Information Fusion 13 (2012) 274–284

Contents lists available at ScienceDirect

Information Fusion

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

Improving record linkage with supervised learning for disclosure risk assessment

Daniel Abril [a], Guillermo Navarro-Arribas [b,*,1], Vicenç Torra [a]

[a] IIIA, Artificial Intelligence Research Institute, CSIC, Spanish Council for Scientific Research, Campus UAB s/n, 08193 Bellaterra, Catalonia, Spain
[b] DEIC, Department of Information and Communications Engineering, UAB, Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain

ARTICLE INFO

Article history:
Received 24 December 2010
Received in revised form 10 May 2011
Accepted 12 May 2011
Available online 30 May 2011

Keywords:
Record linkage
Data privacy

ABSTRACT

In data privacy, record linkage can be used as an estimator of the disclosure risk of protected data. To model the worst case scenario one normally attempts to link records from the original data to the protected data. In this paper we introduce a parametrization of record linkage in terms of a weighted mean and its weights, and provide a supervised learning method to determine the optimum weights for the linkage process. That is, the parameters yielding a maximal record linkage between the protected and original data. We compare our method to standard record linkage with data from several protection methods widely used in statistical disclosure control, and study the results taking into account the performance in the linkage process, and its computational effort.

1. Introduction

Record linkage was initially introduced for database integration in Ref. [15] and further developed in Ref. [35] and with a formal mathematical foundation given in Ref. [19]. It identifies records from different databases (or data sources in general) that refer to the same entity. It is nowadays a popular technique employed by statistical agencies, research communities, and corporations, not only to integrate different databases or data sets in general [41,6], but also for data cleaning and quality [23,2,51] for example, by detecting duplicate records between several data sets [18]. As an indication of the relevance of linked data, both the UK and US governments have launched respectively web portals to centralize (and thus allow linking) different sources of governmental data.[2]

Record linkage has been applied on the estimation of the population size at the US Bureau of the Census [49,50,26], in building big social databases in Ref. [24], in epidemiology and medical studies in Refs. [27,36,20,28], in sociological sciences in Ref. [3], or in counterterrorism in Ref. [21].

More recently, in the context of data privacy, record linkage has emerged as an important technique to evaluate the disclosure risk of protected data. By identifying the links between the protected dataset and the original one, we can evaluate the re-identification risk of the protected data.

Given, for example, some statistical data X providing information about individuals as records of V attributes, and a protected version of the data X', we consider a possible intruder with some prior knowledge of the original data, that is, some values of V. The task of the intruder is to attempt to link the values of the protected data with the ones he knows. If the links can be established, the attacker can re-identify individuals from the protected data, and the protection is said to be broken.

To model the worst case scenario one normally attempts to link records from the original data to the protected data. This gives an estimation of the chances that an attacker will be able to re-identify records in the protected data. The estimation is usually used as a disclosure risk measure of the protection method applied to protect the data. That is, the percentage of correctly linked records between the protected dataset and the original dataset is taken as a measure for the disclosure risk of the data. This approach to measure the disclosure risk of protected data was initially introduced in Ref. [40] and adopted in much of the subsequent literature such as Refs. [10,30,54,42,52]. Note also that sampling is not taken into account in this approach, which means assuming that the intruder knows the sampled individuals in the data set. This is a common practice in the previously cited works.

There are several techniques and types of record linkage, mainly depending on whether they use a distance function to match records, or they use some probabilistic estimation (cf. Ref. [1], some recent techniques can also be found in Refs. [31,29]). We focus our work in distance-based record linkage [39], where the link between records is determined in terms of a distance function between them. Moreover, the distance function is a weighted aggregation of the distance function between the single attribute values of each record.

* Corresponding author.
E-mail addresses: dabril@iiia.csic.es (D. Abril), gnavarro@deic.uab.cat (G. Navarro-Arribas), vtorra@iiia.csic.es (V. Torra).

[1] Work done while the authors was at the Artificial Intelligence Research Institute (IIIA – CSIC).

[2] UK: http://data.gov.uk, USA: http://data.gov.

1566-2535/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2011.05.001

In this paper we present a supervised learning approach to determine the optimal parameters for the distance function between records. That is, the parameters yielding a maximal record linkage between the protected and original data. We have evaluated this novel linkage technique with a wide range of protection methods normally used in statistical disclosure control. Unlike previous work that improved record linkage by tuning it to specific protection methods [47,37,38], our proposal is focused toward generic data protection. By determining the weight of the variables, we can also identify key-variables for record linkage, that is those variables that entail more re-identification risk. Besides improving standard distance-based record linkage, our experiments have also shown other insightful results regarding the computation cost in record linkage. This works also provides a discussion on the appropriateness of a good parametrization of the aggregated distance, and its impact in the record linkage process.

1.1. Contributions and plan of the paper

The contributions presented in this paper depart from providing and analyzing the optimal distance based record linkage between a protected dataset X' and a non-protected dataset Y, which share some variables. The implications of our study in data privacy are detailed.

- Improvement in the linkage as compared to standard distance-based record linkage.
- Identification of key-attributes for record linkage.
- Evaluation of the computational cost, and its implications.

An important issue of our contribution is that our record linkage technique is considered for a generic use. Contrary to previous work, which attempts to improve standard distance-based record linkage [38,37], our proposal can be applied no matter the protection method used. Because of that, we have evaluated our proposal with different protection methods, and provide a discussion on the results.

We introduce record linkage and its use in data privacy in Section 2. Our supervised approach for distance-based record linkage is described in Section 3. Section 4 describes the evaluation of several protection methods with our proposal, and a discussion on the obtained results. Finally, Section 5 concludes the paper and details future work lines.

2. Record linkage in data privacy

In data privacy, record linkage can be used to re-identify individuals from a protected dataset. It serves as an evaluation of the protection method used by modeling the possible attack to be performed on the protected dataset.

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

- *Identifiers*: attributes that can be used to identify the individual unambiguously. A typical example of identifier is the passport number.
- *Quasi-identifiers*: attributes that are not able to identify a single individual when they are used alone. However, when combining several quasi-identifier attributes, they can unequivocally identify an individual. Among the quasi-identifier attributes, we distinguish between confidential (X_c) and non-confidential (X_{nc}), 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.

Before releasing the data, 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 first used in Ref. [10] to compare several protection methods, has also been adopted in other works like Ref. [42].

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 is 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.

Note that this will be the generic scenario. In order to provide a measure of disclosure risk in the protected dataset, one normally considers the same problem where $Y = X$. That is, the approach attempts to link records between the original dataset and the protected one. The percentage of correct links, records that are correctly linked between both datasets, is given as a global measure of the disclosure risk.

There are two main approaches for record linkage:

- *Distance based record linkage (DBRL)*. This approach [39] links each record a to the *closest* record in b. The *closest* record is defined in terms of a distance function.
- *Probabilistic record linkage (PRL)*. 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 [22,8,33] 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 $\gamma(a, b) = (\gamma_1(a, b) \dots \gamma_n(a, b))$, where $\gamma_i(a, b)$ is defined as 1 if $V_i(a) = V_i(b)$ and as 0 if $V_i(a) \neq V_i(b)$. According to some criterion defined over these coincidence vectors, pairs are classified as linked pairs (LP) or non-linked pairs (NP). This concrete method was introduced in Ref. [26], although probabilistic record linkage was first presented in Ref. [19].

Both approaches have been used extensively in the area of data privacy to evaluate the disclosure risk of protected data.

The work in this paper is focused on distance-based record linkage, which is further described in the next section.

2.1. Distance-based record linkage

In distance-based record linkage, the determination of parameters is not easy. Its main point is the definition of a distance. Nevertheless, different distances can be defined, each obtaining different results. Different distances have been considered and tested in the literature. We review the most relevant ones below.

We will use $V_1^X,\ldots,V_n^X$ and $V_1^Y,\ldots,V_n^Y$ to denote the set of variables of file X and Y, respectively. Using this notation, we express the values of each variable of a record a in X as $a=(V_1^X(a),\ldots,V_n^X(a))$ and of a record b in Y as $b=(V_1^Y(b),\ldots,V_n^Y(b))$. $\overline{V_i^X}$ corresponds to the mean of the values of variable X.

