $1 \leqslant v_i < V$ (with $\sum v_i = V$), and then to apply in parallel some microaggregation technique (possibly univariate, if $v_i = 1$) to each block; in general, the final protected dataset in this case will not have k equal elements, because two elements which are in the same cluster for a block of variables are not necessarily in the same cluster for other blocks. As a consequence, this second possibility offers less privacy (k-anonymity is not ensured anymore), but the statistical utility of the protected dataset is higher.

* Corresponding author.
E-mail addresses: jnin@iiia.csic.es (J. Nin), jherranz@iiia.csic.es (J. Herranz), vtorra@iiia.csic.es (V. Torra).

0169-023X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2008.06.014

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

Formally, microaggregation is defined as an optimization problem with some constraints. The constraints are that clusters should have at least k elements. The objective function (to minimize) is the total sum of distances between all the elements to be protected and the centroids of the clusters where the elements are finally assigned. That is, the goal is to minimize the total sum of square errors (SSE). The rationale of this process is to make the protected dataset as similar as possible to the original data, but providing at the same time a sufficient level of privacy for the original data. In the case of univariate microaggregation, there exist polynomial-time methods to obtain an optimal solution of this optimization problem, for example the algorithm in [11] which works over a graph built from the sorted original data. The main drawback of univariate microaggregation, of course, is that it provides a bad level of privacy, due to its high disclosure risk [19].

With respect to the multivariate case, the problem of optimal microaggregation was proved to be NP-hard [20]. Due to this, different heuristic microaggregation schemes are found in the literature. We can distinguish here two approaches. The first one is to design a specific heuristic scheme; a well-know algorithm is MDAV (maximum distance to average vector) [13,8], an iterative algorithm. The second approach is to reduce the multivariate case to the univariate one, by using some method to project variables into a single one. In other words, records are sorted using a new projected value computed from the V variables to microaggregate; after that, an univariate microaggregation algorithm (for example, an optimal one) is applied to the sorted projected records. According to the projection method which is applied to the variables, two such microaggregation schemes have been suggested (but not detailed) in the literature: *PCP-microaggregation* (principal component projection) and *ZScore-microaggregation*. The first contribution of this paper is the first detailed description of these two methods; we also implement them and compare their performance against MDAV.

Since k-anonymity is ensured only in the case of microaggregation with a unique block of variables, the exact level of privacy of specific microaggregation schemes must be carefully studied. The usual technique to do so is to compute the *disclosure risk* of the methods, which considers in particular the number of original (confidential) records that an attacker can link with their protected partners. A correct link between the two datasets implies that the privacy of the respondent is broken. The use of such record linkage methods [25] to study the privacy level of a data protection method presents the advantage that it is an unified approach to deal with different types of scenarios (applicable to both numerical or ordinal data, and even to the case of synthetic data [27]). This methodology is used e.g. in [6,17,30]. Distance-based record linkage (DB-RL) [21] and probabilistic record linkage (P-RL) [14] are the two standard methods for record linkage which are considered when analyzing a particular protection method. However, as shown in a few works [17,26], record linkage can be *tuned* by an intruder, depending on the specific data protection method which is being used, to improve the results of the attack.

In this paper we add a new argument in this direction. Namely, we first show by means of some experiments that the privacy level of PCP-microaggregation and ZScore-microaggregation is quite good, when only the standard record linkage methods DB-RL and P-RL are considered. However, we then design and implement a new ad hoc record linkage method, that we call *projected record linkage*, specifically designed to increase the disclosure risk of these two projection-based multivariate microaggregation schemes. Informally, the projected record linkage method computes the projected values of the protected values, then it uses the standard distance-based record linkage method to compare the projections with the original record to be linked. The experiments described in this paper show that this new record linkage method finds (many) more correct links than DB-RL or P-RL. The conclusion is that the real disclosure risk of PCP-microaggregation and ZScore-microaggregation is underestimated when it is computed using only standard record linkage methods. In order to confirm whether the projected record linkage method is really specific for these two methods, we also apply it to MDAV, obtaining better results than standard methods like DB-RL or P-RL in most of the cases. This last result is quite surprising, since the projected record linkage method is originally designed to work well for projection-based microaggregation.

The rest of the paper is organized as follows. In Section 2, we explain some preliminary concepts like the disclosure risk and the information loss of a microdata protection method, and the basic definitions and some standard protocols for microaggregation. In Section 3, we first propose a formal and generic description of PCP-microaggregation and ZScore-microaggregation. We then perform experiments with two reference databases, Census and EIA, to compute the standard score of MDAV and the two new projection-based microaggregation methods. Then, in Section 4, we present a new record linkage technique based on projections; we experiment this technique with all the considered microaggregation methods, and then we compute two updated and more realistic scores. Finally, in Section 5, we draw the conclusions of this work and present some future work.

2. Preliminaries

In this section, we explain basic concepts about microdata protection methods, including the score, a measure to compare two protection methods, by combining their information loss and their disclosure risk. We also review the definition and some results about microaggregation.

2.1. Microdata protection methods

A dataset X can be viewed as a matrix with n rows (*records*) and V columns (*attributes*), where each row contains V attributes of an individual. The attributes in a dataset can be classified in two different categories, depending on their capability to identify unique individuals, as follows:

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

- Identifiers. The identifier attributes can be used to identify the individual unambiguously. A typical example of identifier is the passport number.
- Quasi-identifiers. A quasi-identifier attribute is an attribute that is not able to identify a single individual when it is used alone. However, when it is combined with other quasi-identifier attributes, they can univocally identify an individual. Among the quasi-identifier attributes, we distinguish between confidential and non-confidential, depending on the kind of information that they contain. An example of non-confidential quasi-identifier attribute would be the zip code, while a confidential quasi-identifier might be the salary.

Normally, before releasing a dataset with confidential attributes, a protection method ρ is applied, leading to a protected dataset X'. Indeed, we will assume the following typical scenario: (i) identifier attributes in X are either removed or encrypted, therefore we will write $X = X_{nc} \| X_c$; (ii) confidential quasi-identifier attributes X_c are not modified, and so we have $X'_c = X_c$; (iii) the protection method itself is applied to non-confidential quasi-identifier attributes, in order to preserve the privacy of the individuals whose confidential data are being released. Therefore, we have $X'_{nc} = \rho(X_{nc})$. This scenario, which was used first in [6] to compare several protection methods, has also been adopted in other works like [27].

2.2. Record linkage

Once the protected dataset X' is released, everybody can see its content $X' = X'_{nc} \| X_c$. We assume now that an intruder obtains from another data source another non-protected dataset $Y = Y_{id} \| Y_{nc}$ which includes one identifier and some (maybe all) of the non-confidential quasi-identifier attributes of some (maybe all) of the individuals whose data are in X. The goal of such an intruder is to find correct links between the protected dataset X' and the non-protected dataset Y using the common attributes between X' and Y (X'_{nc} and Y_{nc}). If the intruder is able to correctly link a record of Y with its corresponding protected record in X', then he will know that the matching (not modified) confidential information X_c belongs to the individual with identifier Y_{id}, breaking therefore the privacy of this individual. Therefore, the disclosure risk (i.e. the level of privacy) of a protection method is directly related to the difficulty of finding correct linkages between original and protected data.

