WP.7
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (i): Microdata

ON METHOD-SPECIFIC RECORD LINKAGE FOR RISK ASSESSMENT

Invited Paper

Prepared by Jordi Nin, Javier Herranz and Vicenç Torra, Artificial Intelligence Research Institute and Spanish National Research Council, Spain

On method-specific record linkage for risk assessment

Jordi Nin, Javier Herranz and Vicenç Torra

IIIA, Artificial Intelligence Research Institute
CSIC, Spanish National Research Council
Campus UAB s/n
08193 Bellaterra (Catalonia, Spain)
{jnin,jherranz,vtorra}@iiia.csic.es

Abstract. Nowadays, the need for privacy motivates the use of methods that permit us to protect a microdata file both minimizing the disclosure risk and preserving the statistical utility.

Nevertheless, research is usually focused on how data utility is preserved, and much less research effort is dedicated to the study of the tools that an intruder might use to compromise the privacy of the data or, in other words, to increase the disclosure risk.

Record linkage is a standard mechanism used to measure the disclosure risk of a microdata protection method. In this paper we present some improvements for the (standard) distance based record linkage. In particular, we test our improvements to evaluate the disclosure risk of rank swapping, which is higher than what was believed up to now. We will also present the results of the application of this approach to microaggregation.

1 Introduction

Nowadays, statistical agencies publish confidential microdata files in the Internet. This data can be accessible for a variety of users, as decision makers, politicians, researchers or general public. However, such publication has to fulfill laws and regulations to preserve the privacy of the respondents.

A good statistical practice is that the released data include a full description of the data as well as the anonymization criteria that has been applied. For instance, all available microdata files in the EUROSTAT web page [12] include a text description explaining all the anonymization criteria applied to the confidential data.

The main goal of data protection methods [1], is to minimize both the *disclosure risk* (DR) and the *information loss* (IL) of the protected released microdata. Disclosure risk measures the capacity of an intruder to obtain some sensitive information about the original dataset from the protected one, and information loss measures the reduction of the statistical utility of the protected microdata with respect to the original one.

Information loss is deeply studied in many works [2, 6, 14], and it is out of the scope of this paper. Although, we will use in our experiments the measures defined in [7] to compare several protection methods.

In this paper, we focus in the way of computing the disclosure risk. Many works [6, 20] use *record linkage* methods [18, 19] for this purpose. Such methods are widely used in the scenario where an intruder has a compete access to the protected data set, whereas he knows some records of the original data set obtained from other data sources (publicly available or not). The aim of the intruder is to use record linkage to link his original records with the corresponding protected records released by the statistical agency. Obviously, the more records are correctly linked, the more disclosure risk has the employed protection method. Some examples of standard record linkage methods are distance based ones and probabilistic ones.

As we have said before, a good practice for statistical agencies is to give a complete description of the anonymization process, therefore, the intruder has a valuable information about how protected data is obtained. For this reason, the common assumption that a real intruder will use a standard record linkage method is quite unrealistic.

Many protection methods like rank swapping [16] or univariate microaggregation [8], protect the data using only *local* information, so that information loss is kept low. For instance, rank swapping has a parameter which limits the swap interval, or univariate microaggregation build the clusters with the k nearest values when the original data is sorted.

In this paper, we present an ad-hoc record linkage method called *Location Record Linkage* (L-RL). Our method exploits such limitations (*i.e.* protection is made locally). Using this knowledge, the intruder can limit the records where the record linkage method is applied, decreasing in this way the probability of finding incorrect links. As a result there is an increase on the number of correct links, and, therefore, an increment in the disclosure risk of such protection methods.

The rest of the paper is organized as follows. In Section 2 we recall the three data protection methods (rank swapping, univariate and multivariate microaggregation) where we have tested the new ad-hoc record linkage technique. Then we explain in Section 3 the basic concepts related to data protection, disclosure risk, information loss and the standard definition of score. a description of the new record linkage method. In Section 4 we describe the Location Record Linkage (L-RL) technique, we define a new score which takes into account L-RL, and we test L-RL with the abovementioned protection methods. Finally, in Section 5, we draw some conclusions and present some future work.