Euclidean (DBRL1): The Euclidean distance is used for attribute-standardized data. Accordingly, the distance between two records a and b is defined by:

$$d(a,b)^2=\sum_{i=1}^{n}\left(\frac{V_i^X(a)-\overline{V_i^X}}{\sigma(V_i^X)}-\frac{V_i^Y(b)-\overline{V_i^Y}}{\sigma(V_i^Y)}\right)^2$$

Euclidean (DBRL2): The Euclidean distance is used for distance-standardized data. Formally, the distance is defined as follows:

$$d(a,b)^2=\sum_{i=1}^{n}\left(\frac{V_i^X(a)-V_i^Y(b)}{\sigma(V_i^X-V_i^Y)}\right)^2$$

Mahalanobis (DBRLM): The Mahalanobis distance is used and applied to the original data with no standardization.

$$d(a,b)^2=(a-b)'[Var(V^X)+Var(V^Y)-2Cov(V^X,V^Y)]^{-1}(a-b)$$

where $Var(V^X)$ is the variance of attributes V^X, $Var(V^Y)$ is the variance of attributes V^Y and $Cov(V^X,V^Y)$ is the covariance between attributes V^X and V^Y.

The computation of $Cov(V^X,V^Y)$ poses one difficulty: how records in X are lined up with records in Y to compute the covariances. Two approaches have been considered in the literature.

DBRLM-COV: In a worst case scenario, it would be possible to know the correct links (a,b). Therefore, the covariance of attributes might be computed with the correct alignment between records.

DBRLM-COV0: It is not possible to know a priori which are the correct matches between pairs of records. Therefore, any pair of records (a,b) are feasible. If any pair of records (a,b) are considered, the covariance is zero.

Kernel (KDBRL): A Kernel-distance is considered. That is, instead of computing distances between records (a,b) in the original n dimensional space, records are compared in a higher dimensional space H. Thus, let $\Phi(x)$ be the mapping of x into the higher space. Then, the distance between records a and b in H is defined as follows:

$$\begin{aligned}d(a,b)^2&=\|\Phi(a)-\Phi(b)\|^2=(\Phi(a)-\Phi(b))^2\\&=\Phi(a)\cdot\Phi(a)-2\Phi(a)\cdot\Phi(b)+\Phi(b)\cdot\Phi(b)\\&=K(a,a)-2K(a,b)+K(b,b)\end{aligned}$$

where K is a kernel function (i.e. $K(a,b)=\Phi(a)\cdot\Phi(b)$).

Experiments have considered kernel functions of the form $K(x,y)=(1+x\cdot y)^d$ for $d>1$. Note that with $d=1$, the kernel record-linkage reduces to the distance-based record linkage with the Euclidean distance.

Taking all this into account, the distance between a and b is defined as:

$$d(a,b)^2=K(a,a)-2K(a,b)+K(b,b)$$

with a kernel function K.

Categorical data: distance-based record linkage for categorical data is not very widespread in the literature. In Ref. [13], the following distances are considered.

For nominal variables distance between records a and b is considered as

$$d(a,b)^2=\begin{cases}0 & \text{if } a=b\\ 1 & \text{if } a\neq b\end{cases}$$

If ordinal variables are considered, given the total order operator $\leqslant_V$ over the range of the variables, and denoting the cardinality of the range as $D(V)$, then,

$$d(a,b)^2=\frac{|c:\ \min(a,b)\leqslant_V c\leqslant_V \max(a,b)|}{|D(V)|}$$

For a description and comparison of these distances as used for distance-based record linkage see Refs. [42,13].

In this paper we consider the parametrization of distance based record linkage using weights to express the importance of the variables in the linkage process. This will be achieved considering a variation of the Euclidean distance using a weighted distance as will be detailed in the next sections. Other distance functions could also be used, like for instance, the Mahalanobis one which could be appropriate in cases when there is an important correlation between variables.

3. Supervised learning for record linkage

In this paper we determine the best weights for achieving the best possible performance in record linkage. To do so, we assume that a particular parametrized distance is used and consider the problem of finding the optimal weights for such parametrization. In the following sections we introduce the parametric distance based on the weighted mean for record linkage and describe a supervised learning approach for the determination of such weights.

3.1. A parametric distance for record linkage

It is well known that the multiplication of the Euclidean distance by a constant will not change the results of any record linkage algorithm. Due to this, we can express the distance DBRL1 given in Section 2.1 as a weighted mean of the distances for the attributes.

In a formal way, we redefine DBRL1 as follows:

$$d(a,b)^2=\sum_{i=1}^{n}\frac{1}{n}\left(\frac{V_i^X(a)-\overline{V^X}_i(a)}{\sigma(V_i^X)}-\frac{V_i^Y(b)-\overline{V^Y}_i(b)}{\sigma(V_i^Y)}\right)^2$$

Now, defining

$$d_i^2(a,b)=\left(\frac{V_i^X(a)-\overline{V^X}_i(a)}{\sigma(V_i^X)}-\frac{V_i^Y(b)-\overline{V^Y}_i(b)}{\sigma(V_i^Y)}\right)^2$$

we can rewrite this expression as

$$d(a,b)^2=AM\left(d_1^2(a,b),\ldots,d_n^2(a,b)\right)$$

where AM is the arithmetic mean $AM(c_1,\ldots,c_n)=\sum_i c_i/n$.

In general, any aggregation operator $\mathbb{C}$ [46] might be used:

$$d(a,b)^2=\mathbb{C}(d_1^2(a,b),\ldots,d_n^2(a,b))$$

From this definition, it is straightforward to consider a weighted version of the DBRL1. Its definition is as follows:

Definition 1. Let $p=(p_1,\ldots,p_n)$ be a weighting vector (i.e. $p_i\geqslant 0$ and $\sum_i p_i=1$). Then, the weighted distance is defined as:

$$dWM_i^2(a,b) = WM_p\left(d_1^2(a,b),\ldots,d_n^2(a,b)\right)$$

where $WM = (c_1,\ldots,c_n) = \sum_i p_i \cdot c_i$.

The interest of this variation is that we do not need to assume that all the attributes are equally important in the re-identification. This would be the case if one of the attributes is a key-attribute, e.g. an attribute where $V_i^X = V_i^Y$. In this case, the corresponding weight would be assigned to one, and all the others to zero. Such an approach would lead to 100% of re-identifications.

Moreover, as we will see later, this definition permits us to apply a supervised learning approach to determine the parameters of the method. In this way, we can tune the distance to have a better performance.

Although in this definition, the parametrized distance is applied to continuous variables, its extension to any type of variable or attribute where a distance function $d(a,b)$ between them could be defined is straightforward. As we will see, we focus our work in continuous variables, but the procedure for assessing other types of variables will be substantially the same.

3.2. Determining the optimal weights

For the sake of simplicity, we presume that each record of A, $(a_1,\ldots,a_N)$, is the protected record of B, $(b_1,\ldots,b_N)$, where N is the total number of records. That is, files are aligned. Then, if $V_k(a_i)$ represents the value of the kth variable of the ith record, we will consider the sets of values $d(V_k(a_i), V_k(b_j))$ for all pairs of records a_i and b_j.

Then, record i is correctly linked using aggregation operator $\mathbb{C}$ when the aggregation of the values $d(V_k(a_i), V_k(b_i))$ for all k is smaller than the aggregation of the values $d(V_k(a_i), V_k(b_j))$ for all $i \neq j$. That is,

$$\mathbb{C}(d(V_1(a_i),V_1(b_i)),\ldots,d(V_n(a_i),V_n(b_i))) < \mathbb{C}(d(V_1(a_i),V_1(b_j)),\ldots,d(V_n(a_i),V_n(b_j))) \quad (1)$$

for all $i \neq j$. Then, the optimal performance of record linkage is achieved when this equation holds for all records i.

To formalize the optimization problem and permit that the solution violates some equations we consider the equation in blocks. We consider a block as the set of equations concerning record i. That is, we define a block as the set of all the distances between one record of the original data and all the records of the protected data. Therefore, we have so many K as the number of rows of our original file. Besides, we need a constant C that multiplies K to avoid the inconsistencies and satisfy the constraint (given by the inequality (1)).