In this paper, we will consider the worst case for the intruder, where there is only one original record in $Y \cap X$, and so the goal of the intruder is to find the protected record in X' which corresponds to this original one. The situation in real life can be different, more favorable to the intruder, who can know more than one of the original records in X. In this case, if the record linkage method for a single original record is already good enough, the intruder can run it in parallel for all the original records he wants to link. Additionally, maybe there exists a more effective record linkage method for the case where many original records are known. For example, if the applied microdata protection method is deterministic and an intruder knows all the original records in X, the best possible record linkage strategy is simply to run again the protection method on X to obtain the same X'; in this case, all the correct links are found and the protection method does not offer any privacy at all. Summing up, if we find record linkage strategies which are already successful in this theoretically worst scenario for the intruder, then we can conclude that the considered protection method does not offer enough privacy, because the situation in real life can be even more advantageous for the intruder.

Two record linkage (also known as *re-identification*) methods are usually considered in the literature:

- *Distance-based record linkage (DB-RL).* It can be applied when a distance between records in the original and protected datasets can be defined. In the standard method proposed in [21], every protected record is linked with the closest original record using the Euclidean distance after normalizing both datasets.
- *Probabilistic record linkage (P-RL).* In this case, the matching algorithm uses the linear sum assignment model to choose which pairs of the original and protected records must be matched. In order to compute this model, the EM (expectation-maximization) algorithm [12,15] is normally used. Informally, for each pair of records (a, b) where a is an original record of the dataset Y and b is a protected record of the dataset X', we define a coincidence vector $\gamma(a,b) = (\gamma_1(a,b) \ldots \gamma_n(a,b))$, where $\gamma_i(a, b)$ is defined as 1 if $V_i(a) = V_i(b)$ and as 0 if $V_i(a) \neq V_i(b)$. According to some criterion defined over these coincidence vectors, pairs are classified as linked pairs (LP) or non-linked pairs (NP). This method was introduced in [14].

2.3. Score computation

A microdata protection method must guarantee a certain level of privacy (low disclosure risk). At the same time, since the goal is to allow third parties to perform reliable statistical computations over the released (protected) data, the protection method must ensure somehow that the protected data are statistically close to the original data. In other words, the amount of information which is lost by releasing the protected data instead of the original one should not be very high.

We thus have two inversely related values for each microdata protection method: the *disclosure risk* (*DR*), which is the risk that an intruder obtains correct relations between the protected and the original data; and the *information loss* (*IL*) caused by the protection method. If one of them increases, the other one necessarily decreases. The two extreme cases are the following ones: (i) if the original microdata are released, then information loss is zero, but the disclosure risk is maximum; (ii) if the original microdata are encrypted and then released, the disclosure risk is (almost) zero, but the information loss is maximum.

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

There are different measures proposed in the literature to evaluate the quality of a data protection method, by taking into account these two values. We will use one of the most popular ones, the *score*, which was presented in [5] and used in many other works [16,30]. The score is a simple and natural way to evaluate the trade-off between the information loss and the disclosure risk because it is defined as the average of these two values. Namely,

$$\text{score} = \frac{IL + DR}{2}.$$

In the score framework, information loss and disclosure risk are computed as follows.

- *Information loss (IL).* The overall IL is computed as $IL = 100(0.2IL_1 + 0.2IL_2 + 0.2IL_3 + 0.2IL_4 + 0.2IL_5)$, where
  (i) IL_1 is the mean absolute error of the original microdata X with respect to the protected data X'.
  (ii) IL_2 is the mean variation of the attribute average vectors.
  (iii) IL_3 is the mean variation of the attribute covariance matrices.
  (iv) IL_4 is the mean variation of the attribute variance vectors.
  (v) IL_5 is the mean variation of the attribute correlation matrices.

- *Disclosure risk (DR).* Three alternatives are considered for measuring the disclosure risk. They are the two variations of record linkage explained in Section 2.2 and the interval disclosure. Half weight is given to record linkage and the remaining half weight is given to interval disclosure. For record linkage, the average of the two methods is computed. Formally, this corresponds to $DR = 0.25 \cdot DLD + 0.25 \cdot PLD + 0.5 \cdot ID$, where
  (i) *DLD*, the *distance-based linkage disclosure risk*, is the average percentage of correctly linked records using distance-based record linkage;
  (ii) *PLD*, the *probabilistic linkage disclosure risk*, is the average percentage of correctly linked records using probabilistic record linkage; and
  (iii) *ID*, the *interval disclosure risk*, is the average percentage of protected values falling into the intervals around their corresponding original values.

These three values are computed over the number of attributes that the intruder is assumed to know.

The idea of the score is to work without prior knowledge about both the use of the data and the possible behaviour of the intruders. For this reason, the score always computes the arithmetic mean of all the considered options (the same weight is assigned to each one). Of course, other choices could be possible (for example, considering $\text{Score}' = 0.3IL + 0.7DR$, if one is more interested in privacy than in utility of the data; or $\text{DR}' = 0.3ID + 0.7LD$, if one thinks that linkage disclosure attacks make more sense than interval disclosure attacks). These modifications would assume some prior knowledge on the use of the data or the behaviour of the intruders, which would go against the philosophy of the score measure.

2.4. Microaggregation

Microaggregation is one of the most popular, studied and used microdata protection methods, because its score values is ranked as one of the best ones [5]. Microaggregation builds small clusters of at least k elements of v variables and replaces the original records by the centroid of the cluster to which the record belongs.

The goal of a microaggregation is to minimize the total sum of square error

$$\text{SSE} = \sum_{i=1}^{c} \sum_{x_{ij} \in C_i} (x_{ij} - \bar{x}_i)^T (x_{ij} - \bar{x}_i),$$

where c is the total number of clusters, C_i is the ith cluster and $\bar{x}_i$ is the centroid of C_i. The restriction is $|C_i| \geqslant k$, for all $i = 1, \ldots, c$.

When the number of attributes is large, the basic microaggregation technique suffers from a low statistical utility, especially if the attributes are not much correlated. This is so because in this case the distances between original records in the dataset and the centroids are quite large. Therefore, a lot of information on the original data is lost and is not included in the released (protected) dataset.

To solve this drawback, the following natural strategy is applied by statistical agencies: the dataset is split into smaller blocks of attributes, and microaggregation is applied separately to each block. In this way, the information loss is lower but at the cost of a loss in the achieved level of privacy. Indeed, the property of k-anonymity is not ensured now. For example, the k records which fall in the same cluster for the first block of attributes, can fall in different clusters for all the other blocks of attributes. So, the resulting protected records will not be equal and no k-anonymity is ensured. The simplest approach for microaggregation is when the size of the attribute blocks is equal to one, in other words, each attribute is protected independently. This corresponds to *univariate microaggregation* or *individual ranking microaggregation*.

In this section we will explain several different algorithms that have been proposed (in more or less detail) for microaggregation. First we will explain a deterministic and optimal algorithm for univariate microaggregation; it will be used later

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

on when we will explain two methods for projection-based multivariate microaggregation: PCP-microaggregation and Zscores-microaggregation. Finally, we will describe one of the most used methods for heuristic microaggregation (specially for the multivariate case, although it can be applied to the univariate case, as well): the MDAV algorithm.

2.4.1. Optimal univariate microaggregation

In [7], two results related to optimal univariate microaggregation are proved, which constrain the search space and make possible a polynomial time algorithm:

- *Result 1.* For any optimal partition, when the elements are sorted according to the variable, the elements in each cluster are contiguous (non overlapping clusters exist).
- *Result 2.* All clusters of any optimal partition have between k and $2k-1$ elements.

