Ann Oper Res (2012) 195:97–110
DOI 10.1007/s10479-011-0989-x

Choquet integral for record linkage

Daniel Abril · Guillermo Navarro-Arribas · Vicenç Torra

Published online: 5 October 2011

Abstract Record linkage is used in data privacy to evaluate the disclosure risk of protected data. It models potential attacks, where an intruder attempts to link records from the protected data to the original data. In this paper we introduce a novel distance based record linkage, which uses the Choquet integral to compute the distance between records. We use a fuzzy measure to weight each subset of variables from each record. This allows us to improve standard record linkage and provide insightful information about the re-identification risk of each variable and their interaction. To do that, we use a supervised learning approach which determines the optimal fuzzy measure for the linkage.

Keywords Data privacy · Record linkage · Choquet integral · Optimization

1 Introduction

Record Linkage techniques identify records from different databases (or data sources in general) that refer to the same entity. It was initially introduced for database integration in Dunn (1946) and further developed in Newcombe et al. (1959), Fellegi and Sunter (1969), and it is nowadays a popular technique used by statistical agencies, research communities, and corporations. Record linkage is mainly used to integrate different databases or data sets in general (Statistics Canada 2010; Colledge 1995; Data.gov 2010; Data.gov.uk 2010), or for data cleaning and quality control (Batini and Scannapieco 2006; Winkler 2003). For example, for detecting duplicate records between several data sets (Elmagarmid et al. 2007). More recently, in the context of data privacy (Lane et al. 2008), record linkage has emerged

D. Abril · G. Navarro-Arribas (✉) · V. Torra
IIIA, Institut d'Investigació en Intel·ligència Artificial—CSIC, Consejo Superior de Investigaciones Científicas, Campus UAB s/n, 08193 Bellaterra, Catalonia, Spain
e-mail: guille@iiia.csic.es

D. Abril
e-mail: dabril@iiia.csic.es

V. Torra
e-mail: vtorra@iiia.csic.es

as an important technique to evaluate the disclosure risk of protected data (Torra et al. 2006; Winkler 2004). By identifying the links between the protected dataset and the original one, we can evaluate the re-identification risk of the protected data (Domingo-Ferrer and Torra 2001).

Among record linkage approaches are those based on a distance function between records, which link records by their closeness. In this paper we introduce a new distance-based record linkage for data privacy which is based on the Choquet integral (Choquet 1953; Torra and Narukawa 2007). This approach allows the use of a fuzzy measure to weight the attributes in each dataset. Moreover, we present a supervised learning approach to determine the fuzzy measure for the linkage.

Our contribution is twofold. On the one hand we improve the performance of more standard distance-based record linkage. On the other hand we provide insightful information about the relevance of each variable, and the interactions between variables, in the linkage process. This is especially important in data privacy. Our method identifies which attributes provide more information for the record linkage, or in other words, which concrete attributes leak more information for the re-identification of individuals and increase the disclosure risk. It can be seen as an evaluation of disclosure risk of each individual attribute. Moreover, since we use a fuzzy measure to weight the attributes, we can also determine the relevance of the interaction of several attributes. This information can be used by the statistical agencies to increase the protection for concrete attributes.

The paper is organized as follows. Section 2 introduces distance-based record linkage in the context of data privacy. In Sect. 3 we describe our contribution, and in Sect. 4 we show our results and evaluation. Finally, Sect. 5 concludes 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 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, 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 Y. 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 $Y_c = X_c$; (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 $Y_{nc} = \rho(X_{nc})$. This scenario, which was first used in Domingo-Ferrer and Torra (2001) to

compare several protection methods, has also been adopted in other works like Torra et al. (2006).

Once the protected dataset Y is released, everybody can see its content $Y = Y_{nc} \| X_c$. We assume now that an intruder obtains from another data source another non-protected dataset $Z = z_{id} \| z_{nc}$ 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 Y and the non-protected dataset Z using the common attributes between Y and Z (y_{nc} and z_{nc}). If the intruder is able to correctly link a record of Z with its corresponding protected record in Y, 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.