The rationale of this approach is as follows. The variable K indicates, for each block, if all the corresponding constraints are accomplished ($K = 0$) or not ($K = 1$). Then, we want to minimize the number of blocks non compliant with the constraints. Then, in this way, we can find the best weights that minimize the number of violations, or in other words, we can find the weights that maximize the number of re-identifications between the original and protected data.

Note that if for a record i, inequality (1) is violated for a certain record j. Then, it does not matter that other records j also violate the same Equation for the same record i. This is so because record i will not be re-identified.

Using these variables, K_i and the constant C are defined as follows:

$$\mathbb{C}(d(V_1(a_i),V_1(b_j)),\ldots,d(V_n(a_i),V_n(b_j))) - -\mathbb{C}(d(V_1(a_i),V_1(b_i)),\ldots, d(V_n(a_i),V_n(b_i))) + CK_i > 0 \quad (2)$$

for all $i \neq j$.

The constant C is used to express the *minimum distance* we require between the correct link and the other incorrect links. The larger it is, the more the correct links are distinguished from the incorrect links.

Using the constraints of the form above, and taking into account what has been explained before, the problem to minimize is as follows.

$$\text{Minimize} \quad \sum_{i=1}^{N} K_i \quad (3)$$

$$\text{Subject to}: \sum_{i=1}^{N}\sum_{j=1}^{N} dWM_i^2 \times (d(V_1(a_i),V_1(b_j)),\ldots,d(V_n(a_i),V_n(b_j))) - (d(V_1(a_i),V_1(b_i)),\ldots,d(V_n(a_i),V_n(b_i))) + CK_i > 0 \quad (4)$$

$$K_i \in \{0,1\} \quad (5)$$

$$\sum_{i=1}^{n} p_i = 1 \quad (6)$$

$$p_i \geqslant 0 \quad (7)$$

where N is the number of records, n the number of variables, dWM_i^2 is the parametrized distance defined in Section 3.1, and $p = (p_1,\ldots,p_n)$ is a weighting vector used in this distance. This problem is a linear optimization problem with linear constraints and the (global) optimum solution can be found with an optimization algorithm.

3.3. Implementation details

In order to minimize the number of non linked records and determine the optimal weights of the problem defined above, we use the simplex optimizer algorithm from the IBM ILOG CPLEX tool [25] (version 12.1).

The problem is first expressed into the MPS (Mathematical Programming System) format, and then, processed with the optimization solver.

If N is the number of records, and n the number of variables of the two data sets X and X'. We have N terms of K_i in the objective function, that is N variables for Eq. (3). The total number of constraints in the optimization problem is $N^2 + N + 1 + n$. There are N^2 constraints from Eq. (4), N for Eq. (5), 1 for Eq. (6), and n for Eq. (7).

4. Results and evaluation

To evaluate our proposal, we have applied the record linkage approach described in Section 3.2 to the data produced by several protection methods.

We test the record linkage with the first 7 variables, and 400 randomly selected records from the 1080 records of Census data (see Section 4.1). Each execution has been performed 10 times, and their means are given as a result. Given the random selection of records in each execution, making 10 executions allows us to take into account all the records.

The tests have been executed in three different machines. One Intel Core 2 6400 at 2.13 GHz, and one Intel Core 2 Quad Q9400 at 2.66 GHz, both with 4 GB of memory and a GNU/Linux 2.6 64 bits. The third one is the Finis Terrae computer [3] composed of 142 HP Integrity rx7640 computing nodes with 16 Itanium Montvale

[3] Centro de Supercomputacin de Galicia, http://www.cesga.es.

cores and 128 GB of memory each, one HP Integrity Superdome node, with 128 Itanium Montvale cores and 1.024 GB of memory, and 1 HP Integrity Superdome node, with 128 Itanium 2 cores and 384 GB of memory. From the Finis Terrae computer we used 16 cores.

Table 1
Attributes of the census dataset. All of them are real valued numeric attributes.

V_1	AFNLWGT	Final weight (2 implied decimal places)
V_2	AGI	Adjusted gross income
V_3	EMCONTRB	Employer contribution for health insurance
V_4	ERNVAL	Business or farm net earnings
V_5	FEDTAX	Federal income tax liability
V_6	FICA	Social security retirement payroll deduction
V_7	INTVAL	Amount of interest income
V_8	PEARNVAL	Total person earnings
V_9	POTHVAL	Total other persons income
V_{10}	PTOTVAL	Total person income
V_{11}	STATETAX	State income tax liability
V_{12}	TAXINC	Taxable income amount
V_{13}	WSALVAL	Amount: total wage & salary

Note that we focus our work in distance-based record linkage, by comparing our proposal to the standard distance-based record linkage. This allows us to test our proposal directly with very similar approaches, which use the same techniques and strategies. Furthermore, a comparison with probabilistic record linkage is not considered for two reasons. As described in Section 4.1 we work with numerical data. For such data, distance-based record linkage is more appropriate than probabilistic-based record linkage as shown in Ref. [43]. Moreover, in Ref. [13] it is concluded that both distance-based and probabilistic record linkage provide very similar results. Those results where based on experiments performed using the same dataset we use in our experiments (described in Section 4.1). With this precedents we can expect probabilistic record linkage to perform very closely to standard distance-based record linkage.

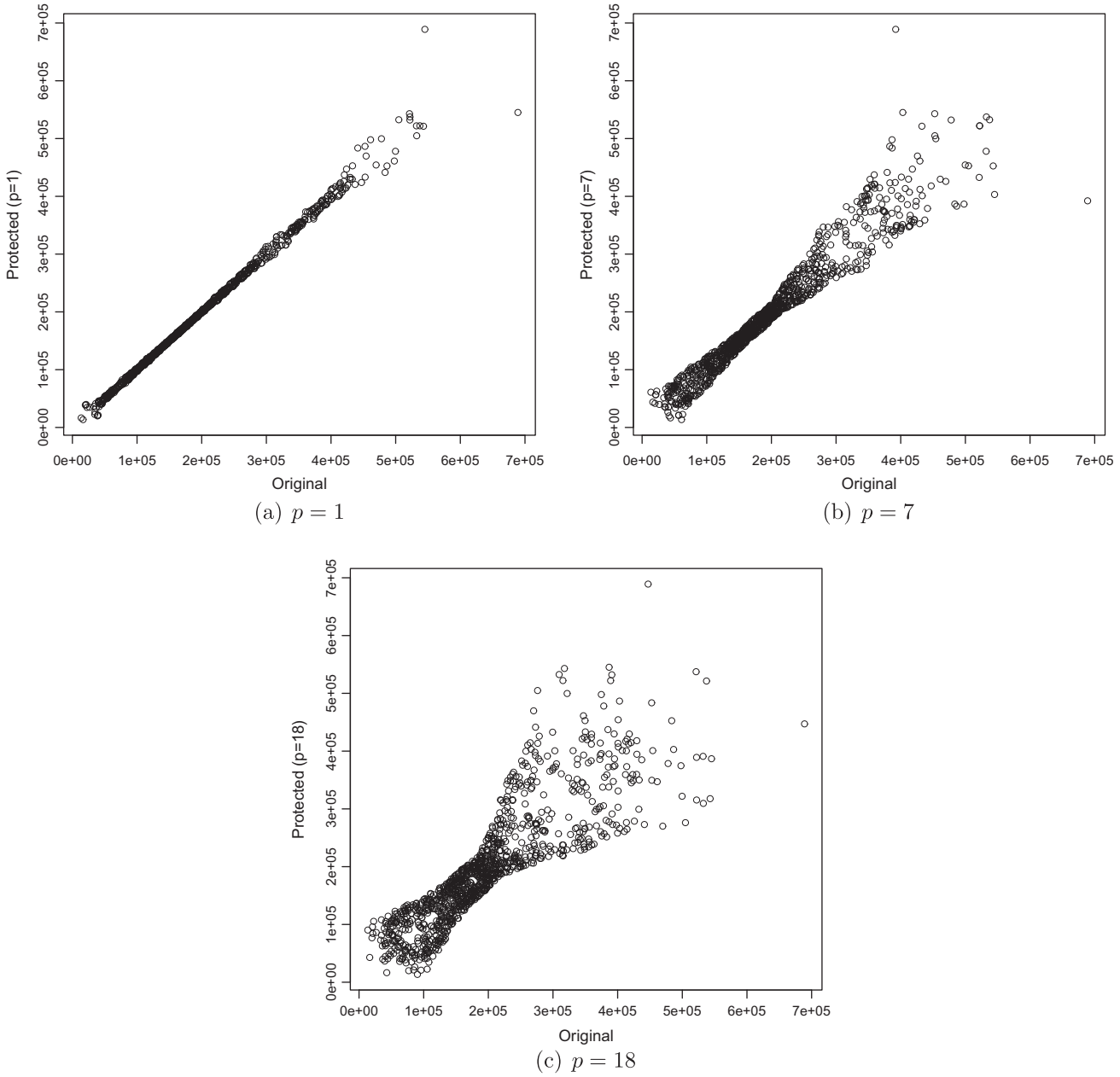

Fig. 1. Comparing protected and non-protected values for variable V_1 using rank swapping for different values of p.

