Project N°: 262608

ACRONYM: Data without Boundaries

DELIVERABLE D11.2

Record linkage approaches for dynamic database integration

WORK PACKAGE 11

Improved Methodologies for Managing Risks of Access to Detailed OS Data

REPORTING PERIOD:	From: Month 1	To: Month 18
PROJECT START DATE:	1st May 2011	DURATION: 48 Months
DATE OF ISSUE OF DELIVERABLE:	June 2013	
DOCUMENT PREPARED BY:	Partner Numbers 21	Partners Names CISC

Combination of CP & CSA project funded by the European Community
Under the programme "FP7 - SP4 Capacities"
Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics

The views and conclusions expressed are those of the author(s) and do not necessarily represent those of the DwB consortium as a whole.

Deliverable D11.2. Record linkage approaches for dynamic database integration

Vicenç Torra*

* Artificial Intelligence Research Institute (IIIA)
Spanish Council for Scientific Research (CSIC)
Campus Universitat Autònoma de Barcelona
08193 Bellaterra, Catalonia, Spain
vtorra@iiia.csic.es

Abstract. In data privacy, record linkage is a well known technique to evaluate the disclosure risk of protected data. Given a file protected by means of a data protection technique and a file that represents the information of an intruder, record linkage can be used to link records in the two files. The larger the number of correctly linked records, the larger the risk of disclosure. Because of that, record linkage is used as a measure of disclosure risk. In this report we give an overview of the work that we have been doing during the last months in the framework of the DwB project, and the results we have obtained.

We first give an outline of disclosure risk measures, and then briefly describe the development of a supervised learning method for distance-based record linkage, which determines the optimum parameters for the linkage process. This corresponds to the consideration of the worst-case scenario because it means to find the parameters that lead to a maximum risk.

1 Introduction

When databases are protected by means of a data protection method that reduces the quality of the data, two aspects are of great importance: information loss (or data utility) and disclosure risk. Measures have been developed to evalute both aspects. Information loss measures evaluate in what extent the analyses a user can do on the protected data are similar to the ones a user can do on the original data. That is, whether e.g. for certain variables the regression coefficients on the protected data would be the same or similar to the regression coefficients on the original data. Disclosure risk measures evaluate in what extent a protected file can

be used by an intruder to learn some information. It is clear that protecting data by means of reducing its quality does not ensure confidentiality. For example, it is clear that when we remove from a database all variables except the social security number and the illness, disclosure is not avoided.

Record linkage is one of the tools used to measure disclosure risk. As we will discuss later, one of its advantages over other approaches is its flexibility and that it can be used in different scenarios. In fact, record linkage can be used in dynamic environments. For example, it can be used when multiple protected versions of the same data are released, or when intruders have information from different sources and combine them to attack a published data set.

This document discusses several aspects of disclosure risk measures. In Section 2 we give an overview of disclosure risk measures. Then, the rest of the paper focuses on the use of re-identification algorithms, one of the approach to define such disclosure risk measures.

2 Disclosure Risk Measures

In general, disclosure risk is defined in terms of the additional confidential information or knowledge that an intruder can acquire from the protected data set. According to [21, 29], disclosure risk can be studied from two perspectives:

- **Identity disclosure.** Disclosure takes place when a respondent is linked to a particular record in the protected data set. This process of linking is known as re-identification (of the respondent).
- **Attribute disclosure.** In this setting it is considered too strong to define disclosure as the disclosure of the identity of the individual. Disclosure takes place when the intruder can learn something new about an attribute of a respondent, even when no relationship can be established between the individual and the data. That is, disclosure takes place when the published data set permits the intruder to increase his accuracy on an attribute of the respondent. This approach was first formulated in [7] (see also [15] and [16]).

Interval disclosure is a measure, proposed in [11] and [12], for attribute disclosure. It is defined according to the following procedure.

Definition 1 Interval disclosure. *Each attribute is independently ranked and a rank interval is defined around the value the attribute takes on each record. The ranks of values within the interval for an attribute around record r should differ less than p percent of the total number of records and the rank in the center of*

the interval should correspond to the value of the attribute in record r. Then, the proportion of original values that fall into the interval centered around their corresponding protected value is a measure of disclosure risk.

A 100 percent proportion means that an attacker is completely sure that the original value lies in the interval around the protected value (interval disclosure).

From our point of view, some attribute disclosure is natural in any release of data, otherwise, it is difficult to find any utility on the publication of data. Note that e.g. almost any regression model leads to attribute disclosure because the goal of a regression model is to estimate the value of a variable from the other ones. Naturally, if the regression model really fits the data, the intruder can use it to infer some values with no much uncertainty.

Because of that, identity disclosure has received much attention in the last years and has been used to evaluate different protection methods. Its formulation needs a concrete scenario. We present it below.

2.1 An Scenario for Identity Disclosure

The typical scenario for identity disclosure considers a protected data set and an intruder having some partial information about the individuals described in the published data set. The protected data set is assumed to be a data file, and it is usual to consider that intruder's information can be represented in the same way. See e.g. [42, 44].

Formally, we consider data sets X with the usual structure of r rows (*records*) and k columns (*attributes*). Naturally, each row contains the values of the attributes for an individual.

The attributes in X can be classified [8, 33, 44] in three non-disjoint categories.