One method for optimal univariate microaggregation was proposed in [11]. Let $V=(v_1,\ldots,v_n)$ be a vector containing the n values of the variable to be protected. Assume that the values are sorted in increasing order: $v_i \leqslant v_j$ if $i<j$. Recall that k is an integer satisfying $1 \leqslant k < n$, which is a parameter of the microaggregation configuration.

Given V and k, a graph $G_{k,n}$ is defined as follows. First, we define one node of G for each element v_i in V, plus one additional node v_0 (this node is later needed to apply the Dijkstra algorithm). Then, for each node v_i, we add to the graph the directed edges $\langle v_i, v_j\rangle$ for all j such that $i+k \leqslant j < i+2k$. The edge $\langle v_i, v_j\rangle$ means that the values $(v_i,\ldots,v_j)$ might form one of the possible clusters (recall Result 1 above). The cost of the edge $\langle v_i, v_j\rangle$ is defined as the within-group sum of square error for such cluster. That is, $\mathrm{SSE}=\sum_{l=i}^{j}(x_l-\bar{x})^2$, where $\bar{x}$ is the average record of the cluster.

Given this graph, the optimal univariate microaggregation is formed by the clusters determined by the edges which are in the shortest path between the nodes G_0 and G_n (this path is found by applying the Dijkstra algorithm).

2.4.2. MDAV microaggregation

The MDAV algorithm [8,13] is an heuristic algorithm for clustering records in a dataset X so that each cluster is constrained to have at least k records. This algorithm can be used for univariate microaggregation and multivariate microaggregation. The MDAV algorithm is described in Algorithm 1.

This generic algorithm can be instantiated for different data types, using appropriate definitions for distance and average. In the most common case, the *most distant record* and the *closest records* are computed using the Euclidean distance, and *the average record* is defined as the arithmetic mean of the records.

Algorithm 1. MDAV

```
Data: X: original dataset, k: integer
Result: X': protected dataset
1  begin
2      Split the dataset X into r sub-datasets {X_i}_{1≤i≤r}, each one with v_i
       attributes of the n records, such that Σ_{i=1}^{r} v_i = V
3      foreach (X_i ∈ X) do
4          while (|X_i| > k) do
5              Compute the average record x̄ of all records in X_i
6              Consider the most distant record x_r to the average record x̄
7              Form a cluster around x_r. The cluster contains x_r together
               with the k − 1 closest records to x_r
8              Replace all the original values in this cluster with the centroid
9              Remove these records from sub-dataset X_i
10             if (|X| > k) then
11                 Find the most distant record x_s from record x_r
12                 Form a cluster around x_s. The cluster contains x_s together
                   with the k − 1 closest records to x_s
13                 Replace all the original values in this cluster with the
                   centroid
14                 Remove these records from sub-dataset X_i
15             Form a cluster with the remaining records
16             Replace all the original values in this cluster with the centroid
17 end
```

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

3. Projection-based microaggregation

The main problem when one tries to extend the optimal univariate solution (see Section 2.4.1) to the case of multivariate microaggregation is how to sort multivariate data. One possibility, as we have seen in the last section, is to order the points with respect to their distance to the global centroid of the data.

A different possibility is to reduce the dimensionality of the problem, from more than one variable to 1 variable, by applying some projection method. Previous references to this idea [5,28] considered only the particular case where all the V variables of the dataset are projected. We detail here a more general method, where projection-based microaggregation is applied in parallel to blocks of v_i variables, where $\sum_i v_i = V$. As we mentioned in the introduction, the resulting protected dataset will not enjoy k-anonymity. However, the information loss of the protected dataset with respect to the original one will be lower.

Algorithm 2. Projected microaggregation

Data: X: original dataset, k: integer
Result: X': protected dataset

```
1 begin
2   Split the dataset X into r sub-datasets {X_i}_{1≤i≤r}, each one with a_i
    attributes of the n records, such that sum_{i=1}^{r} a_i = A
3   foreach (X_i ∈ X) do
4     Apply a projection algorithm to the attributes in X_i, which results
      in an univariate vector z_i with n components (one for each record)
5     Sort the components of z_i in increasing order
6     Apply to the sorted vector z_i the following variant of the
      univariate optimal microaggregation method explained in
      Section 2.4.1: use the algorithm defining the cost of the edges
      ⟨z_{i,s}, z_{i,t}⟩, with s < t, as the within-group sum of square error for
      the a_i-dimensional cluster in X_i which contains the original
      attributes of the records whose projected values are in the set
      {z_{i,s}, z_{i,s+1}, ..., z_{i,t}}
7     For each cluster resulting from the previous step, compute the
      v_i-dimensional centroid and replace all the records in the cluster
      by the centroid
8 end
```

The proposed multivariate microaggregation algorithm is described in Algorithm 2, when applied to a dataset X with n records and V attributes:

Depending on the projection method which is applied to the variables, we will obtain different methods of multivariate microaggregation. Due to the fact that they should preserve as much statistical properties of the data as possible (desirable in the scenario of statistical data protection), the PCP and Zscores projection methods seem to be the best choice. We call the resulting microaggregation algorithms *PCP-microaggregation* and *Zscore-microaggregation*. For completeness, we review now the PCP and Zscores projection methods.

Principal component projection. Formally principal component projection (PCP in short) works as follows: let us assume that values of v variables for n individuals are stored in a matrix X of dimension $n \times v$, where columns contain variables and rows contain individuals. For the sake of simplicity, we will assume here that data are standardized (i.e., the data has $\mu = 0$ and $\sigma = 1$, and so the covariance matrix is $S = 1/nX^{\mathrm{T}}X$).

The first principal component is defined as the linear combination of the variables which has the maximum variance. Therefore, this first principal component will be represented using a vector $z_1 = Xa_1$, for some vector a_1 with v components, to be found. Since the original values have $\mu = 0$, we have that z_1 also has $\mu = 0$, and its variance is

$$\frac{1}{n} z_1^{\mathrm{T}} z_1 = \frac{1}{n} a_1^{\mathrm{T}} X^{\mathrm{T}} X a_1 = a_1^{\mathrm{T}} S a_1. \tag{1}$$

Since S is positive-definite, the variance increases when the module of the vector a_1 does. For this reason, to find a concrete solution for the maximization of the expression (1), some constraint on the module of a_1 is needed; in this case, the search is limited to vectors a_1 with module 1 (i.e. $a_1^{\mathrm{T}} a_1 = 1$). This is equivalent to maximize the following expression, where a Lagrange multiplier has been added to the variance:

$$M = a^{\mathrm{T}} S a_1 - \lambda(a_1^{\mathrm{T}} a_1 - 1). \tag{2}$$

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

To maximize expression (2), the derivative with respect to the a_1 components must be made equal to 0.

$$\frac{\partial M}{\partial a_1} = 2Sa_1 - 2\lambda a_1 = 0. \tag{3}$$

The solution for such equation is $Sa_1 = \lambda a_1$, which implies that a_1 is an eigenvector of the matrix S, and λ is its corresponding eigenvalue. To determine which eigenvalue of S is the right solution, Eq. (3) is left-multiplied with a_1^T, leading to

$$a_1^T S a_1 = \lambda a_1^T a_1 = \lambda.$$

Summing up, λ is the variance of z_1. Since the goal is to maximize the variance, λ is the largest eigenvalue of the matrix S, and its associate eigenvector a_1 defines the coefficients of the projection (PCP). Therefore, the final projected value is

$$\mathrm{PCP} = \sum_{i=1}^{v} a_i x_i.$$

Zscores projection. We assume again that values of the v variables for the n individuals are stored in a matrix X of dimension $n \times v$. Given a record (a row) $(x_1, x_2, \ldots, x_v)$ in X, the sum of Zscores projection is defined as the single element

$$Z = \sum_{i=1}^{v} \frac{x_i - \mu_i}{\sigma_i},$$

where μ_i is the average and σ_i is the variance of the ith variable, computed by taking into consideration all the records in X.