4.1. Testing dataset

We have used the Census dataset from the European CASC project [5], which contains 1080 records and 13 variables, and has been extensively used in other works [12,32,11,53,9].

The dataset was extracted using the Data Extraction System of the US Census Bureau [48]. The records correspond to the *Current Population Survey of the year 1995*, and more precisely to the filegroup *March Questionnaire Supplement – Person Data Files*. All the records with zero or missing values for at least one of the 13 attributes were discarded giving up the final 1080 records. These attributes, all numerical real valued, are described in Table 1.

This dataset has some interesting properties. Records where selected so the number of repeated values was low. Furthermore the 13 variables were selected so values of each one span a wide range.

As previously mentioned, we selected the first 7 variables. Four of them V_1, V_2, V_3, and V_5 have no repeated values. Regarding this issue we wanted to provide a generic record linkage process, so approximately half of the variables had repeated values. Selecting all 7 variables without repeated values, could provide better results, it the scenario will be less realistic, since repeated values are normally expected in this kind of data.

4.2. Protection methods evaluated

The same original dataset (Census) is protected with the following techniques: microaggregation, rank swapping, and additive noise. These methods are described in more detail below.

- *Microaggregation* (*Mic*): it provides privacy by means of clustering the data into small clusters and then replacing the original data by the centroids of the corresponding clusters. Privacy is achieved because all clusters have at least a predefined number of elements, and therefore, there are at least k records with the same value. Note that all the records in the cluster replace a value by the value in the centroid of the cluster. The constant k is a parameter of the method that controls the level of privacy. The larger the k, the more privacy we have in the protected data.
  Microaggregation was originally [7] defined for numerical attributes, but later extended to other domains. For example, to categorical data in Ref. [44] (see also Ref. [11]), and in constrained domains in Ref. [45].
  For the protection of the Census dataset we have used the Euclidean distance to form the clusters, and the arithmetic mean to compute the centroids. We have considered the following variants of microaggregation:
  - Individual ranking (*MicIR*).
  - Z-scores projection (*MicZ*).
  - Principal components projection (*MicPCP*)
  - Multivariate microaggregation:
- Two variables at a time (*Mic2*).
- Three variables at a time (*Mic3*).
- Four variables at a time (*Mic4*).
- Five variables at a time (*Mic5*).
- ix variables at a time (*Mic6*).
- All variables at a time (*MicAll*).

Values of k from 3 to 20 have been considered.

- *Rank swapping*: the values of a variable V_i are ranked in ascending order; then each ranked value of V_i is swapped with another ranked value randomly chosen within a restricted range (e.g. the rank of two swapped values cannot differ by more than p percent of the total number of records). The method was first described for numerical variables in Ref. [34]. We consider values of p from 1 to 6.
- *Additive noise*: Gaussian noise is added to the original data to get the masked data [4]. If the standard deviation of the original variable is σ, noise is generated using a $N(0, p\sigma)$ distribution. We consider values of p from 1 to 16.

To show how the protection methods affect or distort the original data, we provide three plots in Fig. 1 which compare the protected and original values of variable V_1 using the Rank Swapping protection method. The original value is shown (axis x) *versus* the protected value (axis y) for different values of the parameter p.

4.3. Improvement of standard distance-based record linkage

In Tables 2–4 we show the average difference after 10 executions between the true positive rates (percentage of re-identified records) of the weighted mean with optimal weights (dWM^2) and the standard record linkage (from now on denoted as *DBRL*, which corresponds to *DBRL*1 from Section 2.1). We also provide the standard deviation (σ) and the standard error (ϵ), computed as $\epsilon = \sigma/\sqrt{(10)}$, where σ is the standard deviation of the 10 executions. The record linkage has been performed on a training set of 400 records, as described in Section 3.2, and then tested with the

Table 2
Re-identification percentages in the training set for the *RankSwap* protection method. In each case re-identification with both the standard record linkage (*DBRL*) and our proposal (dWM^2) are shown, with the standard deviation (σ) and standard error (ϵ), for 10 execution. Here, p is the percent difference allowed in ranks.

RankSwap						
p	*DBRL*	σ(*DBRL*)	ϵ(*DBRL*)	dWM^2	$\sigma(dWM^2)$	$\epsilon(dWM^2)$
1	99.65	0.00253	0.0008	100	0.00350	0.00111
2	98.525	0.00518	0.00164	99.725	0.00817	0.00258
3	96.975	0.00944	0.00298	98.85	0.00915	0.00289
4	94.65	0.00921	0.00291	97.15	0.00908	0.00287
5	92.85	0.01715	0.00542	95.325	0.01364	0.00431
6	88	0.01668	0.00527	90.825	0.01458	0.00461
Avg.	95.10833	0.01003	0.00317	96.97917	0.00969	0.00307

Table 3
Re-identification percentages in the training set for the *Mic2* protection method. In each case re-identification with both the standard record linkage (*DBRL*) and our proposal (dWM^2) are shown, with the standard deviation (σ) and standard error (ϵ), for 10 execution. Here, k is the minim cluster size for the microaggregation.

Mic2						
k	*DBRL*	σ(*DBRL*)	ϵ(*DBRL*)	dWM^2	$\sigma(dWM^2)$	$\epsilon(dWM^2)$
3	99.975	0.00079	0.00025	100	0	0
4	99.65	0.00269	0.00085	99.9	0.00211	0.00067
5	99.3	0.00511	0.00162	100	0	0
6	99.275	0.00463	0.00147	99.7	0.00329	0.00104
7	99.35	0.00412	0.00130	99.825	0.00265	0.00084
8	98.15	0.00580	0.00183	99.7	0.00284	0.00090
9	98.425	0.00528	0.00167	99.525	0.00322	0.00102
10	98.375	0.00377	0.00119	99.425	0.00206	0.00065
11	97.2	0.00633	0.002	98.725	0.00381	0.00121
12	96.9	0.00592	0.00187	98.525	0.00343	0.00108
13	96.775	0.00786	0.00249	98.375	0.00580	0.00184
14	96.525	0.00924	0.00292	98.1	0.00615	0.00194
15	95.875	0.00637	0.00202	97.975	0.00448	0.00142
16	95.85	0.00709	0.00224	98.15	0.00648	0.00205
17	94.5	0.01041	0.00329	96.75	0.00920	0.00291
18	93.475	0.00901	0.00285	96.175	0.00727	0.00230
19	92.925	0.01444	0.00457	95.325	0.01155	0.00365
20	92.425	0.01068	0.00338	94.95	0.00633	0.002
Avg.	96.94167	0.00664	0.00210	98.39583	0.00474	0.00142

Table 4
Re-identification percentages in the training set for the *Noise* protection method. In each case re-identification with both the standard record linkage (*DBRL*) and our proposal (dWM^2) are shown, with the standard deviation (σ) and standard error (ϵ), for 10 execution. Here p is the parameter of the additive noise (see Section 4.2).

Noise						
p	*DBRL*	$\sigma(DBRL)$	$\epsilon(DBRL)$	dWM^2	$\sigma(dWM^2)$	$\epsilon(dWM^2)$
1	100	0	0	100	0	0
2	100	0	0	100	0	0
4	100	0	0	100	0	0
6	99.875	0.00177	0.00056	100	0	0
8	99.45	0.00369	0.00117	99.9	0.00129	0.00041
10	98.05	0.00633	0.002	99.1	0.00428	0.00135
12	95.6	0.00637	0.00201	97.05	0.00387	0.00123
14	93.85	0.00899	0.00284	95.45	0.00550	0.00174
16	90.025	0.01102	0.00349	92.3	0.00654	0.00207
Avg.	97.42778	0.00424	0.00134	98.2	0.00239	0.00076

same 400 records. We show the results for the protection methods rank swapping (*RankSwap*) in Table 2, microaggregation (*Mic2*) in Table 3, and additive noise (*Noise*) in Table 4 as an example. The average for each protection method is also given as *avg*.

