2006 IEEE International Conference on Fuzzy Systems
Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada
July 16-21, 2006

Establishing a benchmark for re-identification methods and its validation using fuzzy clustering

Vicenç Torra *Senior Member, IEEE*, Josep Domingo-Ferrer *Senior Member, IEEE*

Abstract— **Privacy preserving data mining and statistical disclosure control are related fields with increasing importance nowadays. They aim is to allow the publication of sensible data without compromising the privacy of data respondents. To that end, masking methods have been designed so that data are distorted in a way that preserves confidentiality and data utility. Alternatively, methods have been constructed to generate synthetic data that have properties similar to the ones of the original data.**

At the same time, recent research in re-identification methods (record and variable matching) has been pushed forward due to the current interest on security issues and the huge amount of data stored in databases. However, there is no standard methodology for comparing alternative re-identification methods.

In this paper we propose the use of masking methods and syntethic data generators for building benchmarks for matching methods. We validate our approach using fuzzy clustering.

I. Introduction

In recent years, the importance of re-identification methods has substantially increased due to several factors:

- On one hand, the so-called information explosion causes that the amount of data available at our fingertips is enormous. While in the past only a few records about individuals or households were stored in each circumstance, at present times, large amounts of information are recorded. [28] estimates that the disk storage per person (DSP) expressed in megabytes (MB) was about 472 in 2000 while it was only 28 in 1996 and 0.02 in 1983. Nevertheless, data in a company or institution is typically spread across distributed platforms and systems, and hardly homogeneous. To make data available, reliable and consistent, they have to be integrated. Re-identification methods play a central role in this process.
- On the other hand, current data security requirements stress the need for re-identification. This means linking databases from multiple sources to find relevant information concerning about relevant suspect individuals. To this purpose, data must be found that match some particular requirements.

Beyond the above, let us say, *generic* re-identification methods, algorithms have been developed in some particular fields for re-identification or matching under specific constraints. This is the case, for example, of graph matching algorithms [17], [16] that are commonly used in computer vision.

There is a rich literature on re-identification methods (see *e.g.* [21]). These can be classified according to the type of data they try to re-identify (records, variables), the assumptions made on data (independence between variables, variables in ordinal scales) and the kind of data used in the process (meta-knowledge on the variables).

Nevertheless, there is no standard method for comparing re-identification methods and establishing their performance in an effective way.

In this work, we propose a data generator and a benchmark for comparing re-identification methods. The set of files that define the benchmark can be used to study different aspects of re-identification.

The structure of the paper is as follows. Section II reviews re-identification methods, by sketching the differences between them and highlighting what kind of information is needed for comparing them. Then, in Section III we review some methods developed in statistical disclosure control and privacy preserving data mining for data protection. These methods are the basis of our benchmark. Then, in Section IV we propose the data generators and the benchmark. In this section we also show that fuzzy c-means permits to validate our approach of generating data. The paper finishes in Section V with some conclusions.

II. Re-identification methods

Two major classes of re-identification methods can be distinguished: schema matching and record linkage (record matching) tools. The former attempt to link the schema of two different databases by establishing relationships between the elements that define the schema. On the other hand, record linkage tools attempt to link records in two different files that contain information on the same individuals. Thus, schema matching is re-identification of variables and record linkage is re-identification of records.

We review both classes below, starting with schema matching.

- **Schema matching tools:** These tools are developed to find correspondences between schemata of two or more databases. Although there are several of such tools with a variety of goals, the most developed work is for the case of finding correspondences between variables. These tools are the so-called *attribute* (variable) *correspondence identification methods*. In this case, the goal is to find which variable in a database B corresponds to a particular variable in a database A.
  The class of variable correspondence identification methods is large. Some of the methods are solely based

Vicenç Torra is with the IIIA-CSIC, Campus UAB s/n, 08193 Bellaterra, Catalonia, Spain (phone: +34 935809570; fax: +34 935809661; e-mail: vtorra@iiia.csic.es); Josep Domingo-Ferrer is with the Dept. Enginyeria Informàtica i Matemàtiques, URV, Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain (e-mail: josep.domingo@urv.cat).