- **Identifiers.** These are attributes that *unambiguously* identify the respondent. Examples are passport number, social security number, full name, etc.
- **Quasi-identifiers.** These are attributes that, in combination, can be linked with external information to re-identify some of the respondents. Examples are age, birth date, gender, job, zipcode, etc. Although a single attribute cannot identify an individual, a subset of them can.
- **Confidential.** These are attributes which contain sensitive information on the respondent. For example, salary, religion, political affiliation, health condition, etc.

Using these three categories, an original data set X can be decomposed 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. Let us consider the protected data set X'. X' is obtained from the application of a protection procedure to X. This process takes into account the type of the attributes. It is usual to proceed as follows [42, 23, 24].

- **Identifiers.** To avoid disclosure, identifiers are usually removed or encrypted in a preprocessing step. In this way, information cannot be linked to specific respondents.
- **Confidential.** These attributes X_c are usually not modified. So, we have $X'_c = X_c$.
- **Quasi-identifiers.** They cannot be removed as almost all attributes can be quasi-identifiers. The usual approach to preserve the privacy of the individuals is to apply protection procedures to these attributes. We will use ρ to denote the protection procedure. Therefore, we have $X'_{nc} = \rho(X_{nc})$.

Therefore, we have $X' = \rho(X_{nc})||X_c$. Proceeding in this way, we allow third parties to have precise information on confidential data without revealing to whom the confidential data belongs to.

In this scenario we have identity disclosure when an intruder, having some information described in terms of a set of records and some quasi-identifiers, can correctly link his information with the published data set. That is, he is able to link his records with the ones in the protected data set. Then, if the links between records are correct, he will be able to obtain the right values for the confidential attributes.

Figure 1 represents this situation. A represents the file with data from the protected data set (i.e., containing records from X') and B represents the file with the records of the intruder. B is usually defined in terms of the original data set X, because it is assumed that the intruder has a subset of X. In general, the number of records owned by the intruder and the number of records in the protected data file will differ.

Re-identification is typically achieved using some common quasi-identifiers on both X and X'. They permit to link pairs of records (using record linkage algorithms) from both files, and, then, the confidential attribute is linked to the identifiers. At this point re-identification is achieved.

Formally, following [44, 26, 42] and the notation in Figure 1, the intruder is assumed to know the non-confidential quasi-identifiers $X_{nc} = \{a_1, \ldots, a_n\}$ together with the identifiers $Id = \{i_1, i_2, \ldots\}$. Then, the linkage is between identifiers

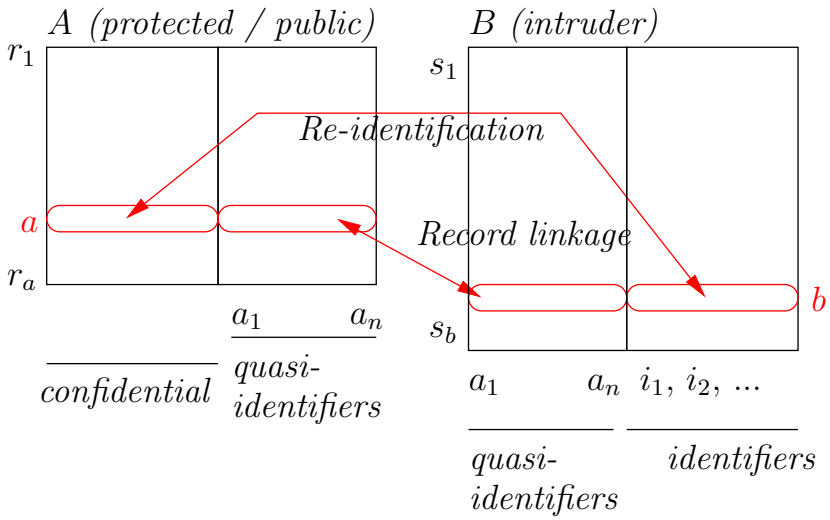

Figure 1: Disclosure Risk Scenario.

$(a_1, \ldots, a_n)$ from the protected data (X'_{nc}) and the same attributes from the intruder (X_{nc}).

2.2 Measures for Identity Disclosure

Two main approaches exists for measuring identity disclosure risk. They are known by uniqueness and re-identification. We describe them below.

- **Re-identification.** Risk is defined as an estimation of the number of re-identifications that might be obtained by an intruder. This estimation is obtained empirically through record linkage algorithms. This approach for measuring disclosure risk goes back, at least, to [39] and [29] (using *e.g.* the algorithm described in [30]). [44, 26, 42] are more recent papers using this approach. This approach is general enough to be applied in different contexts. It can be applied under different assumptions of intruder's knowledge, and under different assumptions on protection procedures. It can even be applied when protected data has been generated using a synthetic data generator (i.e., data is constructed using a particular data model). For example, [44] describes empirical results about using several implementations of distance-based record linkage algorithms on synthetic data. The performance of different algorithms is discussed. [55] considers a similar problem.

Record linkage can also be used in dynamic environments, when multiple protected versions of the same data set are released on the fly, or when intruders have information from different sources and combine them to attack a published data set. See e.g. [40] where we deal with the problem of multiple releases of k-anonymous data sets and k-anonymous relational databases.

- **Uniqueness.** Informally, the risk of identity disclosure is measured as the probability that rare combinations of attribute values in the protected data set are indeed rare in the original population.

  This approach is typically used when data is protected using sampling [53] (i.e., X' is just a subset of X). Note that with perturbative methods it makes no sense to investigate the probability that a rare combination of protected values is rare in the original data set, because *that* combination is most probably *not found* in the original data set.