2 A Review of Protection Methods

2.1 Rank Swapping

Rank swapping is a widely used microdata protection method, which was originally described [16] only for ordinal attributes. However, in the comparisons made in [7], rank swapping was ranked among the best microdata protection methods for numerical attributes.

Rank swapping with parameter p and with respect to an attribute $attr_j$ (i.e., the j-th column of the original dataset X) can be defined as follows: first, the records of X are sorted in increasing order of the values x_{ij} of the considered attribute $attr_j$. For simplicity, assume that the records are already sorted, that is $x_{ij} \leq x_{\ell j}$ for all $1 \leq i < \ell \leq n$. Then, each value x_{ij} is swapped with another value $x_{\ell j}$, randomly and uniformly chosen in the set of still unswapped values in the limited range $i < \ell \leq i + p$. Finally, the sorting step is undone. When rank swapping is applied to a dataset, the algorithm explained above is run for each attribute to be protected, in a sequential way.

2.2 Univariate Microaggregation

Another widely used microdata protection method is microaggregation. Given a data set of a attributes, microaggregation builds small clusters of at least k elements and replaces each original value by the centroid of the cluster to which the element belongs.

A few different approaches exist for microaggregation. The simplest one is when each attribute is protected independently. This corresponds to *univariate microaggregation*. At present, a few optimal univariate microaggregation algorithms have been developed. A good example is [13], where the authors implement an optimal univariate microaggregation using graph operations over a graph built from the confidential data.

2.3 MDAV Microaggregation

The MDAV (Maximum Distance to Average Vector) algorithm [15] is an heuristic algorithm for multivariate microaggregation. MDAV is an iterative algorithm; at each step, it computes first the average record of a set of records and then builds a cluster with the farthest k records of this average record. Then, in the same step, another cluster is built with the farthest k records from the centroid of the new built cluster. Then, all the records of these two clusters are removed, and the process is repeated until all original values are protected.

3 Disclosure Risk Scenario

The main objective of a protection method is to anonymize a data set. A data set can be viewed as a file containing a number of records, where each record contains a

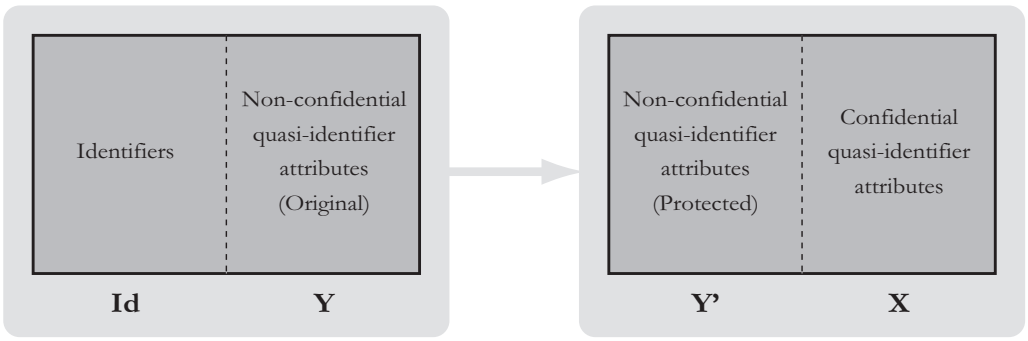

Figure 1: Re-identification scenario.

set of attributes describing an individual. The attributes in the original data set can be classified into two different categories, depending on their capability to identify individuals, as follows:

- **Identifiers.** They can be used to identify the individual unambiguously. A typical example of identifier is the passport number.
- **Quasi-identifiers.** They cannot identify a single individual when used alone. However, when they are combined with others quasi-identifiers attributes, they can uniquely identify an individual. Among the quasi-identifier attributes, we can distinguish between confidential and non-confidential, depending on whether they contain confidential information. An example of non-confidential quasi-identifier attribute is the postal code, while a confidential quasi-identifier is the salary.

When a data set is protected, identifiers are removed or encrypted to prevent an intruder to re-identify individuals easily. Typically, the remaining attributes are released, some of them protected. In this paper, we assume that non-confidential attributes are protected, while confidential attributes are not. This allows third parties to have precise information on confidential data without revealing to whom that confidential data belongs to.