0-7803-9489-5/06/$20.00/©2006 IEEE

on the names of variables and data types while others use some kind of structural information (*e.g.*, [2]) or some characteristics of data instances (see *e.g.* [6] for details). Other methods combine the information in the database with some other background knowledge. This is the case of the system described in [2] that uses knowledge about the variables and terms as represented in the Wordnet [10] dictionary. Recently, hybrid approaches have also been considered that are inspired in the machine learning ensemble methods. They combine the results of several re-identification methods. *E.g.* see COMA [6] that provides an extensible library of variable-matching algorithms and a set of methods to combine matching results.

The process of schema matching can be enhanced by using operators for data transformation. For example, user-defined functions supported in SQL:99 for data transformations are used in [25]. In general, such operations permit new views from an initial table to be defined so that the schema matching procedure can be found more easily. This is the case, for example, when decomposing a *raw* `Address` field in a file, say A, into the following variables: `Address`, `City`, `Zip code`, `State` and `Country`. In this way, the `Zip code` in a file B can easily be matched with the corresponding `Zip code` in file A.

More general methods for schema matching beyond variable correspondence identification methods exist. They correspond to structure-level matching methods. They refer to matching combinations of elements. [24] gives a thorough account of such methods. The simplest case corresponds to establishing a one-to-many (or a many-to-one) relation; in that case, one variable is linked to a set of variables as in the example for the variable `Address`.

- **Record linkage methods:** The goal of record linkage (or record matching) methods is to establish correspondences between records of two or more databases. This is, given two databases A and B, and given a record r in a file A describing an object o, the goal is to find which record in file B (if any) contains information about the same object o. This problem is sometimes referred to as the object identity problem, the duplicate elimination or the merge/purge problem (*e.g.*, [25]).

  Several approaches have been developed for record linkage. The most common one, denoted here by classical record linkage, is applicable when both files A and B share a set of variables. In this case, comparison between records can be done on the basis of these common variables. Types of classical record linkage are probabilistic [32], [33], distance-based [23], [7] and clustering-based [1] methods. Differences between methods are due to the theoretical background and the assumptions on the independence of variables.

  In recent years, an alternative approach has been developed consisting of record linkage when files do not share variables. In this case (see [29], [30]), re-identification relies on the assumption that there is some common structure underlying both files. This is, there is common structural information in both files that can be exploited for re-identification. Partitions are one possible way of expressing structural information. The idea is that, if two records are similar in one file, their corresponding records in the other file should also be similar; if two records are clustered together in one file, one would expect that their corresponding records will be clustered together in the other file.

It must be said that, even though schema matching and record linkage are distinct types of re-identification, most methods can be applied at both record and variable level: if a data set is regarded as a matrix where columns correspond to variables and rows to records, then the data matrix can be transposed, so that variables become records and conversely. Using this fact, the methods described in [29], [9] have been applied to both record and variable re-identification.

Depending on their precise aim, a finer classification of re-identification methods is possible:

- Record level re-identification (common variables). The case of variables with the same semantic meaning but with different domains (*e.g.* different granularity of the terms used in the two data files) is also classified here.
- Record level re-identification (non-common variables)
- Variable level re-identification (seeking to establish a one-to-one mapping between variables in two different schemata)
- Variable decomposition (seeking to establish a many-to-one mapping)
- Schema re-identification (seeking to establish a many-to-many mapping)

Therefore, data generators and data sets used for benchmarking algorithms must be suitable for testing all categories of methods listed above. Moreover, data should be rich enough so that new approaches such as algorithms not assuming independence between variables or attempting to link files with no shared variables can be tested.

III. Methods for data protection

Current methods for data protection can be classified in two large groups: i) masking methods, that apply distortion techniques to original data files to obtain protected data files; ii) synthetic data generators, that use original data to generate models and these models to generate artificial data. We review below both kinds of methods (see also [34] for a recent survey).

A. Masking methods

There exists several methods to generate distorted files from original files in order to preserve confidentiality. We mention below some methods that are appropriate for the purpose of this work.

- **Additive noise:** This method consists of adding noise to the data. Noise is added so that the properties of the original data (*e.g.* means and covariances) are exactly or approximately preserved.