Because of the versatility of re-identification, and its suitability for modelling a large variety of scenarios, we proposed to use this approach in this project.

In the next section we discuss with some detail record linkage algorithms, the ones used for re-identification when we focus on the problem of re-identifying records. Note that other algorithms can also be appropriate for disclosure risk assessment. For example, when the schema of the published data and the one of the data of the intruder are not equal, schema matching algorithms are useful. See [46] for a discussion on scenarios where re-identification algorithms other than record linkage are useful.

2.3 Record Linkage

Record linkage is the process of finding quickly and accurately two or more records distributed in different databases (or data sources in general) that make reference to the same entity or individual. This term was initially introduced in the public health area by [17], when files of individual patients were brought together using name, date-of-birth and other information. As briefly stated above, identifying links between the protected data set and the original one, we can evaluate the re-identification risk of the data by an intruder.

This approach for measuring disclosure risk directly follows the scenario in Figure 1. That is, record linkage consists of linking each record b of the intruder (file B) to a record a in the original file A. The pair (a, b) is a match if b turns out to be the original record corresponding to a. For applying record linkage, the common approach is to use the shared attributes (some quasi-identifiers). As the number of matches is an estimation of the number of re-identifications that an

intruder can achieve, disclosure risk is defined as the proportion of matches among the total number of records in B.

Two main types of record linkage algorithms are described in the literature: distance-based and probabilistic. They are outlined below. For details on these methods see [45].

- **Distance-based record linkage.** Each record b in B is linked to its nearest record a in A. An appropriate definition of a record-level distance has to be supplied to the algorithm to express *nearness*. This distance is usually constructed from distance functions defined at the level of attributes. In addition, we need to standardize attributes as well as assign weights to them.

  [31] proposed distance-based record linkage to assess the disclosure risk for microaggregation. They used Euclidean distance and equal weight for all attributes. Later, in [12], distance-based record linkage (also with Euclidean distance and equal weights) was used for evaluating other masking methods as well. In their empirical work, distance-based record linkage outperforms probabilistic record linkage (See Section 2.3 below).

  The main advantages of using distances for record linkage are simplicity for the implementer and intuitiveness for the user. Another strong point is that subjective information (about individuals or attributes) can be included in the re-identification process by means of appropriate distances.

  The main difficulties for distance-based record linkage are (i) the selection of the appropriate distance function, and (ii) the determination of the weights. In relation to the distance function, for numerical data, the Euclidean distance is the most used distance. Nevertheless, other distances have also been used as e.g. Mahalanobis [44], and some Kernel-based ones [44]. The difficulty of choosing a distance is especially thorny in the cases of categorical attributes and of masking methods such as local recoding where the masked file contains new labels with respect to the original data set. The determination of the weights is also a rellevant problem that is difficult to solve. In the case of the Euclidean distance, it is common to assign equal weights to all attributes, and in the case of the Mahalanobis distance, this problem is avoided because weights are extracted from the covariance matrix. As an alternative, we have developed an approach based on supervised machine learning (see Section 3 for details).

- **Probabilistic record linkage**

  Probabilistic record linkage also links pairs of records (a, b) in data sets A and B, respectively. For each pair, an index is computed. Then, two

thresholds LT and NLT in the index range are used to label the pair as linked, clerical or non-linked pair: if the index is above LT, the pair is linked; if it is below NLT, the pair is non-linked; a clerical pair is one that cannot be automatically classified as linked or non-linked.

When independence between attributes is assumed, the index can be computed from the following two conditional probabilities for each attribute: the probability $P(1|M)$ of coincidence between the values of the attribute in two records a and b given that these records are a real match, and the probability $P(0|U)$ of non-coincidence between the values of the attribute given that a and b are a real unmatch.

To use probabilistic record linkage in an effective way, we need to set the thresholds LT and NLT and estimate the conditional probabilities $P(1|M)$ and $P(0|U)$ used in the computation of the indices. In plain words, thresholds are computed from: (i) the probability $P(LP|U)$ of linking a pair that is an unmatched pair (a *false positive* or *false linkage*) and (ii) the probability $P(NP|M)$ of not linking a pair that is a match (a *false negative* or *false unlinkage*). Conditional probabilities $P(1|M)$ and $P(0|U)$ are usually estimated using the EM algorithm [10].

The original description of probabilistic record linkage can be found in [19] and [20]. [45] describe the method in detail (with examples) and [54] presents a review of the state of the art on probabilistic record linkage. In particular, this latter paper includes a discussion concerning non-independent attributes. A (hierarchical) graphical model has recently been proposed [32] that compares favorably with previous approaches.

Probabilistic record linkage methods are less simple than distance-based ones. However, they do not require rescaling or weighting of attributes. The user only needs to provide the two probabilities $P(LP|U)$ (false positives) and $P(NP|M)$ (false negatives).

The approaches described so far for record linkage do not use any information about the data protection process. That is, they use files A and B and try to re-identify as much records as possible. In this sense, they are general purpose record linkage algorithms.

In the last years, specific record linkage algorithms have been developed. They take advantage of any information available about the data protection procedure. That is, protection procedures are analyzed in detail to find flaws that can be used for computing more efficiently, with larger matching rates, record linkage algorithms. Attacks tailored for two protection procedures are reported in the literature. [47] was the first specific record linkage approach for microaggregation.