There are two main approaches for record linkage

- **Distance based record linkage (DBRL)**. This approach (Pagliuca and Seri 1999) links each record from dataset A to the *closest* record in dataset 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 (Hartley 1958; McLachlan and Krishnan 1997; Jaro 1989) is normally used.

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 V_i^X.

As an example we show two common distance functions used in distance based record linkage which rely on the Euclidean distance. Other distance functions such as the Mahalanobis or Kernel distances, have also been used for record linkage (Torra et al. 2006).

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

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

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

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

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 distance weighted with a fuzzy measure as will be detailed in the next sections.

3 Learning optimal distances for record linkage

In this section we introduce our approach to determine weights associated to each variable yielding an optimal distance-based record linkage. To that end, we first introduce a new distance based on the Choquet integral, and then present the learning process to determine the optimal fuzzy measure to weight the attributes.

3.1 A Choquet integral based 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 Sect. 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

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

we can rewrite this expression as

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

where AM is the arithmetic mean $AM(c_1, \ldots, c_n) = \sum_i c_i/n$. We will denote this distance as $d^2AM(a,b)$.

In general, any aggregation operator $\mathbb{C}$ might be used (Torra and Narukawa 2007):

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

From this definition, we consider a weighted version of DBRL1 using a fuzzy integral as aggregation operator. In this case we rely on a fuzzy measure μ to weight the relevance not only of the single variables, but also of their interaction, and use the Choquet integral as follows.

Definition 1 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:

$$d^2CI_\mu(a,b) = CI_\mu(d_1(a,b)^2, \ldots, d_n(a,b)^2) \tag{1}$$

where $CI_\mu(c_1, \ldots, c_n) = \sum_{i=1}^{n} (c_{s(i)} - c_{s(i-1)})\mu(A_{s(i)})$, given that $c_{s(i)}$ indicates a permutation of the indices so that $0 \leq c_{s(1)} \leq \cdots \leq c_{s(i-1)}$, $c_{s(0)} = 0$, and $A_{s(i)} = \{c_{s(i)}, \ldots, c_{s(n)}\}$.

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 the interaction of different variables is taken into account by the fuzzy measure. Note that previous proposals for weighted distances for record linkage (Torra et al. 2010) only considered the weighted mean and OWA operators, which can only weight individual variables.

3.2 Determining 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.,

$$\begin{aligned} &\mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) \\ &\quad < \mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) \end{aligned} \tag{2}$$

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, (2) 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:

$$\begin{aligned} &\mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) \\ &\quad - \mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + CK_i > 0 \end{aligned} \tag{3}$$

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:

$$\textit{Minimize} \sum_{i=1}^{N} K_i \tag{4}$$

Subject to

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - \mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + CK_i > 0 \tag{5}$$

$$K_i \in \{0, 1\} \tag{6}$$

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 (4). The total number of constraints in the optimization problem is $N^2 + N$. There are N^2 constraints from (5), and N for (6). Note that depending on the aggregation operator $\mathbb{C}$ used, there will be more constraints in the problem.

3.3 Learning the optimal fuzzy measure for record linkage

In the case of the Choquet integral based distance d^2CI introduced in Sect. 3.1, the minimization problem can be defined in a generic form as:

$$\textit{Minimize} \sum_{i=1}^{N} K_i \tag{7}$$

Subject to

$$\sum_{i=1}^{N}\sum_{j=1}^{N} CI_\mu(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - CI_\mu(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + CK_i > 0 \tag{8}$$

$$K_i \in \{0, 1\} \tag{9}$$

$$\mu(\emptyset) = 0 \tag{10}$$

$$\mu(V) = 1 \tag{11}$$

$$\mu(A) \leq \mu(B) \quad \text{when } A \subseteq B \tag{12}$$

To formulate the problem we use the Möbius transform of the fuzzy measure instead of the measure itself. So, we have rewritten the optimization function and also the constraints in terms of the Möbius transformation, following a similar approach as described in Torra and

Narukawa (2007, Chap. 8). Recall, that given a fuzzy measure μ on the set V, its Möbius transform m is defined as:

$$m_\mu(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|} \mu(B)$$

for all $A \subset V$.

