Chapter 1

A SURVEY OF INFERENCE CONTROL METHODS FOR PRIVACY-PRESERVING DATA MINING

Josep Domingo-Ferrer*

Rovira i Virgili University of Tarragona †
UNESCO Chair in Data Privacy
Dept. of Computer Engineering and Mathematics
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
josep.domingo@urv.cat

Abstract Inference control in databases, also known as Statistical Disclosure Control (SDC), is about protecting data so they can be published without revealing confidential information that can be linked to specific individuals among those to which the data correspond. This is an important application in several areas, such as official statistics, health statistics, e-commerce (sharing of consumer data), etc. Since data protection ultimately means data modification, the challenge for SDC is to achieve protection with minimum loss of the accuracy sought by database users. In this chapter, we survey the current state of the art in SDC methods for protecting individual data (microdata). We discuss several information loss and disclosure risk measures and analyze several ways of combining them to assess the performance of the various methods. Last but not least, topics which need more research in the area are identified and possible directions hinted.

Keywords: Privacy, Inference control, Statistical disclosure control, Statistical disclosure limitation, Statistical databases, Microdata.

*This work received partial support from the Spanish Ministry of Science and Education through project SEG2004-04352-C04-01 "PROPRIETAS", the Government of Catalonia under grant 2005 SGR 00446 and Eurostat through the CENEX SDC project. The author is solely responsible for the views expressed in this chapter, which do not necessarily reflect the position of UNESCO nor commit that organization.
†Part of this chapter was written while the author was a Visiting Fellow at Princeton University.

Introduction

Inference control in statistical databases, also known as Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL), seeks to protect statistical data in such a way that they can be publicly released and mined without giving away private information that can be linked to specific individuals or entities. There are several areas of application of SDC techniques, which include but are not limited to the following:

- *Official statistics.* Most countries have legislation which compels national statistical agencies to guarantee statistical confidentiality when they release data collected from citizens or companies. This justifies the research on SDC undertaken by several countries, among them the European Union (*e.g.* the CASC project [8]) and the United States.

- *Health information.* This is one of the most sensitive areas regarding privacy. For example, in the U. S., the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA,[43]) requires the strict regulation of protected health information for use in medical research. In most western countries, the situation is similar.

- *E-commerce.* Electronic commerce results in the automated collection of large amounts of consumer data. This wealth of information is very useful to companies, which are often interested in sharing it with their subsidiaries or partners. Such consumer information transfer should not result in public profiling of individuals and is subject to strict regulation; see [28] for regulations in the European Union and [77] for regulations in the U.S.

The protection provided by SDC techniques normally entails some degree of data modification, which is an intermediate option between no modification (maximum utility, but no disclosure protection) and data encryption (maximum protection but no utility for the user without clearance).

The challenge for SDC is to modify data in such a way that sufficient protection is provided while keeping at a minimum the information loss, *i.e.* the loss of the accuracy sought by database users. In the years that have elapsed since the excellent survey by [3], the state of the art in SDC has evolved so that now at least three subdisciplines are clearly differentiated:

Tabular data protection This is the oldest and best established part of SDC, because tabular data have been the traditional output of national statistical offices. The goal here is to publish *static* aggregate information, *i.e.* tables, in such a way that no confidential information on specific individuals among those to which the table refers can be inferred. See [79] for a conceptual survey and [36] for a software survey.

Dynamic databases The scenario here is a database to which the user can submit statistical queries (sums, averages, etc.). The aggregate information obtained by a user as a result of successive queries should not allow him to infer information on specific individuals. Since the 80s, this has been known to be a difficult problem, subject to the tracker attack [69]. One possible strategy is to perturb the answers to queries; solutions based on perturbation can be found in [26], [54] and [76]. If perturbation is not acceptable and exact answers are needed, it may become necessary to refuse answers to certain queries; solutions based on query restriction can be found in [9] and [38]. Finally, a third strategy is to provide correct (unperturbed) interval answers, as done in [37] and [35].

Microdata protection This subdiscipline is about protecting static individual data, also called microdata. It is only recently that data collectors (statistical agencies and the like) have been persuaded to publish microdata. Therefore, microdata protection is the youngest subdiscipline and is experiencing continuous evolution in the last years.

Good general works on SDC are [79, 45]. This survey will cover the current state of the art in SDC methods for microdata, the most common data used for data mining. First, the main existing methods will be described. Then, we will discuss several information loss and disclosure risk measures and will analyze several approaches to combining them when assessing the performance of the various methods. The comparison metrics being presented should be used as a benchmark for future developments in this area. Open research issues and directions will be suggested at the end of this chapter.

Plan of this chapter

Section 1 introduces a classification of microdata protection methods. Section 2 reviews perturbative masking methods. Section 3 reviews non-perturbative masking methods. Section 4 reviews methods for synthetic microdata generation. Section 5 discusses approaches to trade off infor-

mation loss for disclosure risk and analyzes their strengths and limitations. Conclusions and directions for future research are summarized in Section 6.