More effective algorithms have been proposed in [28, 27] (for either univariate and multivariate microaggregation). [26] describes an algorithm for data protection using rank swapping.

The scenario described above can be relaxed so that the published file and the one of the intruder do not share the set of variables. I.e., there are no common quasi-identifiers in the two files. A few record linkage algorithms have been developed under this premise. In this case, some structural information is assumed to be common in both files. [43] follows this approach. Its use for disclosure risk assessment is described in [13].

In this document we focus on the case that the two files are described in terms of the same schema, and that the intruder applies a distance-based record linkage for attacking the protected data set. Because of that in Section 2.6 we review a few distance functions to be used within the distance-based record linkage algorithm. Then, in Section 3, we review a supervised learning approach for record linkage.

As we will see below, some of the distances are parametric. That is, they depend on some parameters (weights) for the attributes under consideration. This causes that the effectiveness of the record linkage algorithm depends on an appropriate selection of the weights. The *worst-case analysis* corresponds to the case in which an intruder selects the best possible parameterization. Following the approaches found in optimization and supervised machine learning, we define the best possible parameterization as the one that leads to the maximum number of re-identifications. Then, we define a mathematical programming problem in this way, and its solution permits to evaluate the risk of such worst-case scenario.

2.4 Re-identification and Record Linkage: Formalization

The fact that the worst-case scenario is rellevant for disclosure risk assessment relates to the fact that a data protection is acceptable if the risk is low for *any attack*. In other words, when a data file is released, the application by the intruder of any record linkage method with his own data should lead to low bounds of re-identification. Note that even if for most algorithms, re-identifcation bounds are low but there is at least one that leads to disclosure, the intruder might apply this one to compromise the data.

In order to make the concept of re-identification precise, we need a formal definition. That is, we need a definition that makes explicit which are the re-identification methods and which are not. We have worked towards a formalization of re-identification algorithms based on the concepts of probabilities, belief functions, and compatible belief functions [6, 38].

We reproduce below our first definition of re-identification method which is

then refined into the precise concept of compatibility. Informally, we define that an algorithm is a re-identification algorithm when it leads to a probability distribution that is compatible with the true one. We understand the true one as the probability we would have if there is no uncertainty in the data protection process. Then, when uncertainty is present, we accept only those probabilities that represent the existing uncertainty, but we do not accept in any case any arbitrary probability. For example, when an attribute is not available to the intruder, it is clear that some information is lost. But the available attributes constraint possible probabilities. Any re-identification algorithm should take such constraints into account when assigning the probabilities.

Definition 2 *[41] Let ρ be a method for anonymization of databases, X a table with n records indexed by I in the space of tables D and $Y := \rho(X)$ the anonymization of X using ρ. Then a re-identification method is a function that, given a collection of entries y in $\mathcal{P}(Y)$ and some additional information from a space of auxiliary informations A, returns the probability that y are entries from the record with index $i \in I$,*

$$\begin{aligned} r: \quad \mathcal{P}(Y) \times A \quad &\to \qquad\qquad [0,1]^n \\ (y, a) \qquad &\mapsto \quad (P(y \text{ corresponds to record } X[i]) : i \in I)\,. \end{aligned}$$

Consider the objective probability distribution corresponding to the re-identification problem. Then, we require from a re-identification method that it returns a probability distribution that is compatible with this probability, also when missing some relevant information. Compatibility can be modeled in terms of compatibility of belief functions (see [6, 38]).

Let us now consider the more precise definition. As this definition uses the concept of compatibility between two probabilities, we also give this definition. Compatibility is expressed in terms of belief functions and the pignistic transformation. See [51] for more details and the application of these definitions to risk assessment in data protection.

Definition 3 *[52] Given two probabilities P and P', we say that P' is compatible with P if there exists a belief function Bel compatible with P such that P' is the pignistic probability distribution derived from Bel (i.e., $P' = P_{Bel}$).*

Definition 4 *[52] Let ρ be a method for anonymization of databases, X a table in the space of tables D and $Y := \rho(X)$ the anonymization of X using ρ. Let $P_{\rho,X,Y}(x_i|y)$ be the true probability of ρ, X. Then, an algorithm is a record linkage algorithm when it returns a probability distribution P' that is compatible with the true probability $P_{\rho,X,Y}(x_i|y)$.*

2.5 On two Alternative Probabilistic Methods

In Section 2.3 we have reviewed the probabilistic record linkage framework according to Fellegi and Sunter [19]. In a recent work, Skinner [36] compares this approach and the probabilistic modeling framework based on the Poisson log-linear model. This latter approach is described in [18, 37] for the case in which misclassification is not considered. [36] shows that the two frameworks can be seen from a unique perspective. Later, [35] defines disclosure risk measures that take into account misclassification. Note that misclassification is relevant in statistical disclosure control because is introduced on purpose by data protection methods.

Within the framework of this project, [34] has provided empirical evidence of the result in [36], demonstrating how disclosure risk can be assessed for a highly perturbed dataset containing business data from a 1982 Queensland, Australia Survey of Sugar Farms.

2.6 Distances for Distance-Based Record Linkage

In this section we review distances used in distance-based record linkage. We consider the Euclidean distance and the Mahalanobis one. In the definitions below, we 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 V_i^X.

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

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

DBRLM: The Mahalanobis distance between records a and b is defined by:

$$d(a,b)^2 = (a-b)'\Sigma^{-1}(a-b) \tag{2}$$

where, Σ is the covariance matrix. Note that if the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean one.