In this scenario, as shown in Figure 1, an intruder might try to re-identify individuals by obtaining the non-confidential quasi-identifiers (Y) together with the identifiers (Id) from other data sources. Then, applying record linkage between the protected attributes (Y') and the same attributes obtained from other data sources (Y), the intruder might be able to re-identify a percentage of the protected individuals together with their confidential data (X). This is what protection methods try to prevent.

In general, the avoidance of all risk is not possible as this usually implies no information. Instead, we have to find a good trade off between information loss and disclosure risk. The score, presented in [6], was defined to measure this trade off in terms of information loss and disclosure risk measures. We define such measures below.

We will use these measures in our experiments as defined in [7].

- **Information Loss (IL).** Let X and X' be matrices representing the original and the protected data set, respectively. Let V and R be the covariance matrix and the correlation matrix of X, respectively; let $\overline{X}$ be the vector of variable averages for X and let S be the diagonal of V. Define V', R',$\overline{X}'$, and S' analogously from X'. The information loss is computed by averaging the mean variations of $X-X'$,$V-V'$,$S-S'$, and the mean absolute error of $R-R'$ and multiplying the resulting average by 100.

- **Disclosure Risk (DR).** The three different methods were presented in [18] to evaluate this risk: (i) *Distance Linkage Disclosure risk* (DLD), which is the average percentage of linked records using distance based record linkage, (ii) *Probabilistic Linkage Disclosure risk* (PLD), which is the average percentage of linked records using probabilistic based record linkage and (iii) *Interval Disclosure risk* (ID) which is the average percentage of original values falling into the intervals around their corresponding masked values. The three values are computed over the number of attributes that the intruder is assumed to know that, in our case, ranges from one to half of the attributes. The DR is a weighted mean that gives half weights to ID and the other half to linkage disclosure risk. That is:

$$DR = 0.5\,ID + 0.5\,[\frac{DLD + PLD}{2}]$$

- **Score:** The final score is defined as the arithmetic sum of IL and DR, therefore

$$score = 0.5\,IL + 0.5\,DR$$

4 Location Record Linkage

As we stated in Section 1, standard record linkage methods underestimate the real disclosure risk in the real world. Here, we consider a new protection method to be used when the intruder knows that only a subset of the protected records are eligible for being linked with the original one. We will call this method *location record linkage* (L-RL for short).

The rationale of our approach is intuitive: protection methods perturbate the original values in a controlled and predictive way to keep information loss as low as possible. For instance, for a given attribute, standard rank swapping swaps one original value with one of the p following values in the sorted table. Then, if the intruder knows all protected attributes (this is our case), he only needs to compare the original record x_i that he wants to link with $2p$ records in the protected data set (note that a protected value can be either the source or the destination in the swap process). The same problem happens with univariate microaggregation, where, if

original data is sorted, clusters are non-overlaped and the values of each cluster are contiguous.

Obviously, if more than one attribute are known, it is possible to repeat the process for each attribute. Formally, if the original record $x_i = (x_{i1}, \ldots, x_{ic})$ has c attributes $attr_1, \ldots, attr_c$, then, the matching protected record x'_ℓ will necessarily satisfy the condition

$$x'_\ell \in \cap_{1 \leq j \leq c} B(x_{ij}),$$

where $B(x_{ij}$ contains all the protected records whose j-th attribute is one of the $2p$ candidates to have been swapped with x_{ij}. That is, the search of the protected record is reduced to an intersection of the sets of possible protected records. Of course, the more attributes are considered, the less records will be in this intersection, and, therefore, the probability of finding the correct record linkage will increase. However, this is not the main concern, because for some combination of the protected attributes, the intersection gives a unique record: the intruder can be sure that this is the protected record which matches with the considered original record. This is so, because this linkage method does not introduce error probabilities. So, the method guarantees to the intruder that the link is correct.

4.1 Experiments

We have considered two different data sets in our experiments. The first one, called Census, has been extracted using the Data Extraction System (DES) from the U. S. Census Bureau [5]. This dataset contains 1080 records consisting of 13 attributes. The second one, called EIA, was extracted from the U.S. Energy Information Authority [11]. It contains 4092 records consisting of 10 attributes.