1. A classification of microdata protection methods

A microdata set $\mathbf{V}$ can be viewed as a file with n records, where each record contains m attributes on an individual respondent. The attributes can be classified in four categories which are not necessarily disjoint:

- *Identifiers*. These are attributes that *unambiguously* identify the respondent. Examples are the passport number, social security number, name-surname, etc.
- *Quasi-identifiers or key attributes*. These are attributes which identify the respondent with some degree of ambiguity. (Nonetheless, a combination of quasi-identifiers may provide unambiguous identification.) Examples are address, gender, age, telephone number, etc.
- *Confidential outcome attributes*. These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.
- *Non-confidential outcome attributes*. Those attributes which do not fall in any of the categories above.

Since the purpose of SDC is to prevent confidential information from being linked to specific respondents, we will assume in what follows that original microdata sets to be protected have been pre-processed to remove from them all identifiers.

The purpose of microdata SDC mentioned in the previous section can be stated more formally by saying that, given an original microdata set $\mathbf{V}$, the goal is to release a protected microdata set $\mathbf{V}'$ in such a way that:

1 Disclosure risk (*i.e.* the risk that a user or an intruder can use $\mathbf{V}'$ to determine confidential attributes on a specific individual among those in $\mathbf{V}$) is low.

2 User analyses (regressions, means, etc.) on $\mathbf{V}'$ and on $\mathbf{V}$ yield the same or at least similar results.

Microdata protection methods can generate the protected microdata set $\mathbf{V}'$

- either by *masking original data*, *i.e.* generating $\mathbf{V}'$ a modified version of the original microdata set $\mathbf{V}$;

- or by *generating synthetic data* $\mathbf{V}'$ that preserve some statistical properties of the original data $\mathbf{V}$.

Masking methods can in turn be divided in two categories depending on their effect on the original data [79]:

- *Perturbative*. The microdata set is distorted before publication. In this way, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset; such confusion is beneficial for preserving statistical confidentiality. The perturbation method used should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset.

- *Non-perturbative*. Non-perturbative methods do not alter data; rather, they produce partial suppressions or reductions of detail in the original dataset. Global recoding, local suppression and sampling are examples of non-perturbative masking.

At a first glance, synthetic data seem to have the philosophical advantage of circumventing the re-identification problem: since published records are invented and do not derive from any original record, some authors claim that no individual having supplied original data can complain from having been re-identified. At a closer look, some authors (*e.g.*, [80] and [63]) claim that even synthetic data might contain some records that allow for re-identification of confidential information. In short, synthetic data overfitted to original data might lead to disclosure just as original data would. On the other hand, a clear problem of synthetic data is data utility: only the statistical properties explicitly selected by the data protector are preserved, which leads to the question whether the data protector should not directly publish the statistics he wants preserved rather than a synthetic microdata set. We will return to these issues in Section 4.

So far in this section, we have classified microdata protection methods by their operating principle. If we consider the type of data on which they can be used, a different dichotomic classification applies:

- *Continuous*. An attribute is considered continuous if it is numerical and arithmetic operations can be performed with it. Examples are income and age. Note that a numerical attribute does not

necessarily have an infinite range, as is the case for age. When designing methods to protect continuous data, one has the advantage that arithmetic operations are possible, and the drawback that every combination of numerical values in the original dataset is likely to be unique, which leads to disclosure if no action is taken.

- *Categorical.* An attribute is considered categorical when it takes values over a finite set and standard arithmetic operations do not make sense. Ordinal and nominal scales can be distinguished among categorical attributes. In ordinal scales the order between values is relevant, whereas in nominal scales it is not. In the former case, max and min operations are meaningful while in the latter case only pairwise comparison is possible. The instruction level is an example of ordinal attribute, whereas eye color is an example of nominal attribute. In fact, all quasi-identifiers in a microdata set are normally categorical nominal. When designing methods to protect categorical data, the inability to perform arithmetic operations is certainly inconvenient, but the finiteness of the value range is one property that can be successfully exploited.

2. Perturbative masking methods

Perturbative methods allow for the release of the entire microdata set, although perturbed values rather than exact values are released. Not all perturbative methods are designed for continuous data; this distinction is addressed further below for each method.

Most perturbative methods reviewed below (including additive noise, rank swapping, microaggregation and post-randomization) are special cases of matrix masking. If the original microdata set is $\mathbf{X}$, then the masked microdata set $\mathbf{Z}$ is computed as

$$\mathbf{Z} = \mathbf{AXB} + \mathbf{C}$$

where $\mathbf{A}$ is a record-transforming mask, $\mathbf{B}$ is an attribute-transforming mask and $\mathbf{C}$ is a displacing mask (noise)[27].

Table 1.1 lists the perturbative methods described below. For each method, the table indicates whether it is suitable for continuous and/or categorical data.

Additive noise

The noise additions algorithms in the literature are:

- *Masking by uncorrelated noise addition.* The vector of observations x_j for the j-th attribute of the original dataset X_j is replaced by

Table 1.1. Perturbative methods vs data types. "X" denotes applicable and "(X)" denotes applicable with some adaptation

Method	*Continuous data*	*Categorical data*
Additive noise	X	
Microaggregation	X	(X)
Rank swapping	X	X
Rounding	X	
Resampling	X	
PRAM		X
MASSC		X

a vector

$$z_j = x_j + \epsilon_j$$

where ϵ_j is a vector of normally distributed errors drawn from a random variable $\varepsilon_j \sim N(0, \sigma^2_{\varepsilon_j})$, such that $Cov(\varepsilon_t, \varepsilon_l) = 0$ for all $t \neq l$. This does not preserve variances nor correlations

- *Masking by correlated noise addition.* Correlated noise addition also preserves means and additionally allows preservation of correlation coefficients. The difference with the previous method is that the covariance matrix of the errors is now proportional to the covariance matrix of the original data, *i.e.* $\varepsilon \sim N(0, \Sigma_\varepsilon)$, where $\Sigma_\varepsilon = \alpha\Sigma$.