Other distances can be considered. Some of them can be seen as generalizations of the Euclidean distance. In particular, observing the Euclidean distance between two records as the arithmetic mean of the distances between the attributes, we can consider replacing the arithmetic mean by any other mean as e.g. the weighted mean. Some of such distances are reviewed below.

Let us detail this process. First, let us define d_i as the distance of the ith attribute:

$$d_i(a,b)^2 = \left(\frac{V_i^X(a) - \overline{V_i^X}}{\sigma(V_i^X)} - \frac{V_i^Y(b) - \overline{V_i^Y}}{\sigma(V_i^Y)} \right)^2 \quad (3)$$

Then, we can rewrite Equation 1 as

$$d(a,b)^2 = n^2 \cdot AM(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

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

It is easy to prove that a distance-based record linkage using distance $d(a,b)$ will result in the same number of re-identifications that using the expression $d(a,b)/n$. Because of that in the rest of this section we will drop the factor n from the expressions.

In general, any aggregation operator $\mathbb{C}$ [48] can be used to define a distance as follows:

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

From this definition, it is straightforward to consider weighted variations. We consider three variations below, which we have used for disclosure risk assessment. We begin the weighted Euclidean distance, which is based on the weighted mean.

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

$$d^2WM_p(a,b) = WM_p(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

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

Another aggregation operator we have considered is the Choquet integral. The main difference with the weigthed mean is that it uses a fuzzy measure as its parameter. The fuzzy measure permit to represent interactions (e.g. redundancy and complementariness) between the attributes which cannot be represented in the weights of the weighted mean. In short, weighted mean presumes that attributes are independent and the Choquet integral does not.

DBRLCI: Let μ be an unconstrained fuzzy measure on the set of variables V, i.e. $\mu(\emptyset) = 0$, $\mu(V) = 1$, and $\mu(A) \leq \mu(B)$ when $A \subseteq B$ for $A \subseteq V$, and $B \subseteq V$. Then, the Choquet integral distance is defined as:

$$d^2CI_\mu(a,b) = CI_\mu(d_1(a,b)^2, \ldots, d_n(a,b)^2),$$

where CI is the Choquet integral, i.e.,

$$CI_\mu(c_1, \ldots, c_n) = \sum_{i=1}^{n} (c_{s(i)} - c_{s(i-1)})\mu(A_{s(i)}),$$

given that $c_{s(i)}$ indicates a permutation of the indices so that $0 \leq c_{s(1)} \leq \ldots \leq c_{s(i-1)}$, $c_{s(0)} = 0$, and $A_{s(i)} = \{c_{s(i)}, \ldots, c_{s(n)}\}$.

The last approach we have considered is based on the Mahalanobis distance. To do so, firstly, we have to compute the normalized difference between two records $a \in X$ and $b \in Y$, with $d_i(a, b)$ (squared root of Equation 3), and then, use the Mahalanobis distance as an aggregation operator. The corresponding expression is given below.

DBRLQ: Let Σ be an nxn invertible matrix with the role of a covariance matrix. Then, the Mahalanobis distance is defined as:

$$d^2MD^*(a, b) = MD_\Sigma(d_1(a, b), ..., d_n(a, b))$$

where $MD_\Sigma(c_1, ..., c_n) = (c_1, ..., c_n)^T \Sigma^{-1}(c_1, ..., c_n)$.

Note that Σ, is a symmetric matrix. Then, the diagonal of the matrix expresses the relevance of each single variable in the re-identification process, whereas the up or down triangle values of the matrix are the weights that evaluates the interactions between each pair of variables.

The interest of the variations we have presented 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. Note that in DBRLCI and DBRLQ the interaction of different variables is taken into account by the fuzzy measure, in contrast to DBRLW which can only weight the variables individually.

Figure 2 representes a classification of the different distances we have defined. The arithmetic mean is a special case of the weighted mean (when all the weights are equal), and the weighted mean is also a special case of both the Choquet integral and the Mahalanobis distance. For more details see [49].

3 Supervised Learning for Record Linkage

The idea of applying supervised learning for record linkage is to determine the best parameters for an intruder to attack the data. Therefore, this corresponds to the

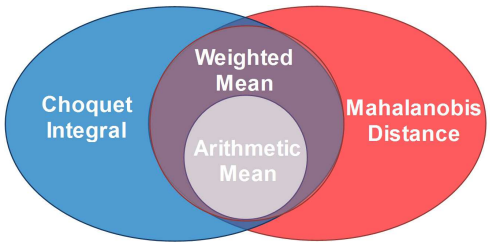

Figure 2: Distances classification

worst-case scenario. We have applied this approach to the distances discussed in the previous section. In the rest of this section we describe the formalization of the problem. That is, which is the problem we solve to find the optimal weights. We also give an overview of the results obtained.

3.1 Determination of the optimal weights

For the sake of simplicity, we presume that each record of X, $X_i = (a_1, \ldots, a_N)$, is the protected record of Y, $Y_i = (b_1, \ldots, b_N)$. 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 $i \neq j$. I.e.,

$$\mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) < \mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) \quad (4)$$

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. I.e. 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.

The rationale of this approach is as follows. We consider a variable K which indicates, for each block, if all the corresponding constraints are satisfied ($K = 0$) or not ($K = 1$). Then, we want to minimize the number of blocks non compliant with the constraints. 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. 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.