Appreciate that the function m is not restricted to the interval [0, 1].

We use the following notation to denote the fuzzy measure of the set of variables V (recall that $|V| = n$). However, instead of $\mu(A)$ to denote the V subsets' measures, we have considered μ_k, with $k \in \{0, \ldots, 2^n - 1\}$, where the index k denotes the subset of variables V determined by the dyadic representation of k. Let $\delta_n^k \delta_{n-1}^k \ldots \delta_1^k$ be the dyadic representation of k, then μ_k denotes the measure of the following set:

$$\mu_k = \mu(\{v_l \in V \mid \delta_l^k = 1 \text{for} l = 1, \ldots, n\})$$

In general, we denote as $\delta(A)$ the index k corresponding to the set A given by its dyadic representation.

We can now express the fuzzy measure as $(\mu_0, \mu_1, \ldots, \mu_{2^n-1})$. Since $\mu_0 = 0$ we only consider the vector $\mu^+ = (\mu_1, \ldots, \mu_{2^n-1})$ and its corresponding Möbius transform $\mathbf{m}^+ = (m_1, \ldots, m_{2^n-1})$.

Then, the Choquet integral defined in Definition 1 can be rewritten in terms of the Möbius transform of the fuzzy measure as:

$$\begin{aligned} CI_\mu(c_1, \ldots, c_n) &= \sum_{i=1}^{n} ((c_{s(i)} - c_{s(i-1)}) \Bigg(\sum_{A \subset A_{s(i)}} m(A) \Bigg) \\ &= \sum_{i=1}^{n} \sum_{A \subset A_{s(i)}} ((c_{s(i)} - c_{s(i-1)}) m(A)) \end{aligned}$$

where $c_{s(i)}$ denotes the ith lowest value in $(c_1, \ldots, c_n)$, and $A_{s(i)} = \{c_{s(i)}, \ldots, c_{s(n)}\}$.

In our case the data vector c is the vector of distances between variables of two records a and b such as $d(a, b) = (d(V_1(a), V_1(b)), \ldots, d(V_n(a), V_n(b)) = (d_1(a, b), \ldots, d_n(a, b))$. We can define the vector $\mathbf{d}^+(a, b) = (d_1^+(a, b), \ldots, d_{2^n-1}^+(a, b))$, where each element corresponds to:

$$d_r^+(a, b) = \sum_{i=1}^{n} (d_{s(i)}(a, b) - d_{s(i-1)}(a, b)) \cdot \tau_{i,r}$$

for $r = 1, \ldots, 2^n - 1$, where

$$\tau_{i,r} = \begin{cases} 1 & \text{if} \delta(B) = k \text{ for } B \subseteq A_{s(i)} \\ 0 & \text{otherwise} \end{cases}$$

So the Choquet integral can be defined in terms of $\mathbf{d}^+$ and $\mathbf{m}^+$ as:

$$CI_\mu(d(a, b)) = \mathbf{d}^+(a, b) \cdot \mathbf{m}^+$$

Now the minimization problem can be expressed as:

$$\textit{Minimize} \sum_{i=1}^{N} K_i \tag{13}$$

Subject to

$$\sum_{i=1}^{N}\sum_{j=1}^{N}[((\mathbf{d}^{+}(a_i,b_j)-\mathbf{d}^{+}(a_i,b_i))\cdot\mathbf{m}^{+})+CK_i]>0$$

$$\text{for all } i\neq j \tag{14}$$

$$K_i\in\{0,1\} \tag{15}$$

$$\sum_{k=1}^{2^n-1} m_k = 1 \tag{16}$$

$$\sum_{B'\subset B} m_k(B') - \sum_{A'\subset A} m_k(A') \geq 0 \quad \text{for all } A\subset B \tag{17}$$

The number of constraints is: N^2 for (14); N for (15); 1 for (16); and $\sum_{k=2}^{n}\binom{n}{k}k$ for (17).