As we are interested in studying the effects of the L-RL, we have computed two different indicators:

Number of linkages. We study the number of correct links that L-RL is able to find using different sets of attributes. We will assume that the intruder knows all the protected records and he has partial knowledge of the attributes.

Score computation. We compute the standard score and a new variant of it which takes into account the L-RL method. Formally, we define these two score as

- **Score$_1$.** The standard score is computed as presented in Section 3. That is

$$score = 0.5\, IL + 0.125\, DLD + 0.125\, PLD + 0.25\, ID$$

- **Score$_2$.** Our variant of the score is defined by the following expression that includes the standard measures as well as the new L-RL (we use

	rs 2	rs 4	rs 6	rs 8	rs 10	rs 12	rs 14	rs 16	rs 18	rs 20
1	38.4	18.0	16.8	10.8	10.8	6.4	6.8	5.2	4.0	4.0
2	497.0	130.2	54.2	29.2	21.8	15.4	13.0	10.4	7.0	6.0
3	1034.2	761.2	420.6	197.2	99.0	60.2	45.2	32.4	28.8	24.2
4	1071.8	959.6	694.2	378.6	199.4	107.2	71.6	58.0	49.4	39.6
5	1076.8	1042.0	925.2	711.6	463.2	281.4	195.0	165.6	131.6	121.2
6	1079 .0	1063.2	1001.6	879.2	681.0	484.2	413.0	340.4	293.6	287.4
7	1079.2	1064.2	1018.6	913.8	733.0	547.4	475.4	432.4	408.6	339.2
8	1079.2	1077.6	1071.8	1042.6	972.0	861.6	701.6	528.6	472.4	386.4
9	1079.6	1077.6	1071.4	1065.6	1036.6	988.8	888.0	766.6	602.0	466.0
10	1079.6	1078.2	1072.1	1066.6	1039.2	996.6	930.8	824.8	677.4	544.0
11	1079.6	1079.1	1073.4	1069.2	1039.4	1001.0	939.4	846.8	706.2	574.0
12	1079.6	1079.1	1073.4	1069.6	1041.2	1002.0	942.0	853.4	726.4	593.8
13	1079.6	1079.1	1076.7	1070.3	1044.8	1004.4	944.2	871.0	745.6	615.4

Table 1: Number of correctly linked records when L-RL is applied to Census data set, protected with rank swapping. The first column shows the number of known attributes.

LLD to denote Location Linkage Disclosure risk) presented in Section 4. That is,

$$score = 0.5\,IL + 0.25\,(\frac{DLD + PLD + LLD}{3}) + 0.25\,ID$$

4.1.1 Rank Swapping

In Tables 1 and 2 we can observe detailed results about the number of correct links obtained by L-RL using different sets of attributes on data protected using rank swapping. It is easy to observe that the more attributes are known by the intruder, the more records are linked. Note that, for the five less protected data sets from Census, an intruder links more than 70% of the records when only half of the attributes are known. Another interesting result with the Census data set is that the intruder is always able to link more than 50% of the records if he knows all the attributes. Similar results are obtained for the EIA data set. For the three less protected datasets, the intruder is able to link more than 50% of records when all the attributes are known.

Tables 3 (Census data set) and 4 (EIA data set) present $score_1$ and $score_2$, as well as the original values of their components before their aggregation. We can observe that the largest disclosure risk measure, for all cases, is LLD. Therefore, it is clear that the L-RL method increases the risk with respect to standard ones for rank swapping.

4.1.2 Univariate Microaggregation

In Table 5 we can observe detailed results about the number of correct links obtained by L-RL using different sets of attributes for data protected using univariate microaggregation. The results show clearly that the intruder is able to link almost