3.1. Standard score for multivariate microaggregation methods

The goal of this paper is to analyze the information loss and the disclosure risk (and so, the score) of these two projection-based microaggregation methods, comparing them with a very popular algorithm for multivariate microaggregation, MDAV. In order to compare them, we have protected two different reference microdata files, proposed in the CASC project [3], with different instances of the three methods, and then we have computed the resulting scores, after having applied the standard information loss and disclosure risk measures (detailed in Section 2.3). The first microdata file, called Census, has been extracted using the data extraction system (DES) from the US Census Bureau [4]. This dataset contains 1080 records with 13

Table 1
Score of different microaggregation methods and parameterizations, applied to Census dataset

Method	*IL*	*DLD*	*PLD*	*ID*	*DR*	Score
Mic2PCP05	80.96	12.93	5.70	42.60	25.96	53.46
Mic2PCP15	92.94	8.46	2.94	35.64	20.67	56.81
Mic2PCP25	84.77	6.61	1.94	32.93	18.60	51.69
Mic2Zscore05	101.95	21.24	9.57	50.69	33.05	67.50
Mic2Zscore15	121.76	16.17	6.11	46.14	28.64	75.20
Mic2Zscore25	122.72	14.61	5.76	44.58	27.38	75.05
Mic2MDAV05	19.30	69.06	49.22	74.77	66.95	43.13
Mic2MDAV15	37.70	45.83	26.67	60.94	48.60	43.15
Mic2MDAV25	47.16	28.56	16.81	51.93	37.31	42.23
Mic3PCP05	57.72	10.15	5.71	43.48	25.71	41.71
Mic3PCP15	71.28	4.35	3.49	37.36	20.64	45.96
Mic3PCP25	72.49	4.07	2.65	35.51	19.44	45.96
Mic3Zscore05	90.72	14.97	11.48	51.89	32.56	61.64
Mic3Zscore15	124.92	9.57	6.94	48.18	28.22	76.57
Mic3Zscore25	128.25	9.23	5.86	46.73	27.14	77.69
Mic3MDAV05	30.66	37.44	33.58	65.21	50.36	40.51
Mic3MDAV15	42.76	22.75	19.38	54.79	37.93	40.34
Mic3MDAV25	56.13	15.86	13.36	51.57	33.09	44.61
Mic4PCP05	72.23	6.48	3.06	45.12	24.94	48.59
Mic4PCP15	91.74	3.43	2.04	40.73	21.73	56.74
Mic4PCP25	92.17	2.92	1.71	39.72	21.02	56.59
Mic4Zscore05	103.98	11.25	7.22	50.64	29.94	66.96
Mic4Zscore15	136.65	6.53	3.75	46.74	25.94	81.29
Mic4Zscore25	133.39	5.69	2.92	45.30	24.80	79.10
Mic4MDAV05	34.67	31.90	24.35	61.37	44.75	39.71
Mic4MDAV15	45.58	15.97	12.31	52.43	33.29	39.43
Mic4MDAV25	54.60	11.20	7.08	45.09	27.12	40.86

Mic*i*var*j* corresponds to microaggregation using variation var (either PCP, Zscore of MDAV) with $v = i$ and $k = j$.

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

attributes each (i.e., 14,040 values to be protected). The second microdata file, called EIA, was extracted from the US Energy Information Authority [29]. It contains 4092 records consisting of 10 attributes (i.e., 40,920 values to be protected).

Tables 1 and 2 show the results of these experiments. Each of the three microaggregation methods has been applied with the following nine parameterizations of the pairs (k,v): $k = 5, 15, 25$ for the minimal number of elements in the resulting clusters, and $v = 2, 3, 4$ for the number of attributes contained in each block of variables to which microaggregation is applied. For example, Mic2Zscore15 refers to the Zscore-microaggregation method applied to blocks of $v = 2$ attributes, with the constraint that resulting clusters must contain at least $k = 15$ records. In the cases where the total number of attributes is not a multiple of v (for example, this always happens with EIA dataset, since 13 is prime), the last non used attributes are non microaggregated and removed from the beginning.

The values of DLD, PLD and ID are average values. We have taken into account different cases, according to the number of variables of the original record(s), to be linked, that the intruder knows. This number varies from 2 to the total number of variables of each dataset. The values in the table are the average of the obtained correct links in all these cases, for each parameterization of each microaggregation method.

For the Census dataset, the best scores are clearly obtained with MDAV; but for the larger EIA dataset, both PCP-microaggregation and (specially) Zscore-microaggregation lead to better scores than MDAV. The lowest score using Zscore-microaggregation is 16.55 (Mic3Zscore05) while the lowest score using MDAV is 29.15 (Mic3MDAV15). Since MDAV produces in (almost) all the cases less information loss (*IL*) than the other two methods, the difference has to come from the disclosure risk part. In effect, the standard methods (distance-based, probabilistic) for record linkage are less effective against the two projection-based microaggregation than against MDAV. This difference in the disclosure risk seems to grow up with respect to the number of records of the dataset, because it is much larger in the case of EIA than in the case of Census (although they are different datasets, so it is impossible to formally conclude anything from this fact).

Summing up, the two projection-based microaggregation methods can be a real alternative to MDAV in some situations, offering a better privacy level against (standard) re-identification attacks. However, as we will see in the next section, these conclusions are not completely right: we will show some new record linkage methods, specially designed for projection-based microaggregation, which increase the real risk of re-identification (and so, the disclosure risk) of these methods. Maybe surprisingly, the new methods are also very effective when applied to MDAV. This will result in important changes to the real disclosure risk, and consequently the real score, of all these multivariate microaggregation methods.

Table 2
Score of different microaggregation methods and parameterizations, applied to EIA dataset

Method	*IL*	*DLD*	*PLD*	*ID*	*DR*	Score
Mic2PCP05	13.90	2.94	6.91	70.04	37.48	25.69
Mic2PCP15	17.24	1.72	2.37	67.67	34.86	26.05
Mic2PCP25	19.98	1.42	1.58	67.21	34.36	27.17
Mic2Zscore05	4.11	33.88	44.25	28.14	33.6	18.86
Mic2Zscore15	4.77	32.05	46.56	28.11	33.71	19.24
Mic2Zscore25	4.95	31.15	49.39	28.08	34.18	19.56
Mic2MDAV05	2.99	35.01	50.80	93.71	68.31	35.65
Mic2MDAV15	5.49	20.02	31.49	86.50	56.13	30.81
Mic2MDAV25	6.35	16.09	26.89	83.88	52.69	29.52
Mic3PCP05	16.08	2.47	2.69	62.79	32.68	24.38
Mic3PCP15	17.76	1.49	1.21	59.41	30.38	24.07
Mic3PCP25	18.49	1.31	0.90	58.49	28.90	24.14
Mic3Zscore05	13.74	7.20	11.99	29.12	19.35	16.55
Mic3Zscore15	15.82	4.66	8.46	29.63	18.09	16.95
Mic3Zscore25	16.76	4.83	7.54	29.80	17.99	17.37
Mic3MDAV05	7.64	21.47	34.53	85.52	56.76	32.20
Mic3MDAV15	9.99	11.33	22.67	79.63	48.31	29.15
Mic3MDAV25	11.12	9.60	18.32	77.63	45.80	28.46
Mic4PCP05	18.25	4.23	4.81	73.22	38.87	28.56
Mic4PCP15	16.39	1.96	2.13	70.48	36.26	26.33
Mic4PCP25	17.27	1.93	1.91	69.66	35.79	26.53
Mic4Zscore05	20.06	6.87	11.85	32.92	21.14	20.60
Mic4Zscore15	21.00	4.52	7.48	33.21	19.61	20.30
Mic4Zscore25	27.50	3.96	6.84	36.28	20.84	24.17
Mic4MDAV05	8.3	25.71	36.78	87.76	59.50	33.90
Mic4MDAV15	19.16	12.66	21.31	81.57	49.28	34.22
Mic4MDAV25	20.11	8.11	14.66	78.28	44.83	32.47

