Post-Masking Optimization of the Tradeoff between Information Loss and Disclosure Risk in Masked Microdata Sets*

Francesc Sebé[1], Josep Domingo-Ferrer[1],
Josep Maria Mateo-Sanz[2], and Vicenç Torra[3]

[1] Universitat Rovira i Virgili, Dept. of Computer Science and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain
{fsebe,jdomingo}@etse.urv.es
[2] Universitat Rovira i Virgili, Statistics and OR Group
Av. Països Catalans 26, E-43007 Tarragona, Catalonia, Spain
jmateo@etseq.urv.es
[3] Institut d'Investigació en Intel·ligència Artificial
Campus de Bellaterra, E-08193 Bellaterra, Catalonia, Spain
vtorra@iiia.csic.es

Abstract. Previous work by these authors has been directed to measuring the performance of microdata masking methods in terms of information loss and disclosure risk. Based on the proposed metrics, we show here how to improve the performance of any particular masking method. In particular, post-masking optimization is 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 can also be used for synthetic microdata generation and can be extended to preservation of all moments up to m-th order, for any m.
Keywords: Microdata masking, Information loss, Disclosure risk, Post-masking optimization, Synthetic microdata generation.

1 Introduction

Statistical offices must guarantee statistical confidentiality when releasing data for public use. Statistical Disclosure Control (SDC) methods are used to to that end[7]. If data being released consist of individual respondent records, called microdata in the official statistics jargon, confidentiality translates to avoiding disclosure of the identity of the individual respondent associated with a published record. At the same time, SDC should preserve the informational content to the maximum extent possible. SDC methods are an intermediate option between encryption of the original data set (no disclosure risk but no informational content released) and straightforward release of the original data set (no confidentiality

* Work partly supported by the European Commission under project IST-2000-25069 "CASC".

J. Domingo-Ferrer (Ed.): Inference Control in Statistical Databases, LNCS 2316, pp. 163–171, 2002.

but maximal informational content released). SDC methods for microdata are also known as masking methods.

In [2,3], a comparison of masking methods for microdata was conducted. The comparison was done by computing a score which weighed information loss against disclosure risk between the original and the masked microdata set. In [1], the score was modified to cope with the case where the number of records in the original and the masked data sets is not the same.

In this paper, we present a post-masking optimization procedure which seeks to modify the masked data set to preserve the first and second-order moments of the original data set as much as possible without increasing the disclosure risk. The better the first and second-order moments are preserved, the better will multivariate statistics on the resulting masked data set mimic those that would be obtained on the original data set. In order to avoid substantially increasing the disclosure risk, a constraint is imposed to guarantee that individual data resulting from post-masking optimization are not too similar to individual original data.

The optimization procedure presented can be combined with any masking method for numerical microdata and can lead to an improvement in terms of the score defined in [3]: the reason is that the information loss is reduced without significantly increasing the disclosure risk. The procedure can also be used in a stand-alone way to produce synthetic data sets with prescribed first and second-order moments. Furthermore, an extension aiming at preservation of all moments up to m-th order for any m is straightforward.

Section 2 sketches the score constructed in [3] with the modifications of [1]. Section 3 describes the post-masking optimization problem and a heuristic hill-climbing procedure to solve it. Computational results reflecting improvement of rankswapped data and microaggregated data are presented in Section 4. Section 5 presents some conclusions and sketches two extensions: 1) how to use the proposed optimization as a synthetic microdata generator; 2) how to preserve all moments up to m-th order for any m.