	rs 2	rs 4	rs 6	rs 8	rs 10	rs 12	rs 14	rs 16	rs 18	rs 20
1	70.3	46.1	43.6	42.4	40.0	37.6	37.6	35.1	37.5	36.3
2	378.4	183.4	145.1	139.8	135.7	135.4	134.6	133.2	133.0	132.3
3	2174.8	338.8	284.1	246.8	236.1	229.1	226.1	224.7	224.3	223.5
4	2827.1	557.0	380.5	327.6	310.2	301.7	298.3	296.0	294.7	294.9
5	3402.9	1441.3	720.9	496.1	423.7	398.7	384.1	373.0	374.0	367.7
6	3582.9	1859.4	856.3	512.4	400.3	415.9	397.0	380.2	378.8	371.3
7	3699.8	2420.6	1325.3	709.6	431.4	423.9	410.1	393.1	424.5	391.6
8	3778.6	2699.0	1631.7	947.3	572.8	448.8	458.3	411.5	445.6	401.3
9	3810.1	2862.2	1808.5	1081.8	654.6	492.2	479.4	420.9	451.9	409.3
10	3831.5	2996.8	1986.3	1221.5	741.5	539.2	507.9	432.6	455.4	411.8

Table 2: Number of correctly linked records when L-RL is applied to EIA dataset, protected with standard rank swapping. The first column shows the number of known attributes.

all the records using only a few attributes. This happens for both data sets. It is clear that univariate microaggregation has a high disclosure risk, greater than the one of rank swapping.

Tables 7 and 8 present, in the same way than in the rank swapping example, $score_1$ and $score_2$, as well as, their components. We can observe that the largest disclosure risk measure is LLD in all cases. Therefore, it is clear that the L-RL method increases the risk with respect to standard ones for univariate microaggregation.

4.1.3 Multivariate Microaggregation

Table 9 presents the number of correct links obtained by L-RL using different sets of attributes for data protected using MDAV multivariate microaggregation. As we can observe in the table, the more groups of attributes are known, the less records are linked. This is due that to the fact that MDAV does not present the same locality problem than univariate microaggregation and rank swapping. In other words, not all the original records are assigned to the cluster represented by the nearest centroid. *I.e.*, some records might be assigned to the second or third nearest cluster. Therefore, L-RL is unsuitable for MDAV. This effect is also illustrated in Tables 10 and 11, where LLD is the lowest disclosure risk value, therefore $score_2$ is lower than $score_1$ and LLD should not be used for the evaluation of the disclosure risk for MDAV.

5 Conclusions

In this paper, we have presented a new type of record linkage designed to exploit the limitations of some protection methods. We have shown that this new method obtains a more accurate disclosure risk evaluation for rank swapping and univariate microaggregation.

	IL	LLD	DLD	PLD	ID	Score_1	Score_2
rs 2	3.89	77.73	73.52	71.28	93.98	43.54	43.98
rs 4	6.54	66.65	58.40	42.92	83.09	36.71	38.04
rs 6	10.57	54.65	43.76	22.49	72.12	31.60	33.39
rs 8	16.54	41.28	32.13	11.74	62.11	29.28	30.89
rs 10	20.18	29.21	23.64	6.03	53.28	27.12	28.32
rs 12	23.46	19.87	18.96	3.46	47.17	26.33	27.05
rs 14	28.93	16.14	15.63	2.06	43.39	27.52	28.13
rs 16	35.16	13.81	13.59	1.29	40.78	29.64	30.17
rs 18	32.52	12.21	11.50	0.83	38.90	27.53	28.03
rs 20	35.12	10.88	10.87	0.59	37.33	28.33	28.75

Table 3: Score calculation for rank swapping using the Census data set. IL stands for Information Loss, LLD stands for Location Linkage Disclosure, DLD stands for Distance Linkage Disclosure, PLD stands for Probability Linkage Disclosure, ID stands for Interval Disclosure, Score_1 is the score computed only using DLD and PLD and Score_2 is the score computed taking into account LLD results.

	IL	LLD	DLD	PLD	ID	Score_1	Score_2
rs 2	4.24	43.27	21.71	16.85	93.10	30.22	32.21
rs 4	9.67	12.54	10.61	4.79	82.09	27.28	27.69
rs 6	14.63	7.69	7.40	2.03	72.21	26.55	26.79
rs 8	18.71	6.12	5.98	1.12	63.90	26.22	26.43
rs 10	22.87	5.60	5.19	0.69	57.09	26.44	26.66
rs 12	26.60	5.39	4.87	0.51	51.64	26.88	27.11
rs 14	29.42	5.28	4.55	0.32	47.49	27.19	27.43
rs 16	32.38	5.19	4.54	0.23	44.19	27.83	28.07
rs 18	34.22	5.20	4.54	0.22	41.42	28.06	28.30
rs 20	36.27	5.15	4.36	0.18	38.97	28.45	28.69

Table 4: Score calculation for rank swapping using the EIA data set.