- **Global recoding:** This method, suitable for categorical data, consists of replacing some of the categories by more general ones. This can be seen as a change in the granularity level. Global recoding can be formalized considering categories in a tree-like structure. Then, categories situated on a leaf are replaced by the one in a node at a higher position in the hierarchy.
- **Microaggregation:** This method, initially developed for numerical data, has been recently incorporated into the μ-Argus [13] system also for categorical data. The idea is to build clusters of similar records and replace records in a cluster by the corresponding centroid. This situation can be seen as a quantization process or as a change in the granularity level (from higher to lower granularity). A variation of this method was defined using fuzzy clustering [31]. The method avoids the disclosure of information about the parameters to the intruder by inspection of the protected data file.
- **Rank swapping:** This method, together with microaggregation, has been shown to be one of the best performers for masking data from the point of view of the trade-off between disclosure risk and information loss. It yields data files with low disclosure risk and low information loss. The method, usable on both numerical and categorical data, consists of replacing values with similar ones (for each variable, values are ranked and then each value is swapped with another one within a restricted range of ranks).

All the above methods can be parameterized so that the parameter corresponds to the degree of distortion to be applied to the file. The following parameterizations have been considered in the literature:

- **Additive noise:** Noise is generated using a $N(0, ps)$ where s is the standard deviation of the original variable and p is a parameter. The larger p, the larger the distortion.
- **Global recoding:** The more categories are recoded, the larger the degree of distortion.
- **Microaggregation:** The minimum number of records in a cluster is the usual parameter. The larger the number of records, the more difficult is re-identification.
- **Rank swapping:** The parameter corresponds to the range of ranks within which swapping is allowed. In other words, the ranks of two swapped values cannot differ by more than p percent of the total number of records.

B. Synthetic data

Publication of simulated —*i.e.* synthetic— data was proposed long ago as a way to guard against statistical disclosure. In fact, as early as 1993, Rubin [26] suggested creating an entirely synthetic data set based on the real survey data and multiple imputation. Specific case studies of synthetic microdata generated by multiple imputation were presented in [18], [19]. Although the results were fairly promising, the multiple imputation approach requires complex models and software, which greatly reduces its appeal in many situations.

In [9], [8] comparisons were presented for measuring the performance of microdata masking methods in terms of information loss and disclosure risk. Based on the proposed measures, it was shown in [27] how to improve the performance of any particular masking method. In particular, post-masking optimization was discussed for preserving as much as possible the moments of first and second order (and thus multivariate statistics) without increasing the disclosure risk. The technique proposed could also be used for synthetic microdata generation and could be extended for preservation of all moments up to m-th order, for any m. The shortcoming of this approach is its computational complexity: the optimization problem is solved using an iterative refinement approach, which may be quite time-consuming when the involved data sets are large.

Latin Hypercube Sampling (LHS) appears in the literature as another method for generating multivariate synthetic data sets. In [14], authors improve the LHS updated technique of [12], but the proposed scheme is still time-intensive even for a moderate number of records. In [4], LHS is used along with a rank correlation refinement to reproduce both the univariate (*i.e.* mean and variance) and multivariate structure (in the sense of rank correlation) of the original data set. This method also permits flexibility in the size of the synthetic data set that is generated. In summary, LHS-based methods rely on iterative refinement, are time-intensive and their running time does not only depend on the number of values to be reproduced, but on the starting values as well.

In [20], a non-iterative method for generating continuous synthetic microdata is proposed. The implementation of this method results in a fast algorithm which *exactly* reproduces the means and the covariance matrix of the original data set and whose running time grows *linearly with the number of records*. Exact preservation of the original covariance matrix implies that variances and Pearson correlations are also exactly preserved in the synthetic data set. Like in any synthetic data generator, the number of records in the synthetic data set can differ from the number of records in the original data set.

IV. On the construction of benchmarks

To construct the benchmarks, we have considered different approaches for each type of problem described in Section II.

A. Record level re-identification: case of common-variables

Masking methods are a suitable tool for generating files for studying re-identification methods at record level. For each original file A, the application of a masking method MM with a given parameterization p yields a masked file $A_{MM,p}$ that contains the same records in A but distorted in a particular way. Records in file $A_{MM,p}$ are, in principle, difficult to re-identify with those in A.

Moreover, for most masking methods, parameter p corresponds to the protection level and, therefore, variations

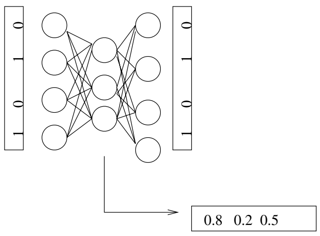

Fig. 1. Graphical representation of the approach of generating files for non-common variables using neural networks.

of this parameter correspond to a *re-identification difficulty*. Roughly speaking, the re-identification difficulty is inverse to the number of successfuls re-identifications that can be obtained.