- *Masking by noise addition and linear transformation.* In [49], a method is proposed that ensures by additional transformations that the sample covariance matrix of the masked attributes is an unbiased estimator for the covariance matrix of the original attributes.

- *Masking by noise addition and nonlinear transformation.* An algorithm combining simple additive noise and nonlinear transformation is proposed in [72]. The advantages of this proposal are that it can be applied to discrete attributes and that univariate distributions are preserved. Unfortunately, as justified in [6], the application of this method is very time-consuming and requires expert knowledge on the data set and the algorithm.

For more details on specific algorithms, the reader can check [5]. In practice, only simple noise addition (two first variants) or noise addition with linear transformation are used. When using linear transformations, a decision has to be made whether to reveal them to the data user to allow for bias adjustment in the case of subpopulations.

With the exception of the not very practical method of [72], additive noise is not suitable to protect categorical data. On the other hand, it is well suited for continuous data for the following reasons:

- It makes no assumptions on the range of possible values for V_i (which may be infinite).
- The noise being added is typically continuous and with mean zero, which suits well continuous original data.
- No exact matching is possible with external files. Depending on the amount of noise added, approximate (interval) matching might be possible.

Microaggregation

Microaggregation is a family of SDC techniques for continous microdata. The rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if records correspond to groups of k or more individuals, where no individual dominates (*i.e.* contributes too much to) the group and k is a threshold value. Strict application of such confidentiality rules leads to replacing individual values with values computed on small aggregates (microaggregates) prior to publication. This is the basic principle of microaggregation.

To obtain microaggregates in a microdata set with n records, these are combined to form g groups of size at least k. For each attribute, the average value over each group is computed and is used to replace each of the original averaged values. Groups are formed using a criterion of maximal similarity. Once the procedure has been completed, the resulting (modified) records can be published.

The optimal k-partition (from the information loss point of view) is defined to be the one that maximizes within-group homogeneity; the higher the within-group homogeneity, the lower the information loss, since microaggregation replaces values in a group by the group centroid. The sum of squares criterion is common to measure homogeneity in clustering. The within-groups sum of squares SSE is defined as

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i)$$

The lower SSE, the higher the within group homogeneity. Thus, in terms of sums of squares, the optimal k-partition is the one that minimizes SSE.

For a microdata set consisting of p attributes, these can be microaggregated together or partitioned into several groups of attributes. Also

the way to form groups may vary. Several taxonomies are possible to classify the microaggregation algorithms in the literature: i) fixed group size [15, 44, 23] vs variable group size [15, 51, 18, 68, 50, 20]; ii) exact optimal (only for the univariate case, [41, 55]) vs heuristic microaggregation; iii) continuous vs categorical microaggregation [75].

To illustrate, we next give a heuristic algorithm called MDAV (Maximum Distance to Average Vector,[23]) for multivariate fixed group size microaggregation on unprojected continuous data. We designed and implemented MDAV for the μ-Argus package [44].

ALGORITHM 1.1 (MDAV)

1 *Compute the average record $\bar{x}$ of all records in the dataset. Consider the most distant record x_r to the average record $\bar{x}$ (using the squared Euclidean distance).*

2 *Find the most distant record x_s from the record x_r considered in the previous step.*

3 *Form two groups around x_r and x_s, respectively. One group contains x_r and the $k-1$ records closest to x_r. The other group contains x_s and the $k-1$ records closest to x_s.*

4 *If there are at least $3k$ records which do not belong to any of the two groups formed in Step 3, go to Step 1 taking as new dataset the previous dataset minus the groups formed in the last instance of Step 3.*

5 *If there are between $3k-1$ and $2k$ records which do not belong to any of the two groups formed in Step 3: a) compute the average record $\bar{x}$ of the remaining records; b) find the most distant record x_r from $\bar{x}$; c) form a group containing x_r and the $k-1$ records closest to x_r; d) form another group containing the rest of records. Exit the Algorithm.*

6 *If there are less than $2k$ records which do not belong to the groups formed in Step 3, form a new group with those records and exit the Algorithm.*

The above algorithm can be applied independently to each group of attributes resulting from partitioning the set of attributes in the dataset.

Data swapping and rank swapping

Data swapping was originally presented as an SDC method for databases containing only categorical attributes [11]. The basic idea behind the method is to transform a database by exchanging values of confidential attributes among individual records. Records are exchanged in such a way that low-order frequency counts or marginals are maintained.

Even though the original procedure was not very used in practice (see [32]), its basic idea had a clear influence in subsequent methods. In [59] and [58] data swapping was introduced to protect continuous and categorical microdata, respectively. Another variant of data swapping for microdata is *rank swapping*, which will be described next in some detail.