Note that if for a record i, Equation (4) 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:

$$\mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - \mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + CK_i > 0 \quad (5)$$

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 these constraints we can define the optimization problem for a given aggregation operator $\mathbb{C}$ as:

$$Minimize \sum_{i=1}^{N} K_i \quad (6)$$

$$Subject\ to:$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - \\ - \mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + CK_i > 0 \quad (7)$$

$$K_i \in \{0, 1\} \quad (8)$$

$$\text{Additional constraints according to } \mathbb{C} \quad (9)$$

where N is the number of records, and n the number of variables. This problem is a linear optimization problem with linear constraints and the (global) optimum solution can be found with an optimization algorithm. More explicitly, it can be considered a mixed integer linear problem (MILP), because it is dealing with integer and real-valued variables in the objective function and the constraints, respectively. Note, that we only have considered aggregation operators with real-valued weights.

If N is the number of records, and n the number of variables of the two data sets X and Y. We have N terms of K_i in the objective function, that is N variables for Equation (6). The total number of constraints in the optimization problem is N^2+N. There are N^2 constraints from Equation (7), and N for Equation (8). Note that depending on the aggregation operator $\mathbb{C}$ used, there will be more constraints in the problem.

3.1.1 Learning the Optimal Weights

Once the optimization problem is defined in general terms, we define in Table 1 the additional constraints which are necessary for each specific aggregation operator. For more details see our recent papers [1, 2, 50].

	d^2WM	d^2CI	d^2MD^{*1}
Additional Constraints	$\sum_{i=1}^{n} p_i = 1$ $p_i \geq 0$	$\mu(\emptyset) = 0$ $\mu(V) = 1$ $\mu(A) \leq \mu(B)$ when $A \subseteq B$	$MD_{\Sigma}(c_1, \ldots, c_n) \geq 0$

Table 1: Additional Constraints for the three variations of the problem.

3.2 Evaluation

We have evaluated our proposal with different protected files using *microaggregation*[9], a well-known microdata protection method, which broadly speaking, provides privacy by means of clustering the data into small clusters of size k, and then replacing the original data by the centroid of their corresponding clusters. This parameter k determines the protection level: the greater the k, the greater the protection and at the same time the greater the information loss.

We have considered files with the following protection parameters:

- *M4-33*: 4 variables microaggregated in groups of 2 with $k = 3$.
- *M4-28*: 4 variables, first 2 variables with $k = 2$, and last 2 with $k = 8$.
- *M4-82*: 4 variables, first 2 variables with $k = 8$, and last 2 with $k = 2$.
- *M5-38*: 5 variables, first 3 variables with $k = 3$, and last 2 with $k = 8$.
- *M6-385*: 6 variables, first 2 with $k = 3$, next 2 with $k = 8$, and last 2 with $k = 5$.
- *M6-853*: 6 variables, first 2 with $k = 8$, next 2 with $k = 5$, and last 2 with $k = 3$.

For each case, we have protected 400 records randomly selected from the Census dataset [5] from the European CASC project [4], which contains 1080 records and 13 variables, and has been extensively used in other works [22, 14, 56].

Note that in all cases, variables have been splitted into subsets and each of these subsets has been microaggregated independently. This masking procedure permits to have lower information loss than microaggregating all variables at the same time, but also some disclosure risk.

Note also that in our experiments we apply different protection degrees to different variables of the same file. These vary between 2 to 8, i.e., values between the lowest protection value and a good protection degree in accordance to [12]. This is especially interesting when variables have different sensitivity.

Table 2 shows the linkage ratio using the standard record linkage method (d^2AM); the Mahalanobis distance (d^2MD); and the three supervised learning approaches: the weighted mean (d^2WM), the Choquet integral (d^2CI) and finally the approach based on the Mahalanobis distance (d^2MD^*) which were described in Section 3.1. The values in the table are the ratio determining the correctly identified records from the total, so a ratio of 1 means a 100% re-identification.

	d^2AM	d^2MD	d^2WM	d^2CI	d^2MD^*[1]
M4-33	0.84	0.94	0.955	0.9575	0.9675
M4-28	0.685	0.9	0.93	0.9375	0.9425
M4-82	0.71	0.9275	0.9425	0.9425	0.9525
M5-38	0.3975	0.8825	0.905	0.9125	0.9225
M6-385	0.78	0.985	0.9925	0.9975	0.9975
M6-853	0.8475	0.98	0.9875	0.9925	0.995

Table 2: Improvement in the linkage ratio.

As it can be appreciated, our proposed methods achieve an important improvement with respect to the standard distances based record linkage. However, the improvement between the three supervised approaches is relatively small, especially between d^2CI and d^2MD*. Although the difference between methods d^2CI and d^2MD^* is small, it is important to bear in mind that the Choquet integral approach is computationally more expensive and complex. This is due to the number of constraints required in the optimization problem. This makes the proposed use of the Mahalanobis distance more effective than the one using the Choquet integral.

In our experiments we have used microaggregation, which is often applied in combination with sampling. In such a case, the disclosure risk computed evaluates only the disclosure risk of the microaggregation step and needs to be combined with the one of sampling. Because of that, the risk computed is an upper bound of the disclosure risk of the whole process.