Mic*i*var*j* corresponds to microaggregation using variation var (either PCP, Zscore of MDAV) with $v = i$ and $k = j$.

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

4. Projected record linkage and its consequences

Recently, some works have appeared which present ad hoc record linkage techniques, specifically designed for some particular microdata protection methods. These techniques have a higher success rate, when looking for correct links, than standard record linkage methods like the probabilistic or the distance-based ones. Some examples can be found in the papers [17,19,26], related to microaggregation and rank swapping.

In some way, the philosophical conclusion of these works is that the standard computations of the disclosure risk (and score) values should be reconsidered, because a smart intruder can find more successful ways to break the privacy of a data protection method, in particular by obtaining more correct links than those obtained when standard record linkage methods are applied. Following this approach, a data protection method would be considered secure if the most effective record-linkage technique against it had a relatively low success rate (for example, below the threshold of 50%). Ideally, such a bound for the maximum number of correct record links should be obtained in an analytical way, independently of the possible strategies of the adversary. But this seems to be very hard for the existing methods for statistical data protection, which are quite heuristic. For this reason, when designing or analyzing a (new) data protection method, it is important to consider and study possible ad hoc record linkage techniques for this method, for example by implementing these techniques to compare their success rates with those of standard techniques.

Since we personally think that this approach is the correct one, we have considered an ad hoc record linkage method, (initially) specific for the two projection-based microaggregation schemes studied in this work. In this section we describe this new record linkage technique, and then we show the experimental results that we have obtained when applying it to the Census and EIA datasets, protected with PCP-microaggregation, Zscore-microaggregation and MDAV. The conclusion is that the real disclosure risk of these three methods is significantly increased.

4.1. The projected record linkage technique

Let X' be the result of applying a data protection method to a dataset X, with n records and V attributes. Let y be an original record of X (obtained by an intruder, possibly from a different dataset Y). The goal of the record linkage method is to find the record $x' \in X'$ which corresponds to the original y.

Our new *projected record linkage* technique (Pro-RL) is specifically designed for the case of microaggregation. Therefore, we can assume that X' is implicitly split into r blocks X'_i of v_i attributes, according to the blocks which have been considered to perform microaggregation. The algorithm of the projected record linkage method is defined in Algorithm 3.

Algorithm 3. Projected record linkage

```
Data: Y: external dataset, X': protected dataset
Result: LP: linked pairs
1 begin
2   foreach X'_i of X' do
3     Apply a projection to the v_i attributes in X'_i, for all the protected
      records. For example, the projection can be PCP or the Zscores
      one (intuitively, if X'_i has been obtained with PCP
      microaggregation, then PCP should be chosen as the projection
      method for record linkage)
4     Apply the same projection to the corresponding v_i attributes of
      the original record y
5   The result of the previous step is a projected original record ỹ and n
    projected protected records x̃', all of them with r values
6   Find the record x̃'_* which is closest to ỹ (for example, according to the
    Euclidean distance)
7   Let x'_* be the protected record whose projection was x̃'_*. Then the
    output link is y ↔ x'_* is added to LP
8 end
```

In some way, the reasoning behind this strategy is that the results of projecting the original data, in X, and the protected data, in X', should be very similar, specially if the projection method applied in the record linkage algorithm has a similar statistical behaviour to the data protection method which transformed X into X'. The experiments that we have implemented, which are explained in the following section, show that this intuition is right.

Although we have explained here a version of the projected record linkage technique which is specific for microaggregation, the idea can be easily extended to work with any data protection method. One can choose to project all the attributes of

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

X into a single projected attribute, or to first split X in disjoint blocks of attributes, and then to project the attributes in each block, independently. We have implemented and run this generic record linkage technique against other protection methods (rank swapping, noise addition) and, differently to what happens in the case of microaggregation, the results do not improve the ones obtained by standard record linkage techniques.

In our opinion, the main reason is that (in general) the protection applied by these methods is done by considering only a value of a record (noise addition) or a single attribute (rank swapping), instead of a group of attributes as in multivariate microaggregation. Therefore, considering projections of attributes does not make so much sense.

4.2. Experimental results: new disclosure risks and scores

We have executed the two projected record linkage methods (using PCP and Zscores as the inherent projection mechanism) against all the protected datasets obtained in the experiments of Section 4.2; that is, the result of applying PCP-microaggregation, Zscore-microaggregation and MDAV, with different parameterizations, to the datasets Census and EIA.

Figs. 1–3 show the percentage of correct links obtained with different record linkage methods, when applied to both the Census and EIA datasets, protected with the three multivariate microaggregation schemes that we analyze in this paper using the most secure configuration (using blocks of four attributes). The percentage is computed by taking into account different scenarios, where the intruder knows different amounts of groups of attributes of the original record(s) to be linked, from 2 to all the groups of attributes.

The results of these experiments can be summarized as follows:

- The new projected record linkage methods obtain, in almost all the cases, more correct links than the other (standard) record linkage methods. The difference is in some cases very significant. For instance, when the Census dataset is protected using the Zscore-microaggregation with $v = 4$ and $k = 25$ (the most protected configuration) DLD is equal to

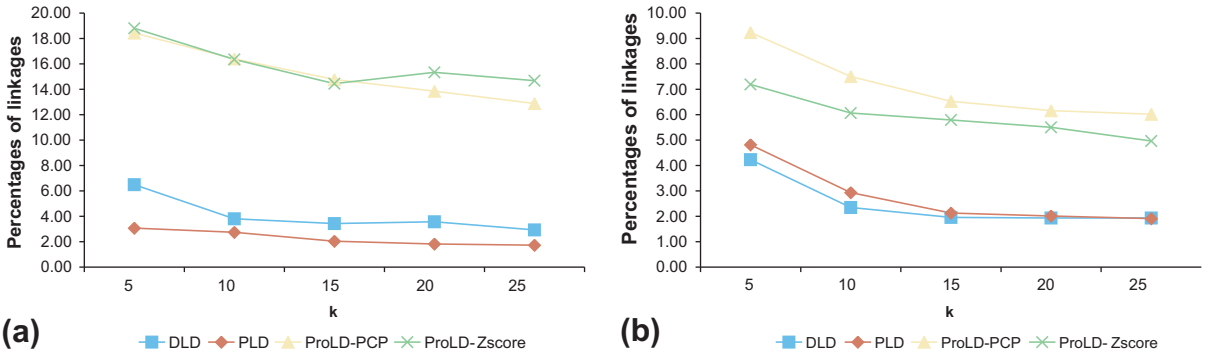


Fig. 1. Percentage of correct links obtained with different record linkage techniques, applied to the Census dataset (a) and EIA dataset (b), protected with PCP-microaggregation with $v = 4$.