Although originally described only for ordinal attributes [40], rank swapping can also be used for any numerical attribute [53]. First, values of an attribute X_i are ranked in ascending order, then each ranked value of X_i is swapped with another ranked value randomly chosen within a restricted range (*e.g.* the rank of two swapped values cannot differ by more than $p\%$ of the total number of records, where p is an input parameter). This algorithm is independently used on each original attribute in the original data set.

It is reasonable to expect that multivariate statistics computed from data swapped with this algorithm will be less distorted than those computed after an unconstrained swap. In earlier empirical work by these authors on continuous microdata protection [21], rank swapping has been identified as a particularly well-performing method in terms of the trade-off between disclosure risk and information loss (see Example 1.4 below). Consequently, it is one of the techniques that have been implemented in the $\mu - Argus$ package [44].

EXAMPLE 1.2 *In Table 1.2, we can see an original microdata set on the left and its rankswapped version on the right. There are four attributes and ten records in the original dataset; the second attribute is alphanumeric, and the standard alphabetic order has been used to rank it. A value of* $p = 10\%$ *has been used for all attributes.* □

Rounding

Rounding methods replace original values of attributes with rounded values. For a given attribute X_i, rounded values are chosen among a set of rounding points defining a *rounding set* (often the multiples of a given base value). In a multivariate original dataset, rounding is usually performed one attribute at a time (*univariate* rounding); however,

Table 1.2. Example of rank swapping. Left, original file; right, rankswapped file

1	K	3.7	4.4	1	H	3.0	4.8
2	L	3.8	3.4	2	L	4.5	3.2
3	N	3.0	4.8	3	M	3.7	4.4
4	M	4.5	5.0	4	N	5.0	6.0
5	L	5.0	6.0	5	L	4.5	5.0
6	H	6.0	7.5	6	F	6.7	9.5
7	H	4.5	10.0	7	K	3.8	11.0
8	F	6.7	11.0	8	H	6.0	10.0
9	D	8.0	9.5	9	C	10.0	7.5
10	C	10.0	3.2	10	D	8.0	3.4

multivariate rounding is also possible [79, 10]. The operating principle of rounding makes it suitable for continuous data.

Resampling

Originally proposed for protecting tabular data [42, 17], resampling can also be used for microdata. Take t independent samples $S_1, \cdots, S_t$ of the values of an original attribute X_i. Sort all samples using the same ranking criterion. Build the masked attribute Z_i as $\bar{x}_1, \cdots, \bar{x}_n$, where n is the number of records and $\bar{x}_j$ is the average of the j-th ranked values in $S_1, \cdots, S_t$.

PRAM

The Post-RAndomization Method (PRAM, [39]) is a probabilistic, perturbative method for disclosure protection of categorical attributes in microdata files. In the masked file, the scores on some categorical attributes for certain records in the original file are changed to a different score according to a prescribed probability mechanism, namely a Markov matrix. The Markov approach makes PRAM very general, because it encompasses noise addition, data suppression and data recoding.

PRAM information loss and disclosure risk largely depend on the choice of the Markov matrix and are still (open) research topics [14].

The PRAM matrix contains a row for each possible value of each attribute to be protected. This rules out using the method for continuous data.

MASSC

MASSC [71] is a masking method whose acronym summarizes its four steps: Micro Agglomeration, Substitution, Subsampling and Calibration. We briefly recall the purpose of those four steps:

1 Micro agglomeration is applied to partition the original dataset into risk strata (groups of records which are at a similar risk of disclosure). These strata are formed using the key attributes, *i.e.* the quasi-identifiers in the records. The idea is that those records with rarer combinations of key attributes are at a higher risk.

2 Optimal probabilistic substitution is then used to perturb the original data.

3 Optimal probabilistic subsampling is used to suppress some attributes or even entire records.

4 Optimal sampling weight calibration is used to preserve estimates for outcome attributes in the treated database whose accuracy is critical for the intended data use.

MASSC in interesting in that, to the best of our knowledge, it is the first attempt at designing a perturbative masking method in such a way that disclosure risk can be analytically quantified. Its main shortcoming is that its disclosure model simplifies reality by considering only disclosure resulting from linkage of key attributes with external sources. Since key attributes are typically categorical, the risk of disclosure can be analyzed by looking at the probability that a sample unique is a population unique; however, doing so ignores the fact that continuous outcome attributes can also be used for respondent re-identification via record linkage. As an example, if respondents are companies and turnover is one outcome attribute, everyone in a certain industrial sector knows which is the company with largest turnover. Thus, in practice, MASSC is a method only suited when continuous attributes are not present.

3. Non-perturbative masking methods

Non-perturbative methods do not rely on distortion of the original data but on partial suppressions or reductions of detail. Some of the methods are usable on both categorical and continuous data, but others are not suitable for continuous data. Table 1.3 lists the non-perturbative methods described below. For each method, the table indicates whether it is suitable for continuous and/or categorical data.

Table 1.3. Non-perturbative methods vs data types

Method	*Continuous data*	*Categorical data*
Sampling		X
Global recoding	X	X
Top and bottom coding	X	X
Local suppression		X

Sampling

Instead of publishing the original microdata file, what is published is a sample S of the original set of records [79].