As it can be appreciated, our proposed method achieves an improvement with respect to the standard distance record linkage. The improvement is however relatively small (about 5%). This leads us to conclude that it is relatively meaningful to use equal weights for estimating the disclosure risk in the scenarios discussed here, especially if we take into account the computation cost (see Section 4.5). A maximum error of 5% of risk can be taken into account by the office before releasing the protected data.

We have also seen that the improvement obtained with our method is related to the percentage of the re-identification. In general for lower or high re-identification percentages the improvement is very low, while for medium percentages, the re-identi fication percentage increases. Fig. 2 shows the difference on the re-identification percentage between our record linkage (using dWM^2) and the standard one (*DBRL*) in accordance with the number of non-identified links. Note that Fig. 2 gives the difference for all computed cases. That is each protection method with each parameter computed 10 times. The main idea of this figure is to show the general behavior of all protection methods. In order to see more clear results, in Fig. 3 we show only two protection methods, *MicAll* and *Mic3*.

The fact that the percentages of improvement of our method depends on the number of re-identified links, has to be taken into account with the computation time analysis presented in Section 4.5.

4.4. Identification of key-attributes

Our method to find the optimal weights in the record linkage, provides at the same time information regarding the relevance of each variable or attribute for the linkage. That is, the attribute with the highest weight value is the one that provides more useful information for the linkage.

This means that we can establish a direct correspondence between the weights associated to each attribute with its disclosure risk, providing thus a disclosure risk estimation for each individual attribute. For example, an attribute with a high weight has a greater disclosure risk. Statistical agencies can then apply specific protection methods to concrete attributes if their individual disclosure risk is greater than expected in relation to the other attributes of the data set.

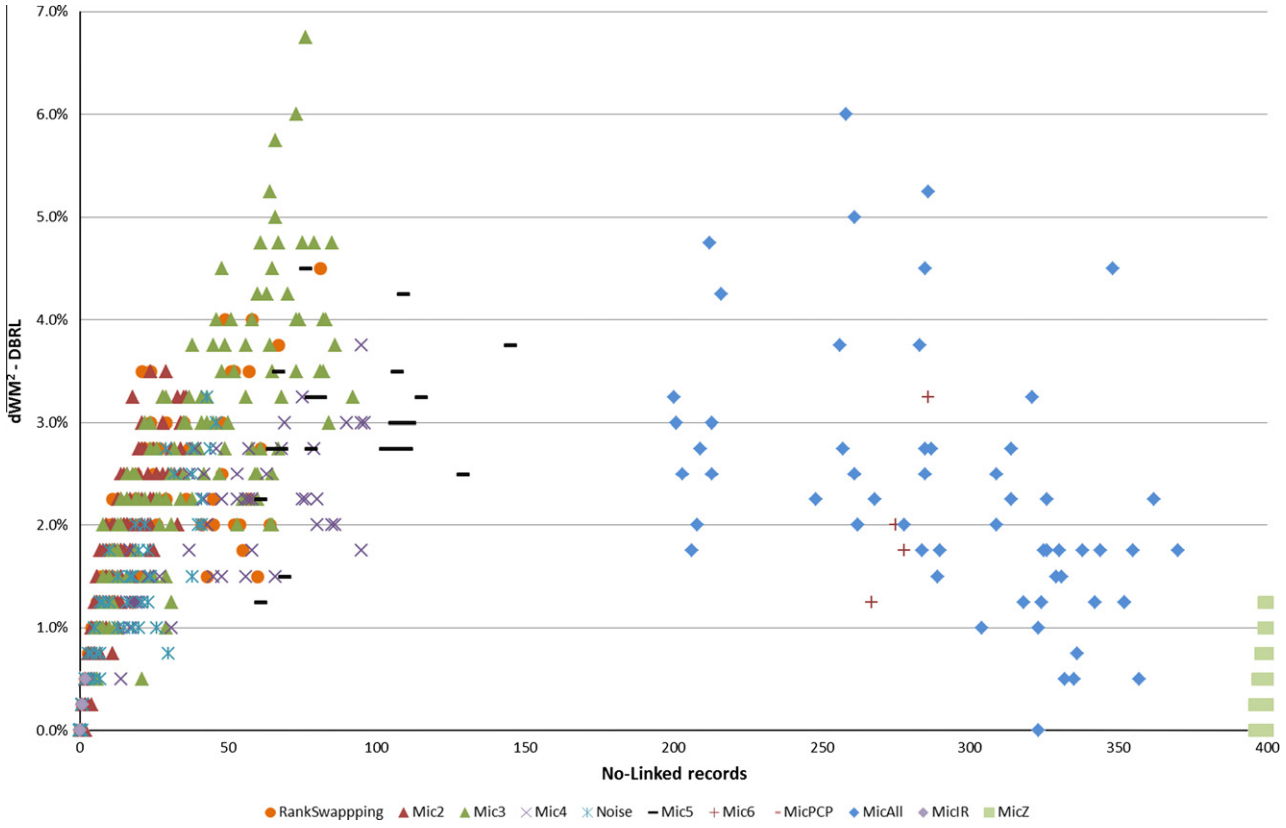

Fig. 2. Improvement for all cases, shown as the difference between the standard record linkage, *DBRL*, and our proposal using dWM^2.

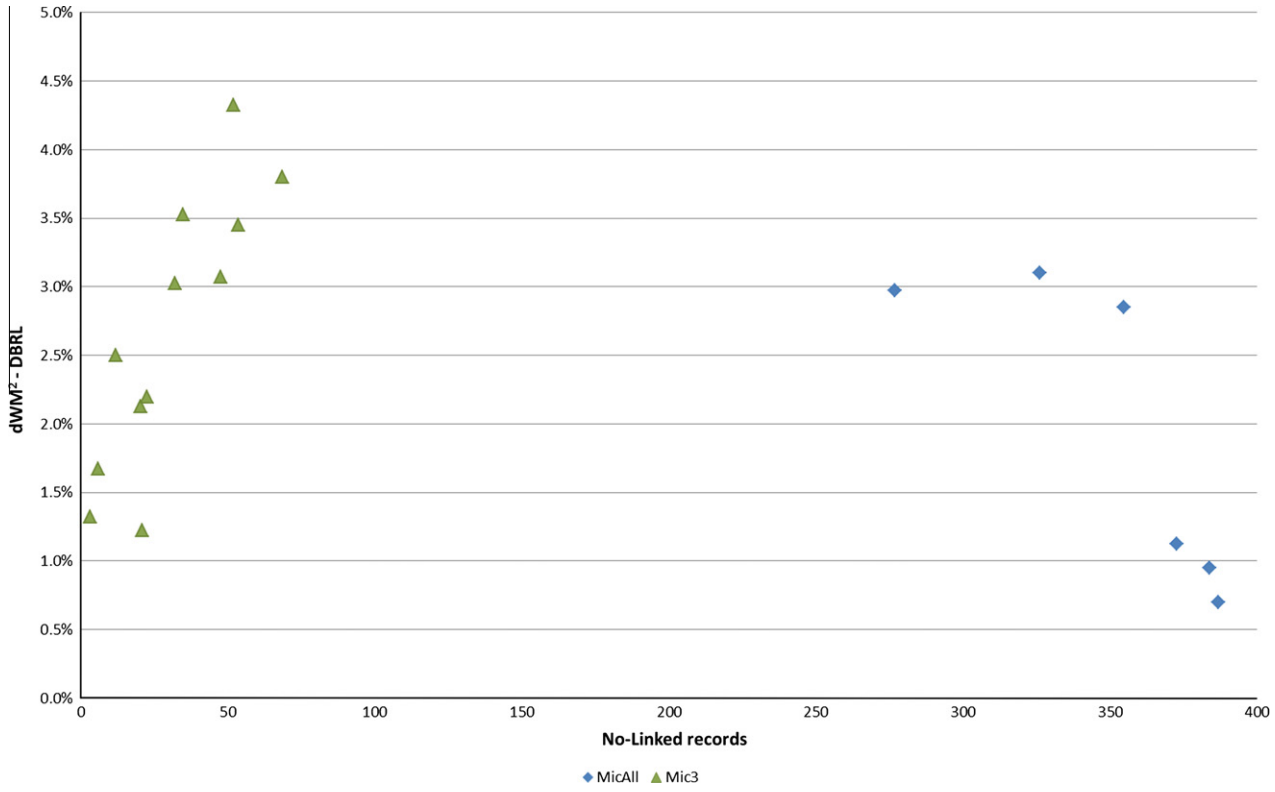

Fig. 3. Improvement for *MicAll* and *Mic3*, shown as the difference between the standard record linkage, *DBRL*, and our proposal using dWM^2. Each dot is the average of 10 executions.