We have also presented some experiments using MDAV microaggregation that prove that in some sense MDAV is immune to the location problem described in this paper.

As future work, we plan to study the disclosure risk of MDAV and other protection methods using other ad-hoc, specific, record linkage methods.

Acknowledgements

Partial support by the Spanish MEC (projects ARES – CONSOLIDER INGENIO 2010 CSD2007-00004 – and eAEGIS – TSI2007-65406-C03-02) is acknowledged. Jordi Nin wants to thank the Spanish Council for Scientific Research (CSIC) for his I3P grant.

	Census				EIA			
k	2 Vars	3 Vars	4 Vars	5 Vars	2 Vars	3 Vars	4 Vars	5 Vars
10	1032	1079	1080	1080	3430	3923	3947	4035
20	892	1070	1077	1079	2609	3780	3872	3980
30	704	1054	1072	1078	1931	3599	3751	3900
40	531	1021	1065	1076	1388	3347	3621	3806
50	379	989	1054	1069	1012	3074	3427	3703

Table 5: Number of correctly linked records when L-RL is applied to Census data set, protected with univariate microaggregation.

Table 6: Score optimal univariate microaggregation using the Census data set

k	IL	LLD	DLD	PLD	ID	Score_1	Score_2
10	1.15	98.87	86.28	86.31	98.36	46.74	47.79
20	2.85	95.32	83.43	83.47	93.43	45.64	46.63
30	3.71	90.46	80.36	80.21	88.41	44.03	44.88
40	4.71	85.49	77.00	76.57	83.69	42.47	43.20
50	5.66	80.81	73.94	73.56	79.41	41.12	41.71

Table 7: Score calculation for optimal univariate microaggregation using the Census data set.

References

[1] Adam, N. R., Wortmann, J. C.,(1989), Security-control for statistical databases: a comparative study, ACM Computing Surveys, Volume: 21, 515-556.

[2] Bertino, E., Nai, I., Parasiliti, L., (2005), A framework for evaluating privacy preserving data mining algorithms, DMKD, Springer, 11:2 121-154.

[3] Brand, R., Domingo-Ferrer, J., and Mateo-Sanz, J. M., (2002) Reference datasets to test and compare sdc methods for protection of numerical microdata. Manuscript for [4].

[4] CASC: Computational Aspects of Statistical Confidentiality, European Project IST-2000-25069, http://neon.vb.cbs.nl/casc.

[5] Data Extraction System, U.S. Census Bureau, http://www.census.gov/

[6] Domingo-Ferrer, J., Torra, V., (2001), Disclosure control methods and information loss for microdata, Pages 91-110 of [10].

[7] Domingo-Ferrer, J., Torra, V., (2001), A quantitative comparison of disclosure control methods for microdata, Pages 111-133 of [10].

[8] Domingo-Ferrer, J., Mateo-Sanz, J. M. (2002) Practical data-oriented microaggregation for statistical disclosure control, KDE, 14 189-201.

	IL	LLD	DLD	PLD	ID	Score_1	Score_2
10	0.32	93.69	72.53	76.74	99.69	43.74	45.33
20	0.80	87.01	55.94	70.32	99.54	41.07	43.06
30	1.42	80.53	43.54	64.93	99.35	39.10	41.30
40	1.62	74.30	35.52	59.78	98.75	37.41	39.63
50	2.07	68.52	30.41	55.08	95.26	35.53	37.68

Table 8: Score calculation for optimal univariate microaggregation using the EIA data set.

	Census					EIA			
	2 GV	3 GV	4 GV	5 GV	6 GV	2 GV	3 GV	4 GV	5 GV
Mic2-05	120	133	94	76	69	1493	1616	1208	922
Mic2-15	42	178	197	201	200	593	1626	1178	1005
Mic2-25	23	91	130	177	205	349	1306	1551	1324
Mic3-05	73	108	159			414	311		
Mic3-15	28	74	199			170	441		
Mic3-25	12	20	99			82	294		
Mic4-05	45	138				449			
Mic4-15	4	44				141			
Mic4-25	2	12				69			

Table 9: Number of correctly linked records when L-RL is applied to MDAV microaggregation. GV stands for the number of groups of variables known by the intruder. Mic$i$-k corresponds to MDAV microaggregation using v variables and clusters of size k.