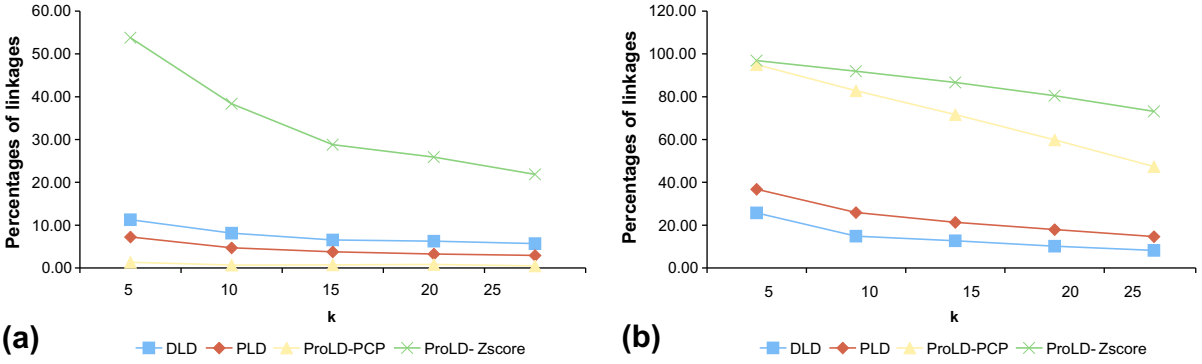


Fig. 2. Percentage of correct links obtained with different record linkage techniques, applied to the Census dataset (a) and EIA dataset (b), protected with Zscore-microaggregation with $v = 4$.

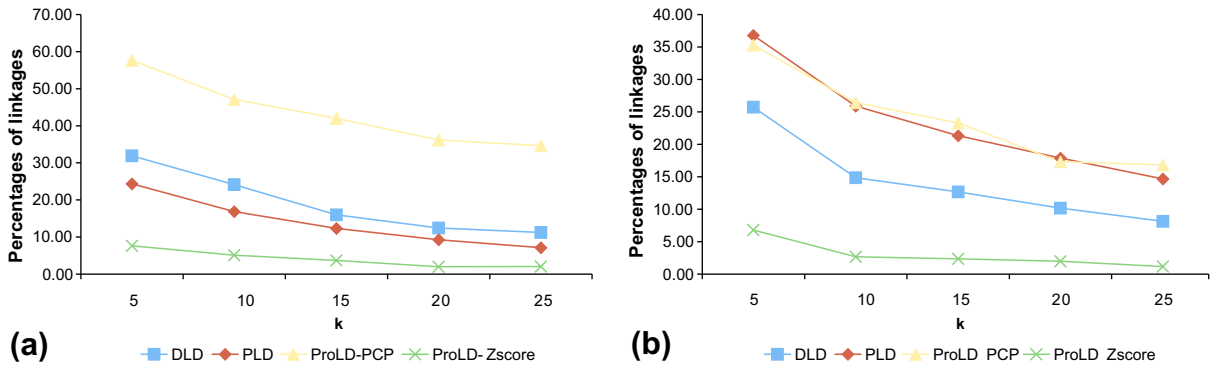


Fig. 3. Percentage of correct links obtained with different record linkage techniques, applied to the Census dataset (a) and EIA dataset (b), protected with MDAV with $v = 4$.

5.69, a very small value, and ProLD is equal to 21.85. A similar situation happens with MDAV, if we observe the configuration with $v = 3$ $k = 15$ (see Table 3), we observe an important increase of the disclosure risk: DLD is equal to 22.75 while ProLD is equal to 60.03, much more than 50% of correctly linked records.

- Not surprisingly, the most effective record linkage method against PCP-microaggregation is the projected one when PCP is used as the inherent projection, and the same happens with Zscore-microaggregation and Zscores projection. For instance, if we compare the ProLD-PCP and ProLD-Zscore in Fig. 2 we observe that the best method is to apply Zscores projection in the record linkage when the dataset is protected using Zscore-microaggregation.
- Using PCP projection in the record linkage method to obtain correct links against Zscore-microaagregation, or vice versa, is not effective at all.

Table 3
New scores of the different microaggregation methods, applied to Census dataset

Method	*IL*	*DLD*	*PLD*	ProLD	*ID*	ScoreOld	ScoreNew	ScoreMax
Mic2PCP05	80.96	12.93	5.70	49.83	42.60	53.46	56.84	63.59
Mic2PCP15	92.94	8.46	2.94	45.67	35.64	56.81	60.14	66.80
Mic2PCP25	84.77	6.61	1.94	43.70	32.93	51.69	54.97	61.54
Mic2Zscore05	101.95	21.24	9.57	99.59	50.69	67.50	74.52	88.55
Mic2Zscore15	121.76	16.17	6.11	96.28	46.14	75.20	82.29	96.48
Mic2Zscore25	122.72	14.61	5.76	90.65	44.58	75.05	81.75	95.17
Mic2MDAV05	19.30	69.06	49.22	80.59	74.77	43.13	44.92	48.49
Mic2MDAV15	37.70	45.83	26.67	69.48	60.94	43.15	45.92	51.46
Mic2MDAV25	47.16	28.56	16.81	61.28	51.93	42.23	45.45	51.88
Mic3PCP05	57.72	10.15	5.71	27.59	43.48	41.71	43.35	46.63
Mic3PCP15	71.28	4.35	3.49	23.46	37.36	45.96	47.59	50.84
Mic3PCP25	72.49	4.07	2.65	20.90	35.51	45.96	47.42	50.35
Mic3Zscore05	90.72	14.97	11.48	87.04	51.89	61.64	67.79	80.09
Mic3Zscore15	124.92	9.57	6.94	67.41	48.18	76.57	81.50	91.36
Mic3Zscore25	128.25	9.23	5.86	60.80	46.73	77.69	82.13	91.01
Mic3MDAV05	30.66	37.44	33.58	75.06	65.21	40.51	43.81	50.40
Mic3MDAV15	42.76	22.75	19.38	60.03	54.79	40.34	43.59	50.09
Mic3MDAV25	56.13	15.86	13.36	53.49	51.57	44.61	47.85	54.33
Mic4PCP05	72.23	6.48	3.06	18.43	45.12	48.59	49.73	52.00
Mic4PCP15	91.74	3.43	2.04	14.77	40.73	56.74	57.74	59.74
Mic4PCP25	92.17	2.92	1.71	12.87	39.72	56.59	57.47	59.23
Mic4Zscore05	103.98	11.25	7.22	53.75	50.64	66.96	70.67	78.09
Mic4Zscore15	136.65	6.53	3.75	28.80	46.74	81.29	83.26	87.21
Mic4Zscore25	133.39	5.69	2.92	21.85	45.30	79.10	80.56	83.48
Mic4MDAV05	34.67	31.90	24.35	57.64	61.37	39.71	42.17	47.09
Mic4MDAV15	45.58	15.97	12.31	41.99	52.43	39.43	41.75	46.40
Mic4MDAV25	54.60	11.20	7.08	34.63	45.09	40.86	42.98	47.23

Mic*i*var*j* corresponds to microaggregation using variation var (either PCP, Zscore of MDAV) with $v = i$ and $k = j$.