2 A Score for Method Comparison

Let n be the number of records in the original data set and n' the number of records in the masked data set. Let d be the number of variables (assumed to be the same in both data sets). Let X be an $n \times d$ matrix representing the original data set: columns correspond to variables and rows correspond to records. Similarly, let X' be an $n' \times d$ matrix representing the masked data set. Let V and V' be the covariance matrices of X and X', respectively; similarly, let R and R' be the correlation matrices. Let $\bar{X}$ and $\bar{X}'$ be the vectors of averages of variables in X and X'. Finally, let S and S' be the vectors of variances of variables in X and X'. Define the mean absolute error of a matrix A' vs another matrix A as the average of the absolute values of differences of corresponding components in both matrices (what "corresponding" means will be discussed below); define the mean variation as the average of absolute differences of corresponding compo-

nents in both matrices with each difference divided by the absolute value of the component in A.

Disclosure risk can be measured using record linkage. Two record linkage methods were used in [3]:

Distance-based record linkage. Let the original and masked data sets consist both of d variables (it is assumed that both data sets contain the same variables). We define that a record in the masked data set corresponds to the nearest record in the original data set, where "nearest" means at shortest d-dimensional Euclidean distance. Assume further that the intruder can only access i key variables of the original data set and tries to link original and masked record based on these i variables. Linkage then proceeds by computing i-dimensional distances between records in the original and the masked data sets (distances are computed using only the i key variables) The variables used are standardized to avoid scaling problems. A record in the masked data set is labeled as "correctly linked" when the nearest record using i-dimensional distance is the corresponding one (*i.e.* the nearest record using d-dimensional distance).

Probabilistic record linkage. Defined in [5], uses a matching algorithm to pair records in the masked and original data sets. The matching algorithm is based on the linear sum assignment model. The definition of "correctly linked" records is the same as in distance-based record linkage. This method is attractive because it only requires the user to provide two probabilities as input: one is an upper bound of the probability of a false match and the other an upper bound of the probability of a false non-match. Unlike distance-based record linkage, probabilistic record linkage does not require rescaling variables nor makes any assumption on their relative weight (by default, distance-based record linkage assumes that all variables have the same weight). Furthermore, probabilistic record linkage can be used on both numerical and categorical data.

A score combining information loss measures with disclosure risk measures can be constructed as follows:

IL Information Loss: 100 times the average of the mean variation of X' vs X (called IL_1), the mean variation of $\bar{X}'$ vs $\bar{X}$ (called IL_2), the mean variation of V' vs V (called IL_3), the mean variation of S' vs S (called IL_4) and the mean absolute error of R' vs R (called IL_5).

DLD Average of DLD-1, $\cdots$, DLD-7. DLD-i is the percent of records correctly linked using distance-based record linkage with Euclidean distance when the intruder knows i key variables of the original file. See below for a discussion on what "correctly linked" means.

PLD Same as DLD, but for probabilistic record linkage.

ID Average percent of original values falling in the intervals around their corresponding masked values. The average is over interval widths from 1% to 10%.

Overall score

$$Overall\_score = 0.5 \cdot IL + 0.125 \cdot DLD + 0.125 \cdot PLD + 0.25 \cdot ID \quad (1)$$

A simplified version of the score can be used with only one record linkage method, namely DLD. In this case, the score is

$$Score = 0.5 \cdot IL + 0.25 \cdot DLD + 0.25 \cdot ID \quad (2)$$

The lower $Score$, the better is a method. In computation of IL_2, IL_3, IL_4 and IL_5 "corresponding" components between the matrices for the original and masked data sets means "referring to the same variables". To compute IL_1 and ID, we need to define a correspondence between records in X and X'. A natural way is to map each published masked record i to the nearest original record $c(i)$, using the d-dimensional Euclidean distance between records (where d is the number of variables in the data sets).

Also, computation of the disclosure risk measure DLD (and PLD) requires defining what correct linkage means. If the masked and the original data sets have d variables, we say that a masked record is correctly linked to an original record if they are at the shortest possible d-dimensional Euclidean distance.

3 Post-Masking Optimization

Once an original data set X has been masked as X', post-masking optimization aims at modifying X' into X'' so that the first and second-order moments of X are preserved as much as possible by X'' while keeping IL_1 around a prescribed value. Near preservation of first and second-order moments results in (constrained) minimization of IL_2, IL_3, IL_4 and IL_5. Regarding IL_1, we cannot pretend to minimize it, because disclosure risk would most likely suffer a dramatic increase: post-masking optimized data would look too much like the original data.