Some masking methods, such as global recoding, are particularly suitable for analyzing the behaviour of re-identification methods when variables have different domains in both files (*i.e.*, when they use different granularity). Global recoding replaces original categories with more general ones. Standard microaggregation has a similar effect for numerical data (replacing several similar numerical data by a single representative of this data - a kind of fine grain quantization).

B. Record level re-identification: case of non common-variables

We consider several methods for generating data for record level re-identification in a situation where records do not share any variables.

The simplest way to re-create the situation of non-common variables is to split a data file into two new data files so that all records are included in both files but only a disjoint half of the variables is in each file. Re-identification in this context is possible whenever there are correlated variables in the original file which are distributed across the two files after splitting. The weaker the correlations among variables across files, the more difficult is re-identification.

We propose here an alternative, but more complex, approach for the same process. It consists on using neural networks to construct artificial variables. This is based on the application of neural networks to data compression (see [11] for details). This approach is represented graphically in Figure 1. Let us first consider the learning phase of the network. In this phase, let us consider an original data file A with N variables and a neural network with N inputs, N outputs and $n << N$ neurons in the hidden layer. Then, we train the neural network with the pairs (a, a). That is, we train the neural network so that when the input is a record a in the file A, the output of the neural network should be equal to its input. That is, the output of the network in this case is also a. Now, once the neural network has learnt the patterns (a, a), it is used to produce the new file. The new file is obtained through the application of the network to each record a in the original file A, and then selecting as the new records in the file the values of the neurons of the hidden layer. In a more formal way, let $h(a)$ denote the values of the hidden layer when the neural network is applied to the record a. Then, we define the new file A' as containing the records $h(a) = (h_1(a), \ldots, h_n(a))$ for all a in A.

In Figure 1 both learning and file-generating phases are represented. In the learning phase, we would train the network with pairs of the form $((1, 0, 1, 0), (1, 0, 1, 0))$ as in the figure. Then, for each record a in A (as $(1, 0, 1, 0)$) we would obtain its $h(a)$ (*e.g.*, $(0.8, 0.2, 0.5)$) and define the new file with such values.

Naturally, the goal is to reidentify a with $h(a)$. As there are no common variables, this is a problem for re-identification with non-shared variables. In this approach, the smaller n, the more difficult is re-identification.

C. Variable level re-identification

The use of masking methods allows the study of variable level re-identification algorithms. However, the generation of synthetic data permits a more challenging study of such algorithms. Such algorithms generate records in such a way that (selected) properties of the original file are present in the records. Re-identification methods can be applied to re-identify the variables that define the new records with the variables used to describe the original records.

We can distinguish the following advantages of using synthetic data for testing and analyzing variable level re-identification:

- The information not explicitly represented in a data model for synthetic data is usually not present in the artificially generated data. In some circumstances, such missing information would make the re-identification process difficult. This is similar to the situation in which re-identification is applied to two files representing two partially different sets. In this case, the model of the data can be partially different.
- Synthetic data can have an arbitrary number of variables. Therefore, the original and the artificial sets can have different number of variables. This situation is quite realistic and is a stress test for re-identification methods.
- Synthetic data allows the creation of data files with an arbitrary number of records. Therefore, the number of records in the original and artificial file are not the same. This is also realistic, but has only a limited relevance for the study of variable level re-identification.

Furthermore, the application of a synthetic data generator to an already masked data set makes variable level re-identification more difficult. This is especially the case when a masking method for changing the granularity, such as global recoding, is used.

It should be underlined that, just as re-identification methods can be applied for both record level and variable level re-identification, some data generators can be used for generating data for both situations. Nevertheless, data generators and generated data have to be used with caution so that data are consistent with experimental observations. For example, variable level re-identification usually has more redundant information than record level, but at the same time all data are more homogeneous (variables are either numerical or categorical, but not mixed).

D. Variable decomposition

One of the most common applications for variable decomposition algorithms is the re-identification of individual addresses. Nevertheless, maintaining public repositories of such data for testing is a rather complex issue due to privacy concerns (although some census data, voting lists and *e.g.* yellow pages can easily be retrieved or legally bought). Thus, address repositories can hardly be used for testing.