Table 5
Weights identifying key-attributes for a file protected with additive noise, where each variable is protected with different values for the parameter p.

Variable	V_1	V_2	V_3	V_4	V_5	V_6	V_7
p	1	2	3	4	5	6	7
Weight	0.97970	0.01484	0.00500	0.0	0.0	0.00046	0.0

Table 6
Re-identification percentages using single variables for a file protected with additive noise, with different values of p for each variable.

Variable	V_1	V_2	V_3	V_4	V_5	V_6	V_7
p	1	2	3	4	5	6	7
Re-identification	29.5%	14.75%	10.5%	6.75%	7%	4.25%	4%

As an example, we consider the case of the original Census data set protected with additive noise. Unlike previous tests, in this case we use different protection parameter for each attribute: attribute V_1 with $p = 1$, V_2 with $p = 2$, V_3 with $p = 3$, and so on. Table 5 shows for each attribute, the weights obtained with our method, and the parameter p of the additive noise used to protect the attribute.

As expected, V_1 is the attribute with a clear higher weight since it is the variable with lower perturbation, and thus, the one that provides better information for the record linkage. Moreover we can attempt to perform the re-identification with single variables. That is, we test the distance-based record linkage using only one variable each time. The results shown in Table 6, show that the re-identification percentages of each variable separately closely relate to the weights previously obtained. It is also interesting to note that single-variable record linkage obtains very poor re-identification results as compared to the record linkage with all 7 variables, which gives a 100% of correct matches.

This approach to identify key records can be compared to the Special Uniques Identification problem [17,16], which identifies records with a high risk of re-identification in a microdata file. In our case, we do not identify the risky records, but the risky variables.

4.5. Considerations on the computation cost

Our experiments also show an interesting behavior regarding the computation cost of the optimal parameter for the distance dWM^2. Fig. 4 shows the time taken to find the solution of 10 executions for each protection method considered. That is, we show the time taken to find the optimal weights (time is given in a logarithmic scale) for all the protection methods, with respect to the number of unlinked records. Note that there are 10 execution for each case. This figure shows the general behavior found in all protection methods. To have a more detailed view, in Fig. 5 we show the average time of 10 executions for the protection methods *MicAll* and *Mic3*.

We can observe that the computation cost depends on the percentage of re-identifications (as determined by the objective function in the x axis). With low and high percentages of re-identifications the cost is very low, even negligible as the percentages reach 0% or 100%. At the same time with medium number of the re-identifications there is a high computational cost, which reaches more than one week for some cases.

Combining these results with the ones described before (cf. Section 4.3) about the improvement of record linkage, we have that a significant improvement of record linkage occurs precisely for the data files where the computational cost is high (1 week or more of computational cost).

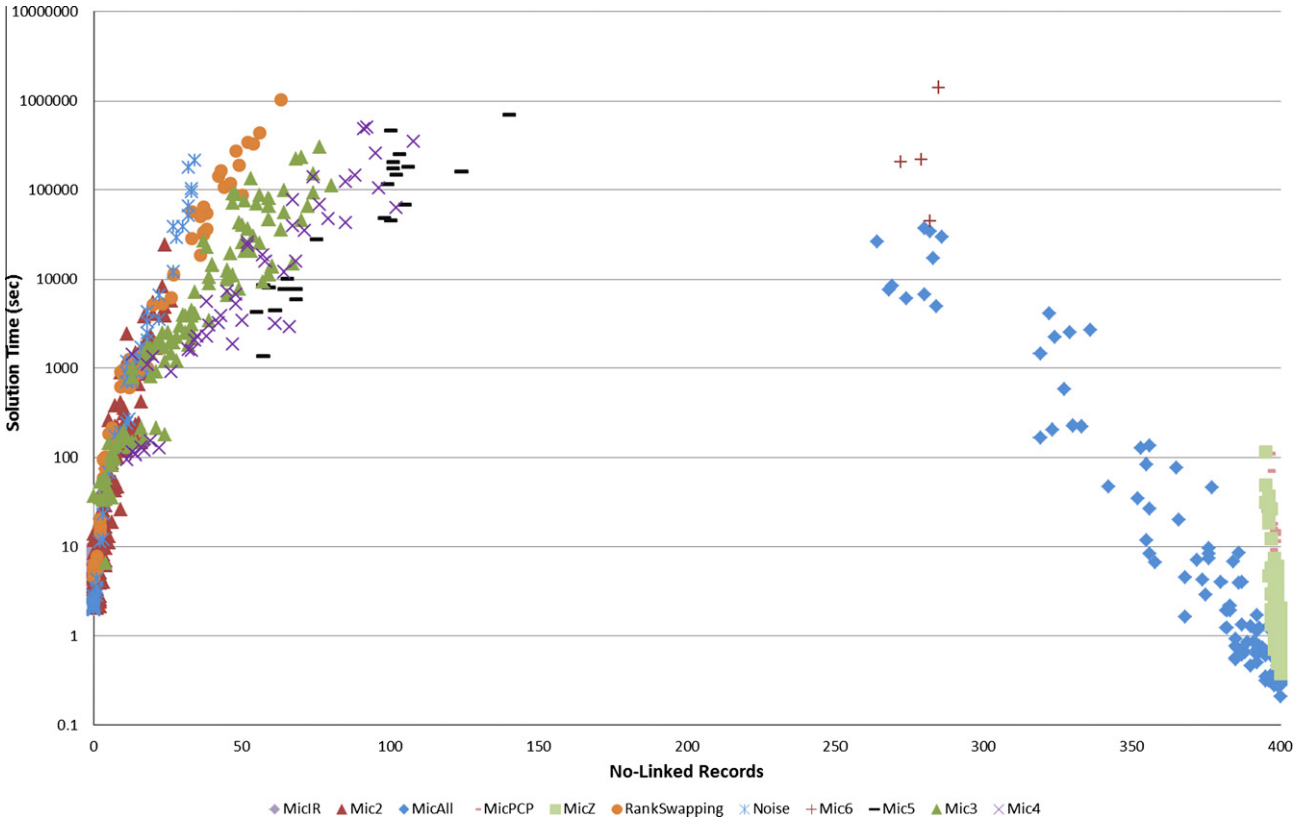

Fig. 4. Computation time for all cases, in terms of the number of non matched links (objective function).