3.1 The Model

The first-order moments of X depend on the sums

$$\frac{\sum_{i=1}^{n} x_{ij}}{n} \text{ for } j = 1, \cdots, d \quad (3)$$

where x_{ij} is the value taken by the j-th variable for the i-th record. The second-order moments of X depend on the sums

$$\frac{\sum_{i=1}^{n} x_{ij}^2}{n} \text{ for } j = 1, \cdots, d \quad (4)$$

$$\frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{n} \text{ for } j, k = 1, \cdots, d \text{ and } j < k \quad (5)$$

Therefore, our goal is to modify X' to obtain a X'' so that the above $2d + d(d-1)/2$ sums are nearly preserved while keeping IL_1 and disclosure risk similar in X' and X''. First, let us compute IL_1 of X' vs X as

$$IL_1 := 100\left(\frac{\sum_{i=1}^{n'}\sum_{j=1}^{d}\frac{|x'_{ij}-x_{c(i),j}|}{|x_{c(i),j}|}}{dn'}\right) \tag{6}$$

where $c(i)$ is the original record nearest to the i-th masked record of X' (d-dimensional Euclidean distance is used). Now let $0 < q \leq 1$ be a parameter and let M be the set formed by the $100q\%$ records of X' contributing most to IL_1 above. Then let us compute the values x''_{ij} of X'' as follows. For $x'_{ij} \notin M$ then $x''_{ij} := x'_{ij}$. For $x'_{ij} \in M$ the corresponding x''_{ij} are solutions to the following minimization problem:

$$\min_{\{x''_{ij}|x'_{ij}\in M\}} \sum_{j=1}^{d}\left(\frac{\sum_{i=1}^{n'} x''_{ij}}{n'} - \frac{\sum_{i=1}^{n} x_{ij}}{n}\right)^2 + \sum_{j=1}^{d}\left(\frac{\sum_{i=1}^{n'} x''^2_{ij}}{n'} - \frac{\sum_{i=1}^{n} x^2_{ij}}{n}\right)^2$$

$$+ \sum_{1\leq j<k\leq d}\left(\frac{\sum_{i=1}^{n'} x''_{ij}x''_{ik}}{n'} - \frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{n}\right)^2 \tag{7}$$

subject to

$$0.99 \cdot p \cdot IL_1 \leq \frac{\sum_{j=1}^{d}\sum_{i=1}^{n'}\frac{|x''_{ij}-x_{C(i),j}|}{|x_{C(i),j}|}}{dn'} \leq 1.01 \cdot p \cdot IL_1 \tag{8}$$

where $p > 0$ is a parameter and $C(i)$ is the original record nearest to the i-th masked record of X'' after optimization. Note that, in general, $C(i) \neq c(i)$, because in general $X'' \neq X'$.

3.2 A Heuristic Optimization Procedure

To solve the minimization problem (7) subject to constraint (8), the following hill-climbing heuristic procedure has been devised:

Algorithm 1 (PostMaskOptim(X,X',p,q,TargetE)).

1. *Standardize all variables in X and X' by using for both data sets the averages and standard deviations of variables in X.*
2. *Compute IL_1 between X and X' according to expression (6).*
3. *Let $TargetIL_1 := p \cdot IL_1$.*
4. *Let $X'' := X'$.*
5. *Rank records in X'' according to their contribution to IL_1. Let M be the subset of the $100q\%$ records in X'' contributing most to IL_1.*
6. *For each record i in X'', determine its nearest record $C(i)$ in X (use d-dimensional Euclidean distance).*