An alternative field of application of such algorithms is re-identification of bibliographical records. Such records have a similar mixed structure of categorical and numerical data. Moreover, there are different conventions for describing them (*e.g.* MLA, Chicago, individual journals) and records in the web do not always follow any standard convention and are error-prone. The advantage of bibliographical repositories is that the entries in them are publicly available and can be used for testing without difficulty.

Therefore, bibliographical records are suitable data for defining benchmarks for re-identification algorithms related to variable decomposition. It is suitable to consider raw text bibliographical records against structured ones using *e.g.* XML or `bibtex` formalisms, or, alternatively, against structured ones in bibliographical databases. Such data permit the study of one-to-many mappings.

E. Schema re-identification

Schema re-identification can also be studied using databases with bibliographical information. In this case, we need to consider databases with different field names. It is also appropriate, and particularly challenging, to consider two databases with raw text bibliographical records.

F. Validating the approach

To test the validity of our approach to generate data for re-identification methods, we apply some indicators to the methods described in this section. Namely, these indicators are either the direct application of re-identification methods (to check that problems of increasing difficulty can be generated) or measures correlated with the re-identification results. We restrict below our description to data generation using masking methods and using the approach based on neural networks.

In Table I, we show results for some categorical data generated using Rank Swapping using data from the DES repository [5] (Data from the American Housing Survey 1993). We show that for increasing values of the possible range for swapping, the number of re-identifications diminish. The rank of two swapped values cannot differ by more than p percent of the total number of records. The results in Table I corresponds to values of p from 1 to 8. Similar results are obtained for numerical data.

TABLE I

PROPORTION OF RE-IDENTIFICATIONS FOR A CATEGORICAL FILE GENERATED FROM THE DES REPOSITORY USING RANK SWAPPING WITH AN INCREASING RANGE FOR SWAPPING.

Proportion Rank Swapping	Re-identified records
1	958.0
2	953.0
3	940.0
4	921.0
5	904.0
6	899.0
7	885.0
8	894.0

To evaluate the approach for data generation in the case of records with non common variables, we have generated several neural networks adjusted to the data according to the procedure described in Section IV-B. This is, we have generated new data files with the outputs of the hidden layer. As we have considered hidden layers of dimensions 1–13 (the original file was described in terms of 13 variables), this generated 13 different files. In this case, a data file (also extracted from the DES repository [5]) with numerical data was considered. The file used was the Current Population Survey corresponding to 1995. For the generation of the new data file, we have used the standard backpropagation algorithm [11].

Then, to check the validity of the approach, we have clustered each file (and the original data file) using the fuzzy c-means algorithm [3]. That is, we have considered the list of tuples in the original data file $a^1, \ldots, a^r$ and the tuples generated by the neural network $h(a^1), \ldots, h(a^r)$. Each a^i has N components $a^i_1, \ldots, a^i_N$ and each generated tuple has as much components as hidden neurons. That is, when the number of hidden neurons is n we have that $h(a^i) = (h_1(a^i), \ldots, h_n(a^i))$. Then, we have clustered both files $a^1, \ldots, a^r$ and $h(a^1), \ldots, h(a^r)$ obtaining c clusters in each file. Figure 2 represents the two files and the results of the clustering (the membership functions to the fuzzy clusters) for $c = 2$.

The next step has been to compute the correlation between the membership degrees of the records in the original file and the ones for each new generated file. Table II displays correlations for the case of this approach applied to the data file previously used in [7] for the dimension of the hidden layer varying between 1 and 13. The number of clusters considered in fuzzy c-means was equal to 2.

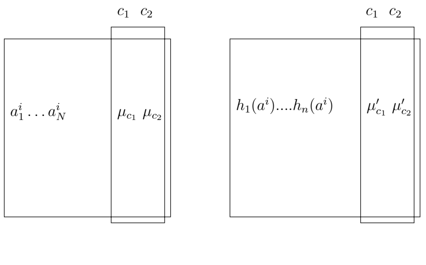

Fig. 2. Clustering the original file and the synthetic file.

TABLE II

CORRELATIONS TO EVALUATE THE APPROACH FOR RECORD LEVEL RE-IDENTIFICATION FOR THE CASE OF NON COMMON-VARIABLES

Dimension Hidden Layer	Correlation
1	0.6930
2	0.7278
3	0.4372
4	0.7579
5	0.7995
6	0.8599
7	0.8633
8	0.9376
9	0.9437
10	0.9460
11	0.9553
12	0.9766
13	0.9811

The rationale of this approach is based on clustering as an unsupervisd machine learning technique. In general, clustering tools are used to extract some knowledge from raw data. In our case, this role is played by the fuzzy c-means. Fuzzy c-means permits to extract some structure from each pair of files: (original data, generated data). As much similar is the underlying structure of the two files, the more similar are the clusters. Then, the more similar are the clusters, the more correlated are the memberships.