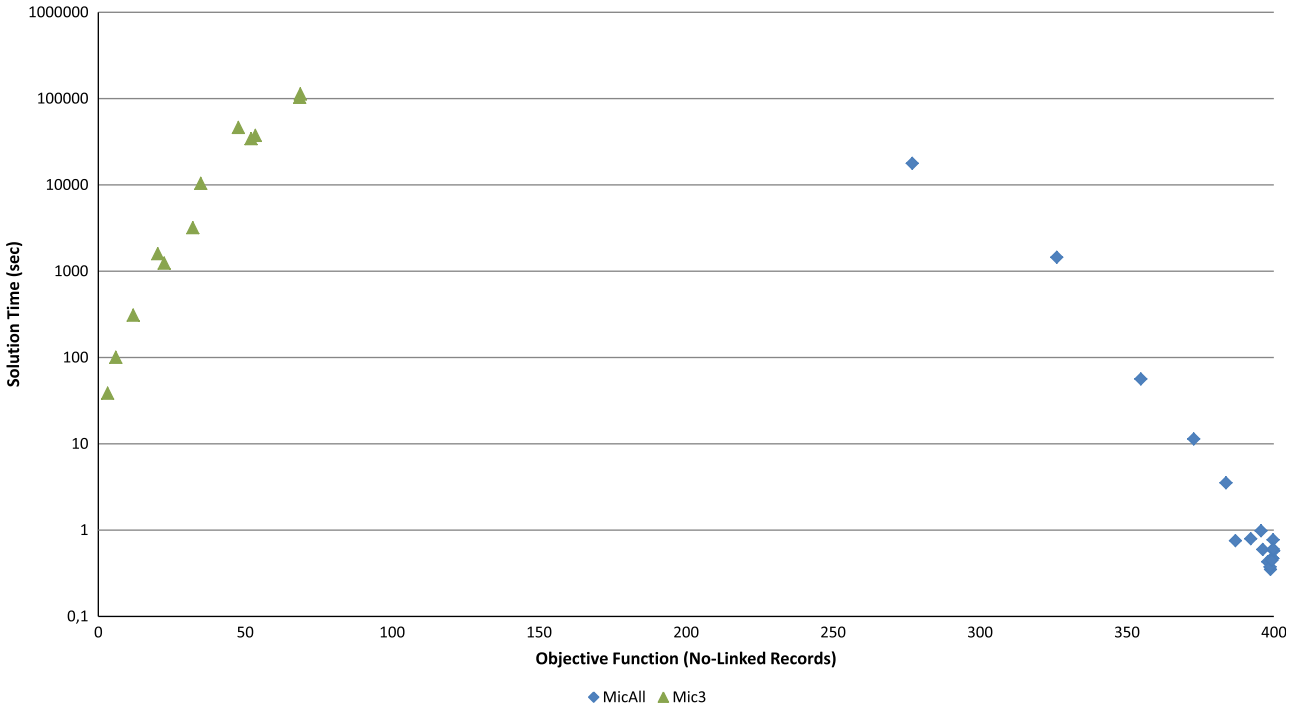

Fig. 5. Computation time for *MicAll* and *Mic3* as the average of 10 executions, in terms of the number of non matched links (objective function).

4.6. Improvements on non-uniformly protected files

As we have seen in the general case, determining optimal weights for the distance-based record linkage does not provides a substantial improvement in the re-identification percentage. There are some cases, where the fact that some attributes are more weighted can have an important impact in the re-identification, these can be seen with non-uniformly protected

Table 7
Re-identification percentage of non-uniformly microaggregated files in the training set.

	DBRL	dWM^2
Mic553-2.8.5	42.025	90
Mic553-8.2.5	25.75	82.6
Mic553-5.3.5	30.75	82.375
Mic2236-8.3.10.5	82.725	97.5

Table 9
Re-identification percentage of non-uniformly microaggregated files for a training set of 400 and testing set of 500 records.

	DBRL	dWM^2
Mic553-2.8.5	39.38	86.94
Mic553-8.2.5	23.66	78.48
Mic553-5.3.5	28.84	77.6
Mic2236-8.3.10.5	80.82	96.46

files. That is, files where some attributes have a higher protection degree than others. This means that some attributes have less perturbation and thus are more useful for re-identification than others.

To illustrate this issue, we have tested our proposal with the same Census dataset microaggregated with different values of k for different groups of attributes in the same file. Similarly to the case described in Section 4.4, we have considered the following files with the given groups of attributes or variables and the given values of k for each one:

- *Mic553-2.8.5*: 3 groups of 5, 5, and 3 attributes, with respective values for k of 2, 8, and 5.
- *Mic553-8.2.5*: 3 groups of 5, 5, and 3 attributes, with respective values for k of 8, 2, and 5.
- *Mic553-5.3.5*: 3 groups of 5, 5, and 3 attributes, with respective values for k of 5, 3, and 5.
- *Mic2236-8.3.10.5*: 4 groups of 2, 2, 3, and 6 attributes, with respective values for k of 8, 3, 10, and 5.

Table 7 shows the re-identification percentage in the training set (400 records randomly selected, giving the average of 10 executions) for the weighted mean with optimal weights (dWM^2) and the standard record linkage (*DBRL*). The table shows that the improvement achieved by the dWM^2 is very important, in some cases more than 50%.

The resulting optimal weight for each attribute clearly reflects the protection applied to each attribute. Table 8 shows the optimal weights for the 7 variables considered in the record linkages process (as described in Section 4.3). In general the weight assigned to each attribute increases when the k corresponding to the attribute is lower. For a lower k we have more weight (clearly illustrated for attribute V_1).

We have also considered another scenario, where we use 400 records for training and 500 different record for testing. Table 9 shows the re-identification percentages in this case. As expected, the dWM^2 provides a better performance than the standard record linkage, *DBRL*. Note also that re-identification percentages are lower if compared to the results using only the training set (see Table 7). Although the difference is not as big as one could expect, this supports the use of the training set as measure of disclosure risk, since it evaluates the worst case. That is, an upper bound or maximum of the disclosure risk using distance-base record linkage.

5. Conclusions

In this paper we have presented and studied the parametrization of distance based record linkage, in the context of data privacy. This is done by extending the Euclidean distance used in standard record linkage with a weighted mean. We have provided a supervised learning approach to determine the optimum weights for such distance, which express the importance of each variable in the linkage process. We have extensively tested our approach with several data sets.

In data privacy and statistical disclosure control, record linkage is used as a disclosure risk estimation of the protected data. This estimation is based on the links between records of the original data and the protected data, that the record linkage method can find. We have tested our approach with some of the most common protection techniques in statistical disclosure control: microaggregation, rank swapping, and additive noise.

Our results show an improvement in the linkage as compared to standard distance based record linkage. Nevertheless, in the general case, the improvement is small, which leads us to conclude that it is relatively meaningful to use equal weights for estimating the disclosure risk in the scenarios discussed here. The low increment in the proportion of correct links can be assumed by the statistical offices as a small increment on the risk of protected data computed using equal weights. There is a concrete case where the increment is very important, though. This is when the attributes of the protected file have different protection degrees. That is, some attributes are more protected (and thus more distorted) than others.

We also show that the computational cost needed to determine the optimal weights, in the general case, depends on the re-identified links (success or failure in the re-identification). For low or high re-identification percentages the cost is negligible, while for medium percentages of re-identification the cost is very high. These results on the computational cost has implications on the analysis of risk, because as shown, a high cost is needed for a small improvement on the performance of an attack.

Table 8
Optimal weights for the non-uniformly microaggregated files.

Attr	V_1	V_2	V_3	V_4	V_5	V_6	V_7
k	2	2	2	2	2	8	8
Mic553-2.8.5	0.84562	0.04475	0.00178	0.00742	0.04884	0.00216	0.04943
k	8	8	8	8	8	2	2
Mic553-8.2.5	0.46971	0.01258	0.00055	0.00033	0.01876	0.01382	0.48425
k	5	5	5	5	5	3	3
Mic553-5.3.5	0.68323	0.01825	0.00123	0.00080	0.02919	0.00424	0.26305
k	8	8	3	3	10	10	10
Mic2236-8.3.10.5	0.39155	0.01172	0.05456	0.39945	0.08340	0.00052	0.05880

Acknowledgments

Partial support by the Spanish MICINN (projects TSI2007-65406-C03-02, TIN2010-15764, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004), and European Commission (project Data without Boundaries (DwB), Grant Agreement Number 262608) is acknowledged.