To solve the problem defined above, we used the simplex optimizer algorithm from the IBM ILOG CPLEX tool (IBM 2010), (version 12.1). The problem is first expressed into the MPS (Mathematical Programming System) format, and then, processed with the optimization solver. Due to requirements of the CPLEX software we also added a constraint for the possible values of the Möbius transform of the fuzzy measure as: $-(n-1)\leq m_k\leq n-1$ for $k=1\ldots(2^n-1)$.

4 Evaluation

To evaluate our proposal we have tested it with different protected files. In each case we attempt to link the protected version of the dataset with the original one to evaluate the disclosure risk. The protection method used is *microaggregation*, 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 centroids of the corresponding clusters. The parameter k determines the protection level: a greater k implies greater protection (and greater information loss). For further information about microaggregation the reader is referred to Defays and Nanopoulos (1993), Torra (2004, 2008), Domingo-Ferrer and Torra (2005). The protection has been performed using the *sdcMicro* R package (Templ 2008; Templ and Petelin 2009).

We have considered files with the following protections:

- *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 variables with $k=3$, next 2 variables with $k=8$, and last 2 with $k=5$.
- *M6-853*: 6 variables, first 2 variables with $k=8$, next 2 variables with $k=5$, and last 2 with $k=3$.
- *M7-999*: 7 variables, first 6 microaggregated in groups of 3. These two groups and the single variable are protected with $k=3$.

In each case, we have protected 400 records randomly selected from the Census dataset (U.S. Census Bureau 2010) from the European CASC project (Brand et al. 2002), which contains

1080 records and 13 variables, and has been extensively used in other works (Domingo-Ferrer et al. 2001, 2006; Laszlo and Mukherjee 2005; Domingo-Ferrer and Torra 2005; Yancey et al. 2002).

Note that in our experiments we apply different protection degrees to different variables. This is especially interesting when variables have different sensitivity. There are more sensitive attributes that need more perturbation (protection) than others.

4.1 Improvement in record linkage

We compare our method based on the Choquet integral d^2CI to the standard record linkage d^2AM which uses a simple mean to aggregate the distances for each variable of the records, and a record linkage using a weighted mean d^2WM to aggregate such distances (as described in Torra et al. 2010). The d^2WM also uses the same supervised learning approach as described in Sect. 3.2 to obtain the optimal weight vector.

Table 1 shows the ratio of linked records for each file and record linkage method. The ratio determines the correctly identified records from the total, so a ratio of 1 means a 100% re-identification.

There is an important increase of the record linkage performance with d^2WM and d^2CI when compared to d^2AM, and a minor increase of d^2CI compared to d^2WM.

In Table 2 we show the computation time, expressed in seconds, for each of the tests corresponding to each protection dataset with the weighted mean (d^2WM) and the Choquet integral (d^2CI) approaches. In this table we can observe an important difference of computation time between both methods. This is so because the Choquet integral approach has to satisfy a higher number of constraints to solve the problem.

Table 1 Improvement in the linkage ratio

	d^2AM	d^2WM	d^2CI
M4-33	0.84	0.955	0.9575
M4-28	0.685	0.93	0.9375
M4-82	0.71	0.9425	0.9425
M5-38	0.3975	0.905	0.9125
M6-385	0.78	0.9925	0.9975
M6-853	0.8475	0.9875	0.9925
M7-999	0.8775	0.915	0.9725

Table 2 Computation time consumed in seconds

	d^2WM	d^2CI
M4-33	0.38	158.52
M4-28	0.37	1626.29
M4-82	0.39	469.53
M5-38	419.79	355923.05(∼ 99 h)
M6-385	4.4	49.69
M6-853	7.7	128.5
M7-999	9019.78(∼ 2.5 h)	21682.29(∼ 6 h)