4 Conclusions

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

[1]This is the supervised learning approach using the Mahalanobis distance.

between records of the original and the protected data. The interest of record linkage for disclosure risk measurement is due to the fact that it can be applied to a large number of scenarios. For example, it can be applied to standard data protection methods but also to the case of synthetic data generators, multiple releases of the same data, and intruders considering dynamic database integration.

In this document we have reviewed our recent results on record linkage. We have outlined the formalization of re-identification and also how we can use optimization and supervised machine learning approaches to study the worst-case scenario. As seen above, we have used different parameterized distances to find the optimal weights for record linkage. The weights we obtain supply the user information about the relevance of the attributes. For other results on dynamic integration more oriented to specific technological solutions as JDBC drivers see e.g. [25, 57].

The results described in this report are focused on numerical data. Nevertheless, record linkage can be applied to other types of data, as e.g. categorical data or time series. Both probabilistic record linkage and distance based record linkage has been applied to categorical data. Distance based record linkage has been applied to time series. In fact, given a certain domain, as soon as a distance function can be defined, distance based record linkage can be applied.

It is also important to note that the approach described in Section 3.1 for the determination of the optimal weights is also applicable to these other types of data. This is so because the optimization problem only depends on the values $d(V_k(a_i), V_k(b_j))$, which only depend on the distance function. Once these values are computed, the learning process can be applied and the optimal weights can be found.

References

[1] Abril, D., Navarro-Arribas, G., Torra, V. (2012) Choquet integral for record linkage. *Annals of Operations Research* 195 97-110.

[2] Abril, D., Navarro-Arribas, G., Torra, V. (2012) Improving record linkage with supervised learning for disclosure risk assessment. *Information Fusion* 13:4 274-284.

[3] Abril, D., Torra, V., Navarro-Arribas, G. (2012) Supervised Learning Using a Symmetric Bilinear Form for Disclosure Risk Evaluation, manuscript.

[4] Brand, R., Domingo-Ferrer, J., Mateo-Sanz, J.M. (2002) Reference datasets to test and compare sdc methods for protection of numerical microdata. *Technical report, European Project IST-2000-25069 CASC*, 2002.

[5] U.S. Census Bureau. Data extraction system.

[6] Chateauneuf, A. (1994) Combination of compatible belief functions and relation of specificity, in Advances in Dempster-Shafer Theory of evidence, Wiley, 97-114.

[7] Dalenius, T. (1977) Towards a methodology for statistical disclosure control, Statistisk Tidskrift 5 429-444.

[8] Dalenius, T. (1986) Finding a needle in a haystack - or identifying anonymous census records, Journal of Official Statistics 2:3 329-336.

[9] Defays, D., Nanopoulos, P. (1993) Panels of enterprises and confidentiality: The small aggregates method. In *Proc. of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204. Statistics Canada, 1993.

[10] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977) Maximum Likelihood From Incomplete Data Via the EM Algorithm, Journal of the Royal Statistical Society 39 1-38.

[11] Domingo-Ferrer, J., Mateo-Sanz, J. M., Torra, V. (2001) Comparing SDC methods for microdata on the basis of information loss and disclosure risk, Pre-proceedings of ETK-NTTS'2001, (Eurostat, ISBN 92-894-1176-5), Vol. 2, 807-826, Creta, Greece.

[12] Domingo-Ferrer, J., Torra, V. (2001) A quantitative comparison of disclosure control methods for microdata, in P. Doyle, J. I. Lane, J. J. M. Theeuwes, L. Zayatz (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland, 111-134.

[13] Domingo-Ferrer, J., Torra, V. (2003) Disclosure Risk Assessment in Statistical Microdata Protection via advanced record linkage, Statistics and Computing, 13 343-354.

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

[15] Duncan, G. T., Lambert, D. (1986) Disclosure-limited data dissemination, Journal of the American Statistical Association, 81 10-18.

[16] Duncan, G. T., Lambert, D. (1989) The risk disclosure for microdata, Journal of Business and Economic Statistics 7 207-217.

[17] Dunn, H. (1946) Record linkage. *American Journal of Public Health*, 36(12):1412–1416.

[18] Elamir, E., Skinner, C.J. (2006) Record-Level Measures of Disclosure Risk for Survey Microdata, Journal of Official Statistics 22 525-539.

[19] Fellegi, I. P., Sunter, A. B. (1969) A theory for record linkage, Journal of the American Statistical Association 64:328 1183-1210.

[20] Jaro, M. A. (1989) Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Association 84:406 414-420.

[21] Lambert, D. (1993) Measures of Disclosure Risk and Harm, Journal of Official Statistics 9 313-331.

[22] Laszlo, M., Mukherjee, S. (2005) Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Trans. on Knowl. and Data Eng.*, 17(7):902–911.

[23] Li, N., Li, T., Venkatasubramanian, S. (2007) T-closeness: privacy beyond k-anonymity and l-diversity, Proc. of the IEEE ICDE 2007.

[24] Machanavajjhala, A., Gehrke, J., Kiefer, D., Venkitasubramanian, M. (2006) L-diversity: privacy beyond k-anonymity, Proc. of the IEEE ICDE.

[25] Mason, T., Lawrence, R. (2005) Dynamic database integration in a JDBC driver, Proc. of the 7th Int. Conf. on Enterprise Information Systems - Databases and Information Systems Integration Track, Miami, FL.

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

[27] Nin, J., Herranz, J., Torra, V. (2008) On the Disclosure Risk of Multivariate Microaggregation, Data and Knowledge Engineering, 67:3 399-412.

[28] Nin, J., Torra, V. (2009) Analysis of the Univariate Microaggregation Disclosure Risk, New Generation Computing, 27 177-194.

[29] Paass, G. (1985) Disclosure risk and disclosure avoidance for microdata, Journal of Business and Economic Statistics 6 487-500.

[30] Paass, G., Wauschkuhn, U. (1985) Datenzugang, Datenschutz und Anonymisierung - Analysepotential und Identifizierbarkeit von Anonymisierten Individualdaten, Oldenbourg Verlag.

[31] Pagliuca, D., Seri, G. (1999) Some results of individual ranking method on the system of enterprise accounts annual survey, Esprit SDC Project, Deliverable MI-3/D2.

[32] Ravikumar, P., Cohen, W. W. (2004) A hierarchical graphical model for record linkage, Proc. of UAI 2004.

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

[34] Shlomo, N. (2011) Assessing Disclosure Risk in Perturbed Microdata, Proc. of Statistics Canada Symposium 2011 Strategies for Standardization of Methods and Tools - How to get there.

[35] Shlomo, N., Skinner, C.J. (2010) Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata, Annals of Applied Statistics 4:3 1291-1310.

[36] Skinner, C.J. (2008) Assessing Disclosure Risk for Record Linkage, Proc. Privacy in Statistical Databases, Lecture Notes in Computer Science 5262 166-176.

[37] Skinner, C.J., Shlomo, N. (2008) Assessing Identification Risk in Survey Microdata Using Loglinear Models, Journal of the American Statistical Association 103:483 989-1001.

[38] Smets, P., Kennes, R. (1994) The transferable belief model, Artificial Intelligence 66 191-234.

[39] Spruill, N. L. (1983) The confidentiality and analytic usefulness of masked business microdata, Proc. of the Section on Survery Research Methods 1983, American Statistical Association, 602-610.

[40] Stokes, K., Torra, V. (2012) Multiple Releases of k-Anonymous Data Sets and k-Anonymous Relational Databases, Int. J. of Unc. Fuzziness and Knowledge Based Systems 20(6) 839-854.

[41] Stokes, K., Torra, V. (2012) n-Confusion: a generalization of k-anonymity, Proc. of the PAIS workshop, Berlin, Germany, 30 March, 2012, 211-215

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

[43] Torra, V. (2004) OWA operators in data modeling and reidentification, IEEE Trans. on Fuzzy Systems 12:5 652-660.

[44] Torra, V., Abowd, J., Domingo-Ferrer, J. (2006) Using mahalanobis distance-based record linkage for disclosure risk assessment. *Lecture Notes in Computer Science*, (4302):233–242.

[45] Torra, V., Domingo-Ferrer, J. (2003) Record linkage methods for multi-database data mining, in V. Torra (ed.) Information Fusion in Data Mining, Springer, 101-132.

[46] Torra, V., Domingo-Ferrer, J., Torres, A. (2003) Data Mining Methods for Linking Data Coming from Several Sources, *Proceedings of the 3rd Joint UN/ECE-Eurostat Work Session on Statistical Data Confidentiality*, Monographs in Official Statistics, (Luxembourg: Eurostat, ISBN 92-894-5766-X, ISSN: 1725-5406), 143-150, 2003. Postproceedings of the UNECE/Eurostat Work Session on Statistical Disclosure Control, Luxembourg, Luxembourg, 7-9 abril 2003.

[47] Torra, V., Miyamoto, S. (2004) Evaluating fuzzy clustering algorithms for microdata protection, PSD 2004, Lecture Notes in Computer Science 3050 175-186.

[48] Torra, V., Narukawa, Y. (2007) *Modeling Decisions: Information Fusion and Aggregation Operators*. Springer.

[49] Torra, V., Narukawa, Y. (2012) On a comparison between Mahalanobis distance and Choquet integral: the Choquet-Mahalanobis operator, Information Sciences 190 56-63.

[50] Torra, V., Navarro-Arribas, G., Abril, D. (2011) Supervised learning approach for distance based record linkage as disclosure risk evaluation, UNECE / Eurostat Work Session on Statistical Confidentiality, 7th Work Session 2011 (Tarragona, Spain).

[51] Torra, V., Stokes, K. (2012) A formalization of record linkage and its application to data protection, Int. J. of Unc. Fuzziness and Knowledge Based Systems, 20:6 in press.

[52] Torra, V., Stokes, K. (2012) A formalization of re-identification in terms of compatible probabilities, submitted.

[53] Willenborg, L., Waal, T. (2001) Elements of statistical disclosure control. Springer-Verlag.

[54] Winkler, W. E. (1993) Matching and record linkage, Statistical Research Division, U. S. Bureau of the Census (USA), RR93/08.

[55] Winkler, W. E. (2004) Re-identification methods for masked microdata. Lecture Notes in Computer Science 3050, pages 216–230, Heidelberg, Berlin. Springer.

[56] Yancey, W. E., Winkler, W. E., Creecy, R. H. (2002) Disclosure risk assessment in perturbative microdata protection. In *Inference Control in Statistical Databases, From Theory to Practice*, volume 2316, pages 135–152, London, UK. Springer-Verlag.

[57] Zhang, Z., Zhao, Z., Cao, Z. (2012) Dynamic Integration System for Heterogeneous Database Based on S2SH, Proc. 2nd Int. Conf. on Computer and Information Application.