Sampling methods are suitable for categorical microdata, but for continuous microdata they should probably be combined with other masking methods. The reason is that sampling alone leaves a continuous attribute V_i unperturbed for all records in S. Thus, if attribute V_i is present in an external administrative public file, unique matches with the published sample are very likely: indeed, given a continuous attribute V_i and two respondents o_1 and o_2, it is highly unlikely that V_i will take the same value for both o_1 and o_2 unless $o_1 = o_2$ (this is true even if V_i has been truncated to represent it digitally).

If, for a continuous identifying attribute, the score of a respondent is only approximately known by an attacker (as assumed in [78]), it might still make sense to use sampling methods to protect that attribute. However, assumptions on restricted attacker resources are perilous and may prove definitely too optimistic if good quality external administrative files are at hand.

Global recoding

This method is also sometimes known as generalization [67, 66]. For a categorical attribute V_i, several categories are combined to form new (less specific) categories, thus resulting in a new V_i' with $|D(V_i')| < |D(V_i)|$ where $|\cdot|$ is the cardinality operator. For a continuous attribute, global recoding means replacing V_i by another attribute V_i' which is a discretized version of V_i. In other words, a potentially infinite range $D(V_i)$ is mapped onto a finite range $D(V_i')$. This is the technique used in the μ-Argus SDC package [44].

This technique is more appropriate for categorical microdata, where it helps disguise records with strange combinations of categorical attributes. Global recoding is used heavily by statistical offices.

EXAMPLE 1.3 *If there is a record with "Marital status = Widow/er" and "Age = 17", global recoding could be applied to "Marital status" to create a broader category "Widow/er or divorced", so that the probability of the above record being unique would diminish. Global recoding can also be used on a continuous attribute, but the inherent discretization leads very often to an unaffordable loss of information. Also, arithmetical operations that were straightforward on the original V_i are no longer easy or intuitive on the discretized V_i'.* □

Top and bottom coding

Top and bottom coding is a special case of global recoding which can be used on attributes that can be ranked, that is, continuous or categorical ordinal. The idea is that top values (those above a certain threshold) are lumped together to form a new category. The same is done for bottom values (those below a certain threshold). See [44].

Local suppression

Certain values of individual attributes are suppressed with the aim of increasing the set of records agreeing on a combination of key values. Ways to combine local suppression and global recoding are discussed in [16] and implemented in the μ-Argus SDC package [44].

If a continuous attribute V_i is part of a set of key attributes, then each combination of key values is probably unique. Since it does not make sense to systematically suppress the values of V_i, we conclude that local suppression is rather oriented to categorical attributes.

4. Synthetic microdata generation

Publication of synthetic —*i.e.* simulated— data was proposed long ago as a way to guard against statistical disclosure. The idea is to randomly generate data with the constraint that certain statistics or internal relationships of the original dataset should be preserved.

We next review some approaches in the literature to synthetic data generation and then proceed to discuss the global pros and cons of using synthetic data.

Synthetic data by multiple imputation

More than ten years ago, it was suggested in [65] to create an entirely synthetic dataset based on the original survey data and multiple imputation. Rubin's proposal was more completely developed in [57]. A

simulation study of it was given in [60]. In [64] inference on synthetic data is discussed and in [63] an application is given.

We next sketch the operation of the original proposal by Rubin. Consider an original microdata set X of size n records drawn from a much larger population of N individuals, where there are background attributes A, non-confidential attributes B and confidential attributes C. Background attributes are observed and available for all N individuals in the population, whereas B and C are only available for the n records in the sample X. The first step is to construct from X a multiply-imputed population of N individuals. This population consists of the n records in X and M (the number of multiple imputations, typically between 3 and 10) matrices of (B, C) data for the $N - n$ non-sampled individuals. The variability in the imputed values ensures, theoretically, that valid inferences can be obtained on the multiply-imputed population. A model for predicting (B, C) from A is used to multiply-impute (B, C) in the population. The choice of the model is a nontrivial matter. Once the multiply-imputed population is available, a sample Z of n' records can be drawn from it whose structure looks like the one a sample of n' records drawn from the original population. This can be done M times to create M replicates of (B, C) values. The result are M multiply-imputed synthetic datasets. To make sure no original data are in the synthetic datasets, it is wise to draw the samples from the multiply-imputed population excluding the n original records from it.

Synthetic data by bootstrap

Long ago, [30] proposed generating synthetic microdata by using bootstrap methods. Later, in [31] this approach was used for categorical data.

The bootstrap approach bears some similarity to the data distortion by probability distribution and the multiple-imputation methods described above. Given an original microdata set X with p attributes, the data protector computes its empirical p-variate cumulative distribution function (c.d.f.) F. Now, rather than distorting the original data to obtain masked data (as done by the masking methods in Sections 2 and 3), the data protector alters (or "smoothes") the c.d.f. F to derive a similar c.d.f. F'. Finally, F' is sampled to obtain a synthetic microdata set Z.