[9] Domingo-Ferrer, J., Torra, V. (2005) Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation, DMKD, 11 195-212.

[10] Doyle, P., Lane, J., Theeuwes, J., Zayatz, L., eds. (2001), Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies, Elsevier Science.

[11] U.S. Energy Information Authority, http://www.eia.doe.gov/cneaf/electricity /page/eia826.html

[12] Statistical Office of the European Communities (EUROSTAT), http://epp.eurostat.ec.europa.eu

[13] Hansen, S., Mukherjee, S. (2003) A Polynomial Algorithm for Optimal Univariate Microaggregation. KDE, 15:4 1043-1044.

[14] Mateo-Sanz, J.M., Domingo-Ferrer, J., Sebé, F., (2005), Probabilistic information loss measures in confidentiality protection of continuous microdata, DMKD, Springer, 11:2, 181-193.

[15] Hundepool, A., Van de Wetering, A., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P.-P., Domingo-Ferrer, J., Torra, V., Brand, R., Giessing,

	IL	LLD	DLD	PLD	ID	Score_1	Score_2
Mic2-05	19.30	9.11	69.06	49.22	74.77	43.13	38.96
Mic2-15	37.70	15.15	45.83	26.67	60.94	43.15	41.39
Mic2-25	47.16	11.59	28.56	16.81	51.93	42.23	41.31
Mic3-05	30.66	10.49	37.44	33.58	65.21	40.51	38.43
Mic3-15	42.76	9.29	22.75	19.38	54.79	40.34	39.36
Mic3-25	56.13	4.04	15.86	13.36	51.57	44.61	43.73
Mic4-05	34.67	8.47	31.90	24.35	61.37	39.71	38.07
Mic4-15	45.58	2.22	15.97	12.31	52.43	39.43	38.44
Mic4-25	54.60	0.65	11.20	7.08	45.09	40.86	40.15

Table 10: Score calculation for optimal MDAV multivariate microaggregation using the Census data set.

	IL	LLD	DLD	PLD	ID	Score_1	Score_2
Mic2-05	2.99	32.01	35.01	50.80	93.71	35.65	34.74
Mic2-15	5.49	26.89	20.02	31.49	86.50	30.81	30.90
Mic2-25	6.35	27.68	16.09	26.89	83.88	29.52	30.03
Mic3-05	7.64	8.86	21.47	34.53	85.52	32.20	30.60
Mic3-15	9.99	7.47	11.33	22.67	79.63	29.15	28.36
Mic3-25	11.12	4.59	9.60	18.32	77.63	28.46	27.68
Mic4-05	8.30	10.97	25.71	36.78	87.76	33.90	32.21
Mic4-15	19.16	3.45	12.66	21.31	81.57	34.22	33.09
Mic4-25	20.11	1.69	8.11	14.66	78.28	32.47	31.66

Table 11: Score calculation for optimal MDAV multivariate microaggregation using the EIA data set.

S. (2003) μ-ARGUS version 3.2 Software and User's Manual. Statistics Netherlands, Voorburg NL, feb 2003.

[16] Moore, R., (1996), Controlled data swapping techniques for masking public use microdata sets, U. S. Bureau of the Census (Unpublished manuscript).

[17] Nin, J., Herranz, J., Torra, V., Rethinking Rank Swapping to Decrease Disclosure Risk, Data and Knowledge Engineering, in press. http://dx.doi.org/10.1016/j.datak.2007.07.006

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

[19] Winkler, W. E., (2003), Data cleaning methods, Proc. SIGKDD 2003.

[20] Yancey, W. E., Winkler, W. E., Creecy, R. H., (2002), Disclosure risk assessment in perturbative microdata protection, Inference Control in Statistical Databases: From Theory to Practice, LNCS, Springer, 2316, 135-152.