Therefore, the use of fuzzy c-means permits to evaluate in what extent the data in the two files contain such underlying structure. What is expected, and as can be observed in Table II this actually occurs, the larger the number of neurons in the hidden layer, the better the correlation is. That is, the larger the number of neurons, the better this hidden layer represents the information in the original file.

The good results in Table II permits to confirm that the method is appropriate to generate benchmarks for this kind of problem.

Two alternatives to fuzzy clustering might be considered for the process of determining the structural similarity between the two files. On the one hand, we could consider crisp clustering as *e.g.* c-means. As in this case the final membership would be crisp, the information contained in the final clusters would be lesser than the one in the fuzzy clustering. The fuzzy degree not only gives information about the membership but also, roughly speaking, about *e.g.* the distance between the object and the cluster center. Such information does not appear in the result of the crisp clustering. On the other hand, we could consider hierarchical clustering. Nevertheless, hierarchical clustering is sensitive to small variations of the data. Due to the fact that the two files contain data from a completely different origin, we considered difficult that the final structures obtained for such clustering algorithm matched. Due to this, we consider that a fuzzy clustering technique is a better approach.

V. CONCLUSIONS

In this work we have argued the need of developing methods to generate data sets for testing re-identification methods. We have described the main requirements for data and proposed several approaches to supply such data. Then, we have evaluated the use of masking methods and a neural networks approach for generating such data.

Additionally, we will consider other fuzzy clustering techniques as variable-size fuzzy c-means [15], [22] that might be able to extract some additional information

ACKNOWLEDGMENTS

This work was partly funded by the Spanish Ministry of Education and Science under project SEG2004-04352-C04-01/02 ”PROPRIETAS”.

REFERENCES

[1] J. Bacher, R. Brand and S. Bender, ‘Re-identifying register data by survey data using cluster analysis: an empirical study’, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10:5 (2002) 589-608.

[2] S. Bergamaschi, S. Castano, M. Vincini and D. Beneventano, ‘Semantic integration of heterogeneous information sources’, *Data and Knowledge Engineering* 36 (2001) 215-249.

[3] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York. 1981.

[4] R. A. Dandekar, M. Cohen and N. Kirkendall, ‘Sensitive micro data protection using latin hypercube sampling technique’, in *Inference Control in Statistical Databases*, vol. LNCS 2316, pp. 245-253, Springer, 2002.

[5] Data Extraction System (DES), U. S. Census Bureau, http://www.census.gov/DES/www/welcome.html

[6] H.-H. Do and E. Rahm, ‘COMA - A system for flexible combination of schema matching approaches’, *Proc. of the 28th VLDB Conference*, Hong-Kong, China, 2002.

[7] J. Domingo-Ferrer and V. Torra, ‘A quantitative comparison of disclosure control methods for microdata’, in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. L. Zayatz, P. Doyle, J. Theeuwes and J. Lane, Amsterdam: North-Holland, 2001, pp. 111-134.

[8] J. Domingo-Ferrer, J. M. Mateo-Sanz and V. Torra, ‘Comparing SDC methods for microdata on the basis of information loss and disclosure risk’, in *Proceedings of ETK-NTTS 2001*, Luxemburg: Eurostat, pp. 807-825, 2001.

[9] J. Domingo-Ferrer, V. Torra, ‘Disclosure risk assessment in statistical disclosure control via advanced record linkage’, Statistics and Computing, 13 (2003) 343-354.

[10] C. Fellbaum, *WordNet: An Electronic Lexical Database*, The MIT Press, 1998.

[11] J. A. Freeman, D. M. Skapura, Neural Networks: algorithms, applications, and programming techniques, Addison-Wesley, 1991.

[12] A. Florian, ‘An efficient sampling scheme: updated latin hypercube sampling’, *Probabilistic Engineering Mechanics*, no. 7, pp. 123-130, 1992.

[13] A. Hundepool, The CASC Project, Lecture Notes in Computer Science 2316, 172-180, 2002.