References

[1] W. Alvey, B. Jamerson (Eds.), Record Linkage Techniques – 1997, Proceedings of an International Workshop and Exposition, Federal Committee on Statistical Methodology, Office of Management of the Budget, 1997.
[2] C. Batini, M. Scannapieco, Data Quality – Concepts, Methodologies and Techniques Series: Data-Centric Systems and Applications, 2006.
[3] T.R. Belin, H. Ishwaran, N. Duan, S. Berry, D. Kanouse, Identifying Likely Duplicates by Record Linkage in a Survey of Prostitutes, Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, Wiley, 2004.
[4] R. Brand, Microdata protection through noise addition, in: Inference Control in Statistical Databases, From Theory to Practice, Lecture Notes In Computer Science, vol. 2316, Springer, Berlin/Heidelberg, 2002, pp. 97–116.
[5] R. Brand, J. Domingo-Ferrer, J.M. Mateo-Sanz, Reference Datasets to Test and Compare SDC Methods for Protection of Numerical Microdata, Technical Report, European Project IST-2000-25069 CASC, 2002.
[6] M. Colledge, Frames and Business Registers: An Overview. Business Survey Methods, Series in Probability and Statistics, Wiley, 1995.
[7] D. Defays, P. Nanopoulos, Panels of enterprises and confidentiality: the small aggregates method, in: Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada, 1993, pp. 195–204.
[8] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society 39 (1) (1977) 1–38.
[9] J. Domingo-Ferrer, J.M. Mateo-Sanz, V. Torra, Comparing sdc methods for microdata on the basis of information loss and disclosure risk, in: Preproceedings of ETK-NTTS 2, Eurostat, 2001, pp. 807–826.
[10] J. Domingo-Ferrer, V. Torra, A Quantitative Comparison of Disclosure Control Methods for Microdata, 2001, pp. 111–133 of [14].
[11] J. Domingo-Ferrer, V. Torra, Ordinal, continuous and heterogeneous *k*–anonymity through microaggregation, Data Mining and Knowledge Discovery 11 (2) (2005) 195–212.
[12] J. Domingo-Ferrer, V. Torra, J.M. Mateo-Sanz, F. Sebe, Empirical disclosure risk assessment of the ipso synthetic data generators, in: Monographs in Official Statistics – Work Session On Statistical Data Confidentiality, Eurostat, 2006, pp. 227–238.
[13] J. Domingo-Ferrer, V. Torra, Validating distance-based record linkage with probabilistic record linkage, in: Topics in Artificial Intelligence, Lecture Notes in Computer Science, vol. 2504, Springer, Berlin/Heidelberg, 2002, pp. 207–215.
[14] P. Doyle, J. Lane, J. Theeuwes, L. Zayatz (Eds.), Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, Elsevier Science, 2001.
[15] H.L. Dunn, Record linkage, American Journal of Public Health 36 (12) (1946) 1412–1416.
[16] M.J. Elliot, A. Manning, R. Ford, A computational algorithm for handling the special uniques problem, International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 5 (10) (2002) 493–509.
[17] M.J. Elliot, A. Manning, The identification of special uniques, in: Proceedings of GSS Methodology Conference, 2001.
[18] A. Elmagarmid, G. Panagiotis, V. Verykios, Duplicate record detection: a survey, IEEE Transactions on Knowledge and Data Engineering 19 (1) (2007) 1–16.
[19] I. Fellegi, A. Sunter, A theory for record linkage, Journal of the American Statistical Association 64 (328) (1969) 1183–1210.
[20] L.E. Gill, OX-LINK: the Oxford medical record linkage system demonstration of the PC version, in: Record Linkage Techniques 1997, Proceedings of an International Workshop and Exposition, Federal Committee on Statistical Methodology, Office of Management of the Budget, 1997, pp. 491.
[21] S. Gomatam, M.D. Larsen, Record linkage and counterterrorism, Chance 17 (1) (2004) 25–29.
[22] H. Hartley, Maximum likelihood estimation from incomplete data, Biometrics 14 (1958) 174–194.
[23] T.N. Herzog, F.J. Scheuren, W.E. Winkler, Data Quality and Record Linkage Techniques, Springer, 2007.
[24] M. Houbiers, Towards a social statistical database and unified estimates at Statistics Netherlands, Journal of Official Statistics 20 (1) (2004) 55–75.
[25] IBM, IBM ILOG CPLEX, High-Performance Mathematical Programming Engine, International Business Machines Corp., 2010. <http://www-01.ibm.com/software/integration/optimization/cplex/>.
[26] 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.
[27] M.A. Jaro, Probabilistic linkage of large public health data files, Statistics in Medicine 14 (1995) 491–498.
[28] D. Krewski, A. Dewanji, Y. Wang, S. Bartlett, J.M. Zielinski, R. Mallick, The Effect of Record Linkage Errors on Risk Estimates in Cohort Mortality Studies, Survey Methodology, vol. 31(1), Statistics Canada, 2005. pp. 13–21.
[29] P. Lahiri, D. Larsen, Regression analysis with linked data, Journal of the American Statistical Association 100 (469) (2005) 222–230.
[30] D. Lambert, Measures of disclosure risk and harm, Journal of Official Statistics 9 (1993) 313–331.
[31] M.D. Larsen, D.B. Rubin, Iterative automated record linkage using mixture models, Journal of the American Statistical Association 96 (2001) 32–41.
[32] M. Laszlo, S. Mukherjee, Minimum spanning tree partitioning algorithm for microaggregation, IEEE Transactions on Knowledge and Data Engineering 17 (7) (2005) 902–911.
[33] G. McLachlan, T. Krishnan, The EM Algorithm and Extensions, Wiley Series in Probability and Statistics, John Wiley & Sons, 1997.
[34] R. Moore, Controlled Data Swapping Techniques for Masking Public Use Microdata Sets, US Bureau of the Census, 1996 (unpublished manuscript).
[35] H.B. Newcombe, J.M. Kennedy, S.J. Axford, A.P. James, Automatic linkage of vital records, Science 130 (1959) 954–959.
[36] H.B. Newcombe, Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business, Oxford University Press, 1998.
[37] J. Nin, J. Herranz, V. Torra, Rethinking rank swapping to decrease disclosure risk, Data & Knowledge Engineering 64 (1) (2008) 346–364.
[38] J. Nin, J. Herranz, V. Torra, On the disclosure risk of multivariate micro-aggregation, Data & Knowledge Engineering 67 (3) (2008) 399–412.
[39] 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.
[40] N.L. Spruill, Measures of confidentiality, in: Proc. Survey Research Section American Statistical Association, 1982, pp. 260–265.
[41] Statistics Canada, Record linkage at Statistics Canada, 2010. <http://http://www.statcan.gc.ca/record-enregistrement/index-eng.htm>.
[42] 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, vol. 4302, Springer, Berlin/Heidelberg, 2006, pp. 233–242.
[43] V. Torra, J. Domingo-Ferrer, Record linkage methods for multidatabase data mining, in: V. Torra (Ed.), Information Fusion in Data Mining, Springer, Berlin, 2003, pp. 99–130.
[44] V. Torra, Microaggregation for categorical variables: a median based approach, in: Proc. Privacy in Statistical Databases (PSD 2004), Lecture Notes in Computer Science, vol. 3050, Springer, Berlin/Heidelberg, 2004, pp. 162–174.
[45] V. Torra, Constrained microaggregation: adding constraints for data editing, Transactions on Data Privacy 1 (2) (2008) 86–104.
[46] V. Torra, Y. Narukawa, Modeling Decisions: Information Fusion and Aggregation Operators, Springer, 2007.
[47] V. Torra, S. Miyamoto, Evaluating fuzzy clustering algorithms for microdata protection, in: Lecture Notes in Computer Science, vol. 3050, Springer, Berlin/Heidelberg, 2004, pp. 175–186.
[48] US Census Bureau, Data Extraction System, 2011. <http://www.census.gov/>.
[49] W.E. Winkler, Advanced methods for record linkage, in: American Statistical Association Proceedings of Survey Research Methods Section, 1994, pp. 467–472.
[50] W.E. Winkler, Matching and record linkage, in: B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, P.S. Kott (Eds.), Business Survey Methods, Wiley Publications, New York, 1995, pp. 355–384.
[51] W.E. Winkler, Data cleaning methods, in: Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation, 2003, pp. 1–6.
[52] W.E. Winkler, Re-identification methods for masked microdata, in: Privacy in Statistical Databases 2004, Lecture Notes in Computer Science, vol. 3050, Springer, Berlin/Heidelberg, 2004, pp. 216–230.
[53] W. Yancey, W.E. Winkler, R. Creecy, Disclosure risk assessment in perturbative microdata protection, in: Inference Control in Statistical Databases, Lecture Notes in Computer Science, vol. 2316, Springer, Berlin/Heidelberg, 2002, pp. 135–152.
[54] W.E. Winkler, Masking and re-identification methods for public use microdata: overview and research problems, in: Privacy in Statistical Databases, Lecture Notes in Computer Science, vol. 3050, Springer, Berlin/Heidelberg, 2004, pp. 231–246.