Synthetic data by Latin Hypercube Sampling

Latin Hypercube Sampling (LHS) appears in the literature as another method for generating multivariate synthetic datasets. In [46], the LHS updated technique of [33] was improved, but the proposed scheme is still time-intensive even for a moderate number of records. In [12], LHS

is used along with a rank correlation refinement to reproduce both the univariate (*i.e.* mean and covariance) and multivariate structure (in the sense of rank correlation) of the original dataset. In a nutshell, 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.

Partially synthetic data by Cholesky decomposition

Generating plausible synthetic values for all attributes in a database may be difficult in practice. Thus, several authors have considered mixing actual and synthetic data.

In [7], a non-iterative method for generating continuous synthetic microdata is proposed. It consists of three methods sketched next. Informally, suppose two sets of attributes X and Y, where the former are the confidential outcome attributes and the latter are quasi-identifier attributes. Then X are taken as independent and Y as dependent attributes. Conditional on the specific confidential attributes x_i, the quasi-identifier attributes Y_i are assumed to follow a multivariate normal distribution with covariance matrix $\Sigma = \{\sigma_{jk}\}$ and a mean vector $x_i B$, where B is a matrix of regression coefficients.

Method A computes a multiple regression of Y on X and fitted Y'_A attributes. Finally, attributes X and Y'_A are released in place of X and Y.

If a user fits a multiple regression model to (y'_A, x), she will get estimates $\hat{B}_A$ and $\hat{\Sigma}_A$ which, in general, are different from the estimates $\hat{B}$ and $\hat{\Sigma}$ obtained when fitting the model to the original data (y, x). IPSO Method B modifies y'_A into y'_B in such a way that the estimate $\hat{B}_B$ obtained by multiple linear regression from (y'_B, x) satisfies $\hat{B}_B = \hat{B}$.

A more ambitious goal is to come up with a data matrix y'_C such that, when a multivariate multiple regression model is fitted to (y'_C, x), *both* sufficient statistics $\hat{B}$ and $\hat{\Sigma}$ obtained on the original data (y, x) are preserved. This is achieved by IPSO Method C.

Other partially synthetic and hybrid microdata approaches

The multiple imputation approach described in [65] for creating entirely synthetic microdata can be extended for partially synthetic microdata. As a result multiply-imputed, partially synthetic datasets are obtained that contain a mix of actual and imputed (synthetic) values. The idea is to multiply-impute confidential values and release non-

confidential values without perturbation. This approach was first applied to protect the Survey of Consumer Finances [47, 48]. In Abowd and Woodcock [1, 2], this technique was adopted to protect longitudinal linked data, that is, microdata that contain observations from two or more related time periods (successive years, etc.). Methods for valid inference on this kind of partial synthetic data were developed in [61] and a non-parametric method was presented in [62] to generate multiply-imputed, partially synthetic data.

Closely related to multiply imputed, partially synthetic microdata is model-based disclosure protection [34, 56]. In this approach, a set of confidential continuous outcome attributes is regressed on a disjoint set of non-confidential attributes; then the fitted values are released for the confidential attributes instead of the original values.

A different approach called hybrid masking was proposed in [13]. The idea is to compute masked data as a combination of original and synthetic data. Such a combination allows better control than purely synthetic data over the individual characteristics of masked records. For hybrid masking to be feasible, a rule must be used to pair one original data record with one synthetic data record. An option suggested in [13] is to go through all original data records and pair each original record with the nearest synthetic record according to some distance. Once records have been paired, [13] suggest two possible ways for combining one original record X with one synthetic record X_s: additive combination and multiplicative combination. Additive combination yields

$$Z = \alpha X + (1 - \alpha) X_s$$

and multiplicative combination yields

$$Z = X^{\alpha} \cdot X_s^{(1-\alpha)}$$

where α is an input parameter in $[0, 1]$ and Z is the hybrid record. [13] present empirical results comparing the hybrid approach with rank swapping and microaggregation masking (the synthetic component of hybrid data is generated using Latin Hypercube Sampling [12]).

Another approach to combining original and synthetic microdata is proposed in [70]. The idea here is to first mask an original dataset using a masking method (see Sections 2 and 3 above). Then a hill-climbing optimization heuristic is run which seeks to modify the masked data to preserve the first and second-order moments of the original dataset as much as possible without increasing the disclosure risk with respect to the initial masked data. The optimization heuristic can be modified to preserve higher-order moments, but this significantly increases computation. Also, the optimization heuristic can take as initial dataset a

random dataset instead of a masked dataset; in this case, the output dataset is purely synthetic.

Pros and cons of synthetic microdata

As pointed out in Section 1, synthetic data are appealing in that, at a first glance, they seem to circumvent the re-identification problem: since published records are invented and do not derive from any original record, it might be concluded that no individual can complain from having been re-identified. At a closer look this advantage is less clear. If, by chance, a published synthetic record matches a particular citizen's non-confidential attributes (age, marital status, place of residence, etc.) and confidential attributes (salary, mortgage, etc.), re-identification using the non-confidential attributes is easy and that citizen may feel that his confidential attributes have been unduly revealed. In that case, the citizen is unlikely to be happy with or even understand the explanation that the record was synthetically generated.

On the other hand, limited data utility is another problem of synthetic data. Only the statistical properties explicitly captured by the model used by the data protector are preserved. A logical question at this point is why not directly publish the statistics one wants to preserve rather than release a synthetic microdata set.