7. *Compute E, where E denotes the objective function in Expression (7).*
8. *While $E \geq TargetE$*
   (a) *Randomly select one value v of a record i_v in $M \subset X''$ and randomly perturb it to get v'. Replace v with v' in record i_v.*
   (b) *Recompute the nearest record $C(i_v)$ in X nearest to the updated i_v.*
   (c) *Let $PreviousIL_1 := IL_1$.*
   (d) *Compute IL_1 between X and X''. To do this, use expression (6) while replacing x'_{ij} by x''_{ij} and $c(i)$ by $C(i)$.*
   (e) *Let $PreviousE := E$.*
   (f) *Recompute E (X'' has been modified).*
   (g) *If $E \geq previousE$ then $undo := true$.*
   (h) *If $IL_1 \notin [0.99 \cdot TargetIL_1, 1.01 \cdot TargetIL_1]$ and $|IL_1 - TargetIL_1| \geq |PreviousIL_1 - TargetIL_1|$ then $undo := true$.*
   (i) *If $undo = true$ then restore the original value v of record i_v and recompute the nearest record $C(i_v)$ in X nearest to i_v.*
9. *Destandardize all variables in X and X'' by using the same averages and standard deviations used in Step 1.*

Note that, by minimizing E, the algorithm above attempts to minimize the information loss IL. No direct action is taken to reduce or control disclosure risk measures DLD and ID, beyond forcing that IL_1 should be in a pre-specified interval to prevent the optimized data set from being dangerously close to the original one. The performance of Algorithm 1 is evaluated *a posteriori*: once E reaches $TargetE$, the algorithm stops and yields an optimized data set for which IL, DLD and ID must be measured.

4 Computational Results

The test microdata set no. 1 of [1] was used. This microdata set was constructed using the Data Extraction System (DES) of the U.S. Census Bureau (`http://www.census.gov/DES`). $d = 13$ continuous variables were chosen and 1080 records were selected so that there were not many repeated values for any of the attributes (in principle, one would not expect repeated values for a continuous attribute, but there were repetitions in the data set).

In the comparison of [2,3], two masking methods were singled out as particularly well-performing to protect numerical microdata: rank swapping [6] and multivariate microaggregation [4]. For both methods, the number of masked records is the same as the number of original records ($n = n' = 1080$). Several experiments have been conducted to demonstrate the usefulness of post-masking optimization to improve on the best (lowest) scores reached by rank swapping and multivariate microaggregation.

The first row of Table 1 shows the lowest score reached by rank swapping for the test microdata set: the score is 25.66 and is reached for parameter value 14 (see [1]). The next rows of the table show scores reached when Algorithm 1 is used with several different values of parameters p (proportion between target IL_1 and initial IL_1) and q (proportion of records in M). The last column shows

the value of the objective function E reached (for all rows but the first one, this is the $TargetE$ parameter of Algorithm 1). The score is computed using Expression (2) and the values of IL, DLD and ID reached are also given in Table 1.

Table 1. Rank-swapping with parameter 14. First row, best score without optimization; next rows, scores after optimization

p	q	$Score$	IL	DLD	ID	E
None	None	25.66	23.83	14.74	40.23	0.419
0.5	0.5	24.45	14.73	20.30	48.03	0.04
0.5	0.3	22.15	13.65	16.30	44.98	0.04
0.5	0.1	21.71	15.26	14.81	41.51	0.09

The first row of Table 2 shows the lowest score reached by multivariate microaggregation for the test data set: the score is 31.86 and is reached for parameter values 4 and 10, that is, when four variables are microaggregated at a time and a minimal group size of 10 is considered (see [1]). The next rows of the table show scores reached when Algorithm 1 is used with several different values of parameters p and q.