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

- When applied to MDAV, the projected record linkage method using PCP is very effective (more than any other method in the case of Census, and only overcome by probabilistic record linkage in some instances of EIA). However, using Zscores as the inherent projection leads to quite bad results. For example, if we observe the MDAV configuration with $v = 3$ and $k = 25$ in the Census dataset (Table 3), we observe that ProLD is equal to 53.49 while DLD and PLD are equal to 15.86 and 13.36. Here, the ProLD is four times higher than standard record linkage disclosure risks.

Obviously, the fact that the projected record linkage technique has a higher success rate than the other record linkage techniques must have a direct impact in the real disclosure risk (and so, in the score) of the studied multivariate microaggregation methods.

One possibility is to compute the disclosure risk, again, as the average of the interval disclosure risk (ID) and the record linkage risk, but now this last value is computed as the average of three values: distance based linkage disclosure risk (DLD), probabilistic linkage disclosure risk (PLD) and projected linkage disclosure risk (ProLD), which is the maximum percentage of correct links obtained by the projected record linkage technique, using either PCP or Zscores as the inherent projection. When the protection method is PCP-microaggregation or MDAV, the maximum is obtained by using PCP as the inherent projection for record linkage; when the protection method is Zscore-microaggregation, the maximum is obtained by using Zscores projection.

Another possibility, maybe more realistic, is to assume that the intruder who wants to break the privacy of the protection method knows which is the most successful strategy to find correct links. For example, after reading this paper, he may know that projected record linkage with PCP as the inherent projection is the best known technique to attack PCP-microaggregation. In this case, it is clear that he will always use this technique to find correct links between original and protected records. Therefore, considering other values to compute the linkage disclosure risk would make no sense in this situation; the real linkage disclosure risk would be defined as the maximum among all the linkage disclosure risk values: DLD, PLD, ProLD.

Summing up, there would be two different alternatives to compute the new scores of these methods. In the first one, that we call *ScoreNew*, the disclosure risk is computed as $DR_New = 0.5 \cdot (0.333 \cdot DLD + 0.333 \cdot PLD + 0.333 ProLD) + 0.5 \cdot ID$. In the second one, that we call *ScoreMax*, the disclosure risk is computed as $DR_Max = 0.5 \cdot \mathrm{MAX}\{DLD, PLD, ProLD\} + 0.5 \cdot ID$. As usual, the final score value is computed as the average of the information loss (*IL*) and the corresponding disclosure risk (DR).

Tables 3 and 4 show the new values of the scores, which strongly depend on the success rate (ProLD) of the new projected record linkage technique. This can be easily verified by comparing the new ScoreNew and ScoreMax with the standard Score, computed by considering only standard (and generic) record linkage techniques. We consider the same parameterizations than in Section 4.2.

Table 4
New scores of the different microaggregation methods, applied to EIA dataset

Method	*IL*	*DLD*	*PLD*	ProLD	*ID*	ScoreOld	ScoreNew	ScoreMax
Mic2PCP05	13.90	2.94	6.91	16.20	70.04	25.69	27.35	28.51
Mic2PCP15	17.24	1.72	2.37	9.17	67.67	26.05	26.98	27.83
Mic2PCP25	19.98	1.42	1.58	8.64	67.21	27.17	28.07	28.95
Mic2Zscore05	4.11	33.88	44.25	98.08	28.14	18.86	26.88	33.61
Mic2Zscore15	4.77	32.05	46.56	94.13	28.11	19.24	27.00	32.94
Mic2Zscore25	4.95	31.15	49.39	88.99	28.08	19.56	26.79	31.74
Mic2MDAV05	2.99	35.01	50.80	54.48	93.71	35.65	38.08	38.54
Mic2MDAV15	5.49	20.02	31.49	36.66	86.50	30.81	32.89	33.53
Mic2MDAV25	6.35	16.09	26.89	32.84	83.88	29.52	31.61	32.36
Mic3PCP05	16.08	2.47	2.69	7.17	62.79	24.38	24.97	25.53
Mic3PCP15	17.76	1.49	1.21	4.84	59.41	24.07	24.49	24.94
Mic3PCP25	18.49	1.31	0.90	3.85	58.49	24.14	24.46	24.83
Mic3Zscore05	13.74	7.20	11.99	91.09	29.12	16.55	27.04	36.92
Mic3Zscore15	15.82	4.66	8.46	79.03	29.63	16.95	26.25	35.07
Mic3Zscore25	16.76	4.83	7.54	71.00	29.80	17.37	25.65	33.58
Mic3MDAV05	7.64	21.47	34.53	35.40	85.52	32.20	33.94	34.05
Mic3MDAV15	9.99	11.33	22.67	20.53	79.63	29.15	30.30	30.57
Mic3MDAV25	11.12	9.60	18.32	15.35	77.63	28.46	29.18	29.55
Mic4PCP05	18.25	4.23	4.81	9.24	73.22	28.56	29.19	29.74
Mic4PCP15	16.39	1.96	2.13	6.52	70.48	26.33	26.90	27.45
Mic4PCP25	17.27	1.93	1.91	6.01	69.66	26.53	27.04	27.55
Mic4Zscore05	20.06	6.87	11.85	96.85	32.92	20.60	31.85	42.47
Mic4Zscore15	21.00	4.52	7.48	86.61	33.21	20.30	30.57	40.46
Mic4Zscore25	27.50	3.96	6.84	73.17	36.28	24.17	32.82	41.11
Mic4MDAV05	8.30	25.71	36.78	35.29	87.76	33.90	35.10	35.28
Mic4MDAV15	19.16	12.66	21.31	23.29	81.57	34.22	35.55	35.79
Mic4MDAV25	20.11	8.11	14.66	16.81	78.28	32.47	33.56	33.83

Micivar*j* corresponds to microaggregation using variation var (either PCP, Zscore of MDAV) with $v = i$ and $k = j$.

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

After looking at these tables, one can conclude that the real scores of multivariate microaggregation methods are not as good as one could think (see the classification of data protection methods in [6]). In particular, the percentage of correct links are now over the threshold of 50% in many cases, so it is not clear if these methods offer the desired level of privacy.

5. Conclusions and future work

In this paper we have analyzed some multivariate microaggregation methods. Namely, for the two projection-based ones (PCP-microaggregation and Zscore-microaggregation), we have described and implemented the first generic version of them, where not all the attributes must be necessarily microaggregated together. We have computed the standard score of these two methods, to see that they can be competitive when compared to (probably) the most popular scheme for multivariate microaggregation, MDAV.

Then, we have designed a new record linkage technique, called projected record linkage, as an ad hoc tool to find correct links between an original record and protected records, when the applied protection method is a projection-based microaggregation. The new technique is much more effective than standard techniques (like distance-based or probabilistic record linkage). The most surprising result, however, has been obtained when we have applied the new record linkage technique to MDAV, because it also overcomes the other record linkage methods in most of the cases. The conclusions of these experiments are that the real disclosure risk (and score) of these multivariate microaggregation methods is higher than what was believed up to now. In other words, the level of privacy that they offer is debatable.

As future work, other microaggregation methods [9], less popular than MDAV, or maybe other data protection methods different from microaggregation, could be subjected to the new projected record linkage technique, to see what is its real scope. From a constructive point of view, one should try to design new protection methods (for microaggregation or not) which resist even ad hoc record linkage attacks.