One possible justification for synthetic microdata would be if valid analyses could be obtained on a number of subdomains, *i.e.* similar results were obtained in a number of subsets of the original dataset and the corresponding subsets of the synthetic dataset. Partially synthetic or hybrid microdata are more likely to succeed in staying useful for subdomain analysis. However, when using partially synthetic or hybrid microdata, we lose the attractive feature of purely synthetic data that the number of records in the protected (synthetic) dataset is independent from the number of records in the original dataset.

5. Trading off information loss and disclosure risk

Sections 1 through 4 have presented a plethora of methods to protect microdata. To complicate things further, most of such methods are parametric (*e.g.*, in microaggregation, one parameter is the minimum number of records in a cluster), so the user must go through two choices rather than one: a primary choice to select a method and a secondary choice to select parameters for the method to be used. To help reducing the *embarras du choix*, some guidelines are needed.

Score construction

The mission of SDC to modify data in such a way that sufficient protection is provided at minimum information loss suggests that a good SDC method is one achieving a good tradeoff between disclosure risk and information loss.

Following this idea, [21] proposed a score for method performance rating based on the average of information loss and disclosure risk measures. For each method M and parameterization P, the following score is computed:

$$Score(\mathbf{V}, \mathbf{V}') = \frac{IL(\mathbf{V}, \mathbf{V}') + DR(\mathbf{V}, \mathbf{V}')}{2}$$

where IL is an information loss measure, DR is a disclosure risk measure and $\mathbf{V}'$ is the protected dataset obtained after applying method M with parameterization P to an original dataset $\mathbf{V}$.

In [21] and [19] IL and DR were computed using a weighted combination of several information loss and disclosure risk measures. With the resulting score, a ranking of masking methods (and their parameterizations) was obtained. In [81] the line of the above two papers was followed to rank a different set of methods using a slightly different score.

To illustrate how a score can be constructed, we next describe the particular score used in [21].

EXAMPLE 1.4 *Let X and X' be matrices representing original and protected datasets, respectively, where all attributes are numerical. Let V and R be the covariance matrix and the correlation matrix of X, respectively; let $\bar{X}$ be the vector of attribute averages for X and let S be the diagonal of V. Define V', R', $\bar{X}'$, and S' analogously from X'. The Information Loss (IL) is computed by averaging the mean variations of $X-X'$, $\bar{X}-\bar{X}'$, $V-V'$, $S-S'$, and the mean absolute error of $R-R'$ and multiplying the resulting average by 100. Thus, we obtain the following expression for information loss:*

$$IL = \frac{100}{5}\Big(\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij}-x'_{ij}|}{|x_{ij}|}}{np} + \frac{\sum_{j=1}^{p}\frac{|\bar{x}_j-\bar{x'}_j|}{|\bar{x}_j|}}{p} +$$

$$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{|v_{ij}-v'_{ij}|}{|v_{ij}|}}{\frac{p(p+1)}{2}} + \frac{\sum_{j=1}^{p}\frac{|v_{jj}-v'_{jj}|}{|v_{jj}|}}{p} + \frac{\sum_{j=1}^{p}\sum_{1\le i\le j}|r_{ij}-r'_{ij}|}{\frac{p(p-1)}{2}}\Big)$$

The expression of the overall score is obtained by combining information loss and information risk as follows:

$$Score = \frac{IL + \frac{(0.5DLD + 0.5PLD) + ID}{2}}{2}$$

Here, DLD (Distance Linkage Disclosure risk) is the percentage of correctly linked records using distance-based record linkage [19], PLD (Probabilistic Linkage Record Disclosure risk) is the percentage of correctly linked records using probabilistic linkage [29], ID (Interval Disclosure) is the percentage of original records falling in the intervals around their corresponding masked values and IL is the information loss measure defined above.

Based on the above score, [21] found that, for the benchmark datasets and the intruder's external information they used, two good performers among the set of methods and parameterizations they tried were: i) rankswapping with parameter p around 15 (see description above); ii) multivariate microaggregation on unprojected data taking groups of three attributes at a time (Algorithm 1.1 with partitioning of the set of attributes). □

Using a score permits to regard the selection of a masking method and its parameters as an optimization problem. This idea was first used in the above-mentioned contribution [70]. In that paper, a masking method was applied to the original data file and then a post-masking optimization procedure was applied to decrease the score obtained.

On the negative side, no specific score weighting can do justice to all methods. Thus, when ranking methods, the values of all measures of information loss and disclosure risk should be supplied along with the overall score.