Table 2. Multivariate microaggregation with parameters 4 and 10. First row, best score without optimization; next rows, scores after optimization

p	q	$Score$	IL	DLD	ID	E
None	None	31.86	22.48	22.14	60.34	0.122
0.5	0.5	26.96	14.16	21.06	58.54	0.008
0.5	0.3	27.39	14.74	21.29	58.80	0.008
0.5	0.1	28.03	14.94	21.83	60.38	0.008

When looking at the results on rankswapped data (Table 1), we can observe the following:

- There is substantial improvement of the score: 21.71 for post-masking optimization with $p = 0.5$ and $q = 0.1$ in front of 25.66 for the initial rankswapped data set.
- The lower q (*i.e.* the smaller the number of records altered by post-masking optimization), the better is the score. In fact, the score for $q = 0.1$ is lower than for $q = 0.3, 0.5$ even if the target E for $q = 0.1$ is less stringent (higher) than for the other values of q.
- Post-masking optimization improves the score by reducing information loss IL and hoping that disclosure risks DLD and ID will not grow. In fact,

Table 1 shows that DLD and ID increase in the optimized data set with respect to the rankswapped initial data set. The lower q, the lower is the impact on the rankswapped initial data set, which results in a smaller increase in the disclosure risk. This small increase in disclosure risk is dominated by the decrease in information loss, hence the improved score.

The results on microaggregated data (Table 2) are somewhat different. The following comments are in order:

- Like for rankswapping, there is substantial improvement of the score: 26.96 for post-masking optimization with $p = 0.5$ and $q = 0.5$ in front of 31.86 for the initial microaggregated data set.
- The higher q, the better is the score. This can be explained by looking at the variation of IL, DLD and ID. Microaggregated data are such that there is room for decreasing IL while keeping DLD and ID at the same level they had in the initial microaggregated data set. In this respect, we could interpret that, multivariate microaggregation being "less optimal" than rank swapping, we should not be afraid of changing a substantial number of values because this can still lead to improvement.

5 Conclusions and Extensions

The procedure presented here is designed to minimize information loss between a masked data set and its original version. Although disclosure risk is not explicitly considered in the minimization model described in Subsection 3.1, there is a constraint on IL_1 whose purpose is to prevent the optimized masked data set from being too close to the original one.

The described post-masking optimization can be applied to improve any masking method. We have demonstrated improvement in the case of two microdata masking methods which already had been identified as the best performers for numerical microdata protection. This is a substantial step forward in optimizing the tradeoff between information loss and disclosure risk in microdata protection.

The application of the proposed technique can be extended in at least two directions:

- *Synthetic microdata generation.* Algorithm 1 can be used on a stand-alone basis to generate synthetic microdata. To do this, let input parameter p be small, let $q := 1$ and let input parameter X' be a random data set with the same number of variables as the original data set X. The resulting synthetic data set will be such that the objective function E reaches the pre-specified value $TargetE$ and IL_1 is $p\%$ of the initial (big) IL_1.
- *Preservation of all moments up to m-th order.* Given a positive integer m, all what is needed to preserve all moments up to m-th order is to add to the objective function (7) terms corresponding to the squared differences between the i-th order sums of X and X'', for $i = 1$ to m.

References

1. R.A. Dandekar, J. Domingo-Ferrer, and F. Sebé, "LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection", in *Inference Control in Statistical Databases*, LNCS 2316, Springer 2002, pp. 153–162.
2. J. Domingo-Ferrer, J.M. Mateo-Sanz, and V. Torra, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk", *Proc. of ETK-NTTS 2001.* Luxemburg: Eurostat, pp. 807–825, 2001.
3. J. Domingo-Ferrer and V. Torra, "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure and Data Access*, eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz. Amsterdam: North-Holland, pp. 111-133, 2001.
4. J. Domingo-Ferrer and J.M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control", *IEEE Transactions on Knowledge and Data Engineering,* vol. 14, pp. 189–201, 2002.
5. M.A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*, vol. 84, pp. 414–420, 1989.
6. R. Moore, "Controlled data swapping techniques for masking public use microdata sets", U. S. Bureau of the Census, 1996 (unpublished manuscript).
7. L. Willenborg and T. de Waal, *Elements of Statistical Disclosure Control.* New York: Springer-Verlag, 2001.