Acknowledgements

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) and by the Government of Catalonia (Grant 2005-SGR-00093) is acknowledged. Jordi Nin wants to thank the Spanish Council for Scientific Research (CSIC) for his I3P grant.

References

[1] N.R. Adam, J.C. Wortmann, Security-control for statistical databases: a comparative study, ACM Computing Surveys 21 (1989) 515–556.
[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu, *k*-Anonymity: Algorithms and hardness, Technical report, Stanford University, 2004.
[3] CASC: Computational Aspects of Statistical Confidentiality, European Project IST-2000-25069. <http://neon.vb.cbs.nl/casc>.
[4] Data Extraction System, US Census Bureau. <http://www.census.gov/>.
[5] J. Domingo-Ferrer, V. Torra, Disclosure control methods and information loss for microdata, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, 2001, pp. 91–110.
[6] J. Domingo-Ferrer, V. Torra, A quantitative comparison of disclosure control methods for microdata, in: P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, 2001, pp. 111–133.
[7] J. Domingo-Ferrer, J.M. Mateo-Sanz, Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering 14 (2002) 189–201.
[8] J. Domingo-Ferrer, A. Martínez-Ballesté, J.M. Mateo-Sanz, F. Sebé, Efficient multivariate data-oriented microaggregation, The VLDB Journal 15 (2006) 355–369.
[9] J. Domingo-Ferrer, F. Sebé, A. Solanas, A polynomial-time approximation to optimal multivariate microaggregation, Computers and Mathematics with Applications 55 (2008) 714–732.
[10] F. Felsö, J. Theeuwes, G. Wagner, Disclosure limitation in use: Results of a survey, 2001, pp. 17–42.
[11] S. Hansen, S. Mukherjee, A polynomial algorithm for optimal univariate microaggregation, Transactions on KDE 15 (4) (2003) 1043–1044.
[12] H. Hartley, Maximum likelihood estimation from incomplete data, Biometrics 14 (1958) 174194.
[13] A. Hundepool, A. Van de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, S. Giessing, μ-ARGUS version 3.2 Software and User's Manual Statistics, Netherlands, Voorburg NL, February 2003.
[14] M.A. Jaro, Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Society 84 (406) (1989) 414–420.
[15] G. McLachlan, T. Krishnan, The EM Algorithm and Extensions, Wiley Series in Probability and Statistics, John Wiley & Sons, 1997.
[16] P. Medrano-Gracia, J. Pont-Tuset, J. Nin, V. Muntés-Mulero, Ordered dataset vectorization for linear regression on data privacy, Modeling Decisions for Artificial Intelligence, Lecture Notes in Artificial Intelligence, vol. 4617, Springer-Verlag, 2007. pp. 361–372.
[17] J. Nin, J. Herranz, V. Torra, Rethinking rank swapping to decrease disclosure risk, Data and Knowledge Engineering 64 (1) (2008) 346–364.
[18] J. Nin, J. Herranz, V. Torra, How to group attributes in multivariate microaggregation, International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems 161 (2008) 121–138.
[19] J. Nin, V. Torra, Analysis of the Univariate Microaggregation Disclosure Risk, submitted for publication.
[20] A. Oganian, J. Domingo-Ferrer, On the complexity of optimal microaggregation for statistical disclosure control, Statistical Journal of United Nations Economic Commission for Europe 18 (4) (2000) 345–354.
[21] D. Pagliuca, G. Seri, Some results of individual ranking method on the system of enterprise accounts annual survey, Esprit SDC Project, Delivrable MI-3/D2, 1999.
[22] P. Samarati, L. Sweeney, Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression, SRI Intl. Tech. Rep., 1998.
[23] L. Sweeney, Achieving *k*-anonymity privacy protection using generalization and suppression, International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems 10 (5) (2002) 571–588.
[24] L. Sweeney, *k*-Anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems 10 (5) (2002) 557–570.

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014

[25] V. Torra, J. Domingo-Ferrer, Record linkage methods for multidatabase data mining, Information Fusion in Data Mining (2003) 101–132.
[26] V. Torra, S. Miyamoto, Evaluating fuzzy clustering algorithms for microdata protection, Lecture Notes in Computer Science 3050 (2004) 175–186.
[27] V. Torra, J.M. Abowd, J. Domingo-Ferrer, Using Mahalanobis distance-based record linkage for disclosure risk assessment, Privacy in Statistical Databases 2006, Lecture Notes in Computer Science, vol. 4302, Springer-Verlag, 2006. pp. 233–242.
[28] À. Torres, Contributions to microaggregation for statistical data protection, Ph.D. Thesis, Polytechnical University of Catalonia, 2003.
[29] US Energy Information Authority. <http://www.eia.doe.gov/cneaf/electricity/page/eia826.html>.
[30] W.E. Yancey, W.E. Winkler, R.H. Creecy, Disclosure risk assessment in perturbative microdata protection, Inference Control in Statistical Databases: From Theory to Practice, Lecture Notes in Computer Science, vol. 2316, Springer-Verlag, 2002. pp. 135–152.



Jordi Nin (Barcelona, Catalonia, 1979; BSc 2004, MSc 2007, PhD 2008 all in Computer Science) is a post-doctoral researcher at the Artificial Intelligence Research Institute (IIIA-CSIC) near Barcelona, Catalonia, Spain. His fields of interests are privacy technologies, machine learning and soft computing tools. He has been involved in several research projects funded by the Catalan and Spanish governments and the European Community. His research has been published in specialized journals and major conferences (around 30 papers).



Javier Herranz obtained his PhD in Applied Mathematics in 2005, in the Technical University of Catalonia (UPC, Barcelona, Spain). After that, he spent 9 months in the École Polytechnique (France) and 9 months in the Centrum voor Wiskunde en Informatica (CWI, The Netherlands), as a post-doctoral researcher, granted with an ERCIM fellowship. From January 2007, he works as a post-doctoral researcher at IIIA-CSIC (Bellaterra, Spain). His research interests include the design and analysis of cryptographic protocols and the study of privacy preserving operations involving databases.



Vicenç Torra (Barcelona, Catalonia, 1968; BSc 1991, MSc 1992, PhD 1994 all in Computer Science) is an Associate Research Professor at the Artificial Intelligence Research Institute (IIIA-CSIC) near Barcelona, Catalonia, Spain. His fields of interests are privacy technologies, decision making, information fusion and soft computing tools. He has had led several research projects funded by Catalan, Spanish, European and US sources. Research has been published in specialized journals (more than 60 articles in ISI JCR journals and LNCS) and major conferences. He has written a book on Artificial Intelligence (in Catalan) and another, with Y. Narukawa, entitled "Modeling Decisions: Information Fusion and Aggregation Operators" (Springer). He is co-editor of the Transactions on Data Privacy and member of the editorial board of Fuzzy Sets and Systems, Journal of Privacy Technology, Journal of Advanced Computational Intelligence and Intel. Informatics, and Mathware and Soft Computing. In 2004, he was co-founder with Y. Narukawa of the conference series Modeling Decisions for Artificial Intelligence (MDAI), whose successive editions he has co-chaired ever since. He was also a Co-chairman of the Privacy in Statistical Databases conference (PSD 2004).

Please cite this article in press as: J. Nin et al., On the disclosure risk of multivariate microaggregation, Data Knowl. Eng. (2008), doi:10.1016/j.datak.2008.06.014