Table 3 Fuzzy measure for *M4-28*

k	μ_k	k	μ_k
0000	0.000000	1000	0.999168
0001	0.154253	1001	0.999799
0010	0.003651	1010	0.999800
0011	0.207612	1011	0.999899
0100	0.039292	1100	0.999268
0101	0.154353	1101	0.999899
0110	0.096721	1110	0.999900
0111	0.207712	1111	0.999999

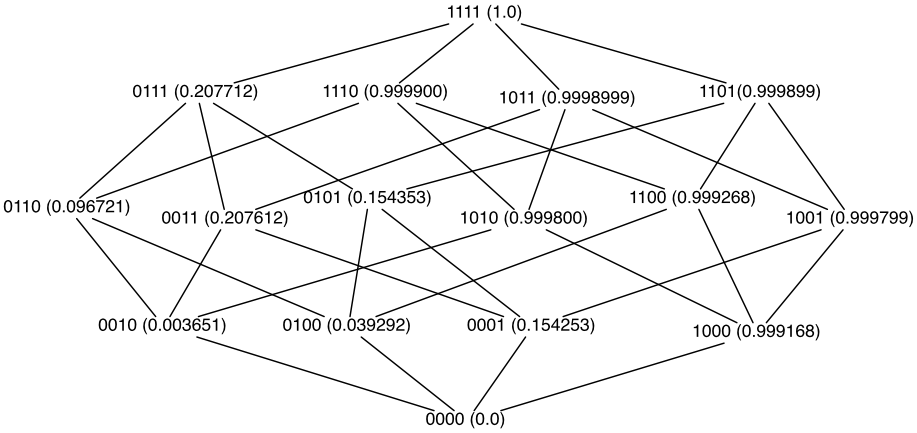

Fig. 1 Fuzzy measure lattice for *M4-28*. Dyadic representation of the set and measures in brackets

4.2 Relevance of attributes determined by the optimal weights

Once we solve the optimization problem with d^2CI (c.f. Sect. 3.3) for a concrete dataset, we can reconstruct the original fuzzy measure from the Möbius transform obtained using (18).

$$\mu(A) = \sum_{B \subseteq A} m(B) \tag{18}$$

for all $A \subseteq V$.

This fuzzy measure provides valuable information about the relevance of each subset of variables in the re-identification process that maximize the number of correctly linked records.

Table 3 shows the fuzzy measure obtained for *M4-28* and Fig. 1 shows the lattice representation of this fuzzy measure. Each subset of variables $A \subseteq V$ is represented by the dyadic representation of its index k as described in Sect. 3.3. We can see that the first variable provides a high degree of information for the re-identification. Note that all subsets which include it have a weight greater than 0.999. This variable has been protected with $k = 2$, which preserves more information (is less distorted) than the last two variables (recall that they are protected with $k = 8$). It is also interesting to note that the highest weight of a two element subset is the one which includes the first and third variables. Each of these variables

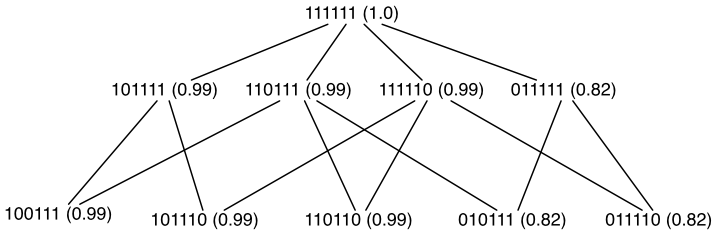

Fig. 2 Partial fuzzy measure lattice for *M6-385* including all measures with values larger than 0.1

Table 4 Fuzzy measure for *M6-853*

k	μ_k	k	μ_k
000000	0.0000000000	010000	0.0004214525
000001	0.0007378150	010001	0.0008378150
000010	0.0000000000	010010	0.0220367576
000011	0.0220367576	010011	0.0221367576
000100	0.0000000000	010100	0.0640252746
000101	0.0640252746	010101	0.0641252746
000110	0.0640252746	010110	0.0641252746
000111	0.0641252746	010111	0.8247279668
001000	0.0019057155	011000	0.0127378590
001001	0.0160771077	011001	0.0213887564
001010	0.0220367576	011010	0.0221367576
001011	0.0221367576	011011	0.0222367576
001100	0.0020057155	011100	0.0641252746
001101	0.0641252746	011101	0.0642252746
001110	0.0641252746	011110	0.8247279668
001111	0.0642252746	011111	0.8248279668
100000	0.0081683003	110000	0.0221158311
100001	0.0221158311	110001	0.0222158311
100010	0.0221158311	110010	0.0222158311
100011	0.0222158311	110011	0.0424704875
100100	0.0082683003	110100	0.0641252746
100101	0.0641252746	110101	0.0642252746
100110	0.0641252746	110110	0.9998000000
100111	0.9998000000	110111	0.9999000000
101000	0.0082683003	111000	0.0222158311
101001	0.0222158311	111001	0.0223158311
101010	0.0423704875	111010	0.0424704875
101011	0.0424704875	111011	0.0425704875
101100	0.0083683003	111100	0.0642252746
101101	0.0642252746	111101	0.0643252746
101110	0.9998000000	111110	0.9999000000
101111	0.9999000000	111111	1.0000000000

Table 5 Weight vector for *M6-853*, when using d^2WM

k	*weight*
100000	0.016809573957189
010000	0.00198841786482128
001000	0.00452923777074791
000100	0.138812880222131
000010	0.835523953314578
000001	0.00233593687053289

correspond to different protected blocks (one with $k = 2$ and the other with $k = 8$). So our approach is useful to detect that to combine variables with complementary information is useful in re-identification.

Also interesting is to observe the case of the fuzzy measure for the files with 6 variables. Table 4 shows the fuzzy measure for *M6-853*, and Fig. 2 the lattice representation of the measure for all subsets with a weight $\mu_k \geq 0.1$.

Note, for example that the sets of four elements (leaves from last row in Fig. 2) all include at least one element of each block of variables. That is, one element of the variables microaggregated with $k = 8$, one with $k = 5$, and one with $k = 3$. As stated before all these variables provide complementary information, which helps in the linkage process.

We also show in Table 5 the weights obtained for the same dataset if we compute the weighted mean distance d^2WM. In this case the most important variable seems be the 5th one. It comes as no surprise that this variable is present in all the measures from Fig. 2. Note also that measures for sets which differ in the presence of the second and last variables are approximately the same. These variables do not seem to provide useful information for the record linkage.

5 Conclusions

In this paper we have introduced a distance based record linkage for the evaluation of the disclosure risk in data privacy. Our proposal uses the Choquet integral and a fuzzy measure to determine the relevance of each variable (and the interaction between variables) in the linkage process. We have provided a supervised learning approach to determine the optimal fuzzy measure for the linkage, which also provides information about the variables and their interactions.

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.

References

Batini, C., & Scannapieco, M. (2006). *Data quality: concepts, methodologies and techniques series (data-centric systems and applications)*. New York: Springer

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.

Choquet, G. (1953). Theory of capacities. *Annales de L'Institut Fourier*, *5*, 131–295.
Colledge, M. (1995). *Frames and business registers: an overview. business survey methods. Wiley series in probability and statistics*. New York: Wiley.
Data.gov.uk (2010). UK Government.
Data.gov (2010). USA Government.
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, statistics*, Canada (pp. 195–204).
Domingo-Ferrer, J., & Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. Lane, J. Theeuwes, & L. Zayatz (Eds.), *Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies* (pp. 111–133). Amsterdam: Elsevier.
Domingo-Ferrer, J., & Torra, V. (2005). Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, *11*(2), 195–212.
Domingo-Ferrer, J., Mateo-Sanz, J. M., & Torra, V. (2001). Comparing sdc methods for microdata on the basis of information loss and disclosure risk. In *Preproceedings of ETK-NTTS 2001* (Vol. 2, pp. 807–826). Luxembourg: Eurostat.
Domingo-Ferrer, J., Torra, V., Mateo-Sanz, J. M., & Sebe, F. (2006). Empirical disclosure risk assessment of the ipso synthetic data generators. In *Monographs in official statistics-work session on statistical data confidentiality* (pp. 227–238). Luxembourg: Eurostat.
Dunn, H. L. (1946). Record Linkage. *American Journal of Public Health*, *36*(12), 1412–1416.
Elmagarmid, A., Panagiotis, G., & Verykios, V. (2007). Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 1–16.
Fellegi, I., & Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210.
Hartley, H. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, *14*, 174–194.
IBM (2010). IBM ILOG CPLEX, High-performance mathematical programming engine. International Business Machines Corp. http://www-01.ibm.com/software/integration/optimization/cplex/.
Jaro, M. A. (1989). Advances in record linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Society*, *84*(406), 414–420.
Lane, J., Heus, P., & Mulcahy, T. (2008). Data access in a cyber world: making use of cyberinfrastructure. *Transactions on Data Privacy*, *1*(1), 2–16.
Laszlo, M., & Mukherjee, S. (2005). Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, *17*(7), 902–911.
McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions. Wiley series in probability and statistics*. New York: Wiley.
Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science*, *130*, 954–959.
Pagliuca, D., & Seri, G. (1999). Some results of individual ranking method on the system of enterprise accounts annual survey. Esprit SDC Project, Delivrable MI-3/D2.
Statistics Canada (2010). Record linkage at Statistics Canada. http://www.statcan.gc.ca/record-enregistrement/index-eng.htm.
Templ, M. (2008). Statistical disclosure control for microdata using the R-Package sdcMicro. *Transactions on Data Privacy*, *1*(2), 67–85.
Templ, M., & Petelin, T. (2009). A graphical user interface for microdata protection which provides reproducibility and interactions: the sdcMicro GUI. *Transactions on Data Privacy*, *2*(3), 207–224.
Torra, V. (2004). Microaggregation for categorical variables: a median based approach. In *Lecture notes in computer science: Vol. 3050. Proc. privacy in statistical databases (PSD 2004)* (pp. 162–174). Berlin: Springer.
Torra, V. (2008). Constrained microaggregation: adding constraints for data editing. *Transactions on Data Privacy*, *1*(2), 86–104.
Torra, V., & Narukawa, Y. (2007). *Modeling decisions: information fusion and aggregation operators*. Berlin: Springer.
Torra, V., Abowd, J. M., & Domingo-Ferrer, J. (2006). Using Mahalanobis distance-based record linkage for disclosure risk assessment. In *Lecture notes in computer science: Vol. 4302. Privacy in statistical databases 2006* (pp. 233–242). Berlin: Springer.
Torra, V., Navarro-Arribas, G., & Abril, D. (2010). On the applications of aggregation operators in data privacy. In *Advances in soft computing (integrated uncertainty management and applications): Vol. 68. International symposium on integrated uncertainty management and applications* (pp. 479–488).
U.S. Census Bureau (2010). Data Extraction System. http://www.census.gov/.

Winkler, W. E. (2003). Data cleaning methods. In *Ninth ACM SIGKDD international conference on knowledge discovery and data mining*.

Winkler, W. E. (2004). Re-identification methods for masked microdata. In *Lecture notes in computer science: Vol. 3050. Privacy in statistical databases, PSD 2004* (pp. 216–230). Berlin: Springer.

Yancey, W., Winkler, W., & Creecy, R. (2002). Disclosure risk assessment in perturbative microdata protection. In *Lecture notes in computer science: Vol. 2316. Inference control in statistical databases* (pp. 135–152). Berlin: Springer.