[14] D. E. Huntington and C. S. Lyrintzis, ‘Improvements to and limitations of Latin hypercube sampling’, *Probabilistic Engineering Mechanics*, vol. 13, no. 4, pp. 245-253, 1998.

[15] H. Ichihashi, K. Honda, N. Tani, (2000), Gaussian mixture PDF approximation and fuzzy c-means clustering with entropy regularization, Proc. of the 4th Asian Fuzzy System Symposium, May 31-June 3, Tsukuba, Japan, 217–221.

[16] H. Kalviainen and E. Oja, ‘Comparisons of attributed graph matching algorithms for computer vision’. *Proc. of STEP-90, Finnish Artificial Intelligence Symposium*, pages 354–368, Oulu, Finland, June 1990. Also at http://citeseer.nj.nec.com/kalviainen90comparisons.html

[17] M. Karpinski, *Fast Parallel Algorithms for Graph Matching Problems: Combinatorial, Algebraic & Probabilistic Approach*, Oxford University Press, 1998.

[18] A. B. Kennickell, ‘Multiple imputation and disclosure control: the case of the 1995 Survey of Consumer Finances’, in *Record Linkage Techniques*, Washington DC: National Academy Press, 1999, pp. 248-267.

[19] A. B. Kennickell, 'Multiple imputation and disclosure protection: the case of 1995 Survey of Consumer Finances', in *Statistical Data Protection*, Luxemburg: Office for Official Publication of the European Communities, 1999, pp. 381-400.

[20] J. M. Mateo-Sanz, A. Martínez-Ballesté and J. Domingo-Ferrer, 'Fast generation of accurate synthetic microdata', in J. Domingo-Ferrer and V. Torra (eds.) *Privacy in Statistical Databases*, Berlin: Springer-Verlag, Lecture Notes in Computer Science 3050 (2004) 298-306.

[21] A. McCallum, S. Tejada, D. Quass, (Eds.), *Proc. 1st Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, Washington D.C., 2003 (in conjunction with 9th ACM SIGKDD Int. Conf. on KD &DM). Also at http://csaa.byu.edu/kdd03cleaning.html

[22] S. Miyamoto, K. Umayahara, (2000), Fuzzy *c*-means with variables for cluster sizes, 16th Fuzzy System Symposium, Akita, Sept.6-8, 537–538 (in Japanese).

[23] D. Pagliuca and G. Seri, *Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey*, Esprit SDC Project, Deliverable MI-3/D2, 1999.

[24] E. Rahm and P. A. Bernstein, A survey of approaches to automatic schema matching, *The VLDB Journal*, 10 (2001) 334-350.

[25] E. Rahm and H. Hai Do, Data Cleaning: Problems and Current Approaches, *Bulletin of the Technical Committee on Data Engineering*, 23:4 (2000) 3-13.

[26] D. B. Rubin, 'Discussion on statistical disclosure limitation', *Journal of Official Statistics*, vol. 9, no. 2, pp. 461-468, 1993.

[27] F. Sebé, J. Domingo-Ferrer, J. Mateo-Sanz and V. Torra, 'Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets' in J. Domingo-Ferrer (ed.) *Inference Control in Statistical Databases*, Berlin: Springer-Verlag, 2002. Lecture Notes in Computer Science 2316.

[28] L. Sweeney, 'Information explosion', in *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, eds. P. Doyle, J. I. Lane, J. M. Theeuwes and L. M. Zayatz, Elsevier, 43–74, 2001.

[29] V. Torra, Towards the re-identification of individuals in data files with Non-common variables, Proc. of the 14th European Conference on Artificial Intelligence (ECAI2000) (IOS Press, ISBN 1 58603 013 2), 326-330, Berlin, Germany, 2000.

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

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

[32] W. E. Winkler, 'Matching and record linkage', in B. G. Cox (Ed.), *Business Survey Methods*, New York: Wiley, pp. 355-384, 1995.

[33] W. E. Winkler, The State of Record Linkage and Current Research Problems, U.S. Census Bureau, Research Report 99/04. 1999 Also at http://www.census.gov/srd/www/byname.html

[34] W. E. Winkler, 'Masking and re-identification methods for public-use microdata: overview and research problems', in J. Domingo-Ferrer and V. Torra (eds.) *Privacy in Statistical Databases*, Berlin: Springer-Verlag, Lecture Notes in Computer Science 3050 (2004) 231-246.