R-U maps

A tool which may be enlightening when trying to construct a score or, more generally, optimize the tradeoff between information loss and disclosure risk is a graphical representation of pairs of measures (disclosure risk, information loss) or their equivalents (disclosure risk, data utility). Such maps are called R-U confidentiality maps [24, 25]). Here, R stands for disclosure risk and U for data utility. According to [25], "in its most basic form, an R-U confidentiality map is the set of paired values (R, U), of disclosure risk and data utility that correspond to various strategies for data release" (*e.g.*, variations on a parameter). Such (R, U) pairs are typically plotted in a two-dimensional graph, so that the user can easily grasp the influence of a particular method and/or parameter choice.

k-anonymity

A different approach to facing the conflict between information loss and disclosure risk is suggested by Samarati and Sweeney [67, 66, 73, 74]. A protected dataset is said to satisfy k-anonymity for $k > 1$ if, for each combination of quasi-identifier values (*e.g.* address, age, gender, etc.), at least k records exist in the dataset sharing that combination. Now if, for a given k, k-anonymity is assumed to be enough protection, one can concentrate on minimizing information loss with the only constraint that k-anonymity should be satisfied. This is a clean way of solving the tension between data protection and data utility. Since k-anonymity is usually achieved via generalization (equivalent to global recoding, as said above) and local suppression, minimizing information loss usually translates to reducing the number and/or the magnitude of suppressions.

k-anonymity bears some resemblance to the underlying principle of microaggregation and is a useful concept because quasi-identifiers are usually categorical or can be categorized, *i.e.* they take values in a finite (and ideally reduced) range. However, re-identification is not necessarily based on categorical quasi-identifiers: sometimes, numerical outcome attributes —which are continuous and often cannot be categorized— give enough clues for re-identification (see discussion on the MASSC method above). Microaggregation was suggested in [23] as a possible way to achieve k-anonymity for numerical, ordinal and nominal attributes. A similar idea called data condensation had also been independently proposed by [4] to achieve k-anonymity for the specific case of numerical attributes.

Another connection between k-anonymity and microaggregation is the NP-hardness of solving them optimally. Satisfying k-anonymity with minimal data modification has been shown to be NP-hard in [52], which is parallel to the NP-hardness of optimal multivariate microaggregation proven in [55].

6. Conclusions and research directions

Inference control methods for privacy-preseving data mining are a hot research topic progressing very fast. There are still many open issues, some of which can be hopefully solved with further research and some which are likely to stay open due to the inherent nature of SDC.

We first list some of the issues that we feel can be and should be settled in the near future:

- Identifying a comprehensive listing of data uses (*e.g.* regression models, association rules, etc.) that would allow the definition of data use-specific information loss measures broadly accepted

by the community; those new measures could complement and/or replace the generic measures currently used. Work in this line has been started in Europe in 2006 under the CENEX SDC project sponsored by Eurostat.

- Devising disclosure risk assessment procedures which are as universally applicable as record linkage while being less greedy in computational terms.
- Identifying, for each domain of application, which are the external data sources that intruders can typically access in order to attempt re-identification. This would help data protectors figuring out in more realistic terms which are the disclosure scenarios they should protect data against.
- Creating one or several benchmarks to assess the performance of SDC methods. Benchmark creation is currently hampered by the confidentiality of the original datasets to be protected. Data protectors should agree on a collection of non-confidential original-looking data sets (financial datasets, population datasets, etc.) which can be used by anybody to compare the performance of SDC methods. The benchmark should also incorporate state-of-the-art disclosure risk assessment methods, which requires continuous update and maintenance.

There are other issues which, in our view, are less likely to be resolved in the near future, due to the very nature of SDC methods. As pointed out in [22], if an intruder knows the SDC algorithm used to create a protected data set, he can mount algorithm-specific re-identification attacks which can disclose more confidential information than conventional data mining attacks. Keeping secret the SDC algorithm used would seem a solution, but in many cases the protected dataset itself gives some clues on the SDC algorithm used to produce it. Such is the case for a rounded, microaggregated or partially suppressed microdata set. Thus, it is unclear to what extent the SDC algorithm used can be kept secret.

Other data security areas where slightly distorted data are sent to a recipient who is legitimate but untrusted also share the same concerns about the secrecy of protection algorithms in use. This is the case of watermarking. Teaming up with those areas sharing similar problems is probably one clever line of action for SDC.

References

[1] J. M. Abowd and S. D. Woodcock. Disclosure limitation in longitudinal linked tables. In P. Doyle, J. I. Lane, J. J. Theeuwes, and L. V. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 215–278, Amsterdam, 2001. North-Holland.

[2] J. M. Abowd and S. D. Woodcock. Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 290–297, Berlin Heidelberg, 2004. Springer.

[3] N. R. Adam and J. C. Wortmann. Security-control for statistical databases: a comparative study. *ACM Computing Surveys*, 21(4):515–556, 1989.

[4] C. C. Aggarwal and P. S. Yu. A condensation approach to privacy preserving data mining. In E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, E. Ferrari, editors, *Advances in Database Technology - EDBT 2004*, vol. 2992 of *Lecture Notes in Computer Science*, pages 183-199, Berlin Heidelberg, 2004. Springer.

[5] R. Brand. Microdata protection through noise addition. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 97–116, Berlin Heidelberg, 2002. Springer.

[6] R. Brand. Tests of the applicability of Sullivan's algorithm to synthetic data and real business data in official statistics, 2002. European Project IST-2000-25069 CASC, Deliverable 1.1-D1, http://neon.vb.cbs.nl/casc.

[7] J. Burridge. Information preserving statistical obfuscation. *Statistics and Computing*, 13:321–327, 2003.

[8] CASC. Computational aspects of statistical confidentiality, 2004. European project IST-2000-25069 CASC, 5th FP, 2001-2004, http://neon.vb.cbs.nl/casc.

[9] F. Y. Chin and G. