WP. 7
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Tarragona, Spain, 26-28 October 2011)

Topic (i): Disclosure risk assessment

Supervised learning approach for distance based record linkage as disclosure risk evaluation

Prepared by Vicenç Torra and Daniel Abril, Artificial Intelligence Research Institute (IIIA), Spanish Council for Scientific Research (CSIC), Spain and
Guillermo Navarro-Arribas, Department of Information and Communications Engineering (DEIC), Spain

Supervised Learning Approach for Distance Based Record Linkage as Disclosure Risk Evaluation

Vicenç Torra*, Guillermo Navarro-Arribas**, Daniel Abril***

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

** Department of Information and Communications Engineering (DEIC), Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Catalonia, Spain, guillermo.navarro@uab.cat

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

Abstract. In data privacy, record linkage is a well known technique to evaluate the disclosure risk of protected data. It is used to evaluate the number of linked records between a data set and its protected version. In this paper we give an overview of the work that we have been doing during the last months. We describe the development of a supervised learning method for distance-based record linkage, which determines the optimum parameters for the linkage process. We also present an evaluation and a comparison between three different alternatives of such method. They are based on the weighted mean, the Choquet integral and a variation of the Mahalanobis distance and also with other standard distances to evaluate the risk.

1 Introduction

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 [13], when files of individual patients were brought together using name, date-of-birth and other information. In the following years, this idea was developed in [21, 20, 15], and nowadays it is a popular technique used by statistical agencies, research communities and corporations. Record linkage is one of the existing preprocessing techniques used for data cleaning [19, 27], and it is also used to control the quality of the data [5]. For example, data sources could be analyzed to deal with dirty data like duplicate records [14], data entry mistakes, transcription errors, lack of standards for recording data fields, etc. Moreover,

it is nowadays a popular technique employed to integrate different data sets that provide information regarding to the same entities [10, 7].

In the last years, record linkage techniques have also emerged in the data privacy context. Many governments agencies and companies need to collect and analyze sensitive data about individuals. So, it is fundamental to provide security to statistical databases against disclosure of confidential information. Privacy preserving data mining [4] and Statistical Disclosure Control [26] research on methods and tools for ensuring the privacy of these data. Record linkage permits the evaluation of disclosure risk of protected data [23, 28]. By identifying links between the protected data set and the original one, we can evaluate the re-identification risk of the data by an intruder. For example, [11] defines a score using the combination of disclosure risk techniques, to evaluate the risk of re-identification, and another method, which readily quantified the information loss of a protected data set using analytical measures (either generic or data-use-specific).

In this paper we focus on distance-based record linkage. We give an overview of a supervised learning approach developed for three different parametrized distances, the weighted mean, the Choquet integral [9] and the Mahalanobis distance [18]. We show the suitability of our proposals, testing it in the field of data privacy and comparing all the developed methods with the currently standard methods to evaluate the disclosure risk.

The outline of this paper is as follows. In section 2, we review some concepts needed in the rest of the paper. In section 3, we describe the supervised learning approach for distance based record linkage with the three alternative parametrized distances. The evaluation of the approach is introduced in section 4. Finally, Section 5 presents the conclusions of the paper.

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 attacks 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, 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, but that can unequivocally identify an individual when combining several of them. 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 to the dataset X, leading to a protected dataset Y. Indeed, we will assume the following typical scenario with respect

to the protection method ρ: (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 $Y_c = X_c$; (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 $Y_{nc} = \rho(X_{nc})$ and so, $Y = \rho(X_{nc})||X_c$. This scenario, which was first used in [11] to compare several protection methods, has also been adopted in other works like [23].

There are two extensively used approaches for record linkage to evaluate the disclosure risk of protected data. The **Probabilistic record linkage (PRL)** [16] and the **Distance based record linkage (DBRL)** [22], which links each record from dataset A to the *closest* record in dataset B. The *closest* record is defined in terms of a distance function.

The work in this paper is focused on distance based record linkage.

2.1 Distance-Based Record Linkage

In this section we give the definition of two distances currently used as a record linkage techniques in the data privacy field. The first one relies on the Euclidean distance and the second on the Mahalanobis distance.

To do so 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: Distance based record linkage using the Mahalanobis distance is as follows:

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

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

3 Supervised Learning for Record Linkage

The idea of applying supervised learning with parametrized distances for record linkage is to determine the best parameters for achieving the best possible number of linked records. We deal with different parametrized distances, in this section we introduce three approaches to determine the weights associated to each variable, and also their interactions, depending on the complexity of the parametrized distance yielding an optimal distance-based record linkage. Firstly, we present three different parametrized distances with different number of parameters, i.e. the greater the number is, the better defined is the problem, then, we define how to determine the optimal weights of a parametric distance by means of an optimization problem and finally, we explain how to adapt the general problem to determine the best parameters for each distance presented.

3.1 Parametric Distances

There are lots of parametrized distances in the literature and most of them obtain different results when are applied to the same problem. Therefore, we have focused on three different very relevant types of parametric distances.

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 Euclidean distance as a weighted mean of the distances for the attributes.

Defining,

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

we can rewrite, Equation 1 as

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

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

In general, any aggregation operator $\mathbb{C}$ [24] might be used:

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

From this definition, it is straightforward to consider weighted variations. We consider three variations below.

Definition 1 *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 used is the Choquet integral (Definition 2). From a definitional point of view, its main difference with the previous tool is the use of fuzzy measures. Choquet integral and fuzzy measures permit to express information like redundancy, complementariness, and interactions among the variables, which are not reflected in the weighted mean. Therefore, tools that use fuzzy measures to represent background knowledge permit the consideration of variables that are not independent.

Definition 2 *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 relies 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:

Definition 3 *Let Σ be an $n \times n$ weighting matrix, instead of a covariance matrix as is used in Equation 2. 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 reidentification 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 these variations 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 Definition 2 and 3 the interaction of different variables is taken into account by the fuzzy measure, in contrast to Definition 1 which can only weight the variables individually.

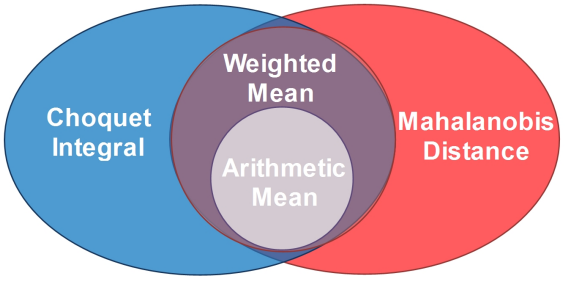

Figure 1: Distances classification

Figure 1 shows the classification of the different distances that we have explained. As you can see arithmetic mean is a special case of weighted mean and at the same time the weighted mean is also a shared special case between the Choquet integral and the Mahalanobis distance. For more details see [25].

3.2 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.

3.2.1 Learning the Optimal Weights

Once the optimization problem is defined in general terms, we define in Table 1 the additional constraints which are necessary to add for each specific aggregation operator explained in Section 3.1. More details and deeper explanations can be found in [1, 2, 3].

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

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

4 Evaluation

We have evaluated our proposal with different protected files using *microaggregation*[10], 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 [8] from the European CASC project [6], which contains 1080 records and 13 variables, and has been extensively used in other works [17, 12, 29].

Note 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 [11]. 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 3.2. 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.

5 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 between records of the original and the protected data.

In this paper we have introduced a supervised learning for distance based record linkage. Our proposal uses a supervised learning approach relying on three different parametrized distances to determine the optimal weights for the linkage. Moreover, these weights supply information about the relevance of the data attributes and depending on the approach used we obtain different accuracy types of information. Furthermore, we have evaluated these supervised learning approaches obtaining better results than the standard methods.

Acknowledgements

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.

Some of the results described in this paper have been obtained using the Centro de Supercomputación de Galicia (CESGA). This partial support is gratefully acknowledged.

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

References

[1] D. Abril, G. Navarro-Arribas, and V. Torra. Choquet integral for record linkage. *Annals of Operations Research*, 2011. In press.

[2] D. Abril, G. Navarro-Arribas, and V. Torra. Improving record linkage with supervised learning for disclosure risk assessment. *Information Fusion*, 2011. In press.

[3] D. Abril, G. Navarro-Arribas, and V. Torra. Supervised learning using mahalanobis distance for record linkage. In Radko Mesiar Bernard De Baets and Luigi Troiano, editors, *Proc. of 6th International Summer School on Aggregation Operators - AGOP2011*, pages 223–228. Lulu.com, 2011.

[4] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pages 439–450. ACM Press, 2000.

[5] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc., 2006.

[6] R. Brand, J. Domingo-Ferrer, and 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.

[7] Canada. Record linkage at statistics canada, 2010.

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

[9] G. Choquet. Theory of capacities, annales de l'institut fourier. 5:131–295, 1953.

[10] D. Defays and P. Nanopoulos. 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.

[11] J. Domingo-Ferrer and V. Torra. *A quantitative comparison of disclosure control methods for microdata*, pages 111–133. Elsevier, 2001.

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

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

[14] A. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.

[15] I. Fellegi and A. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[16] M. A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.

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

[18] P. C. Mahalanobis. On the generalised distance in statistics. In *Proceedings National Institute of Science, India*, volume 2, pages 49–55, April 1936.

[19] A. Mccallum and B. Wellner. Object consolodation by graph partitioning with a conditionally-trained distance metric. *In Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 19–24, 2003.

[20] H. B. Newcombe and J. M. Kennedy. Record linkage: making maximum use of the discriminating power of identifying information. *Commun. ACM*, 5(11):563–566, 1962.

[21] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 130:954–959, 1959.

[22] D. Pagliuca and G. Seri. Some results of individual ranking method on the system of enterprise acounts annual survey. *Esprit SDC Project, Delivrable MI-3/D2*, 1999.

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

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

[25] V. Torra and Y. Narukawa. On independece, expectation and distances: Choquet integrals and mahalanobis distance. *Modeling Decisions for Artificial Intelligence*, 2010.

[26] L. Willenborg and T. Waal. Elements of statistical disclosure control. Springer-Verlag, 2001.

[27] W. E. Winkler. Data cleaning methods. *Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.

[28] W. E. Winkler. Re-identification methods for masked microdata. volume 3050, pages 216–230, Heidelberg, Berlin, 2004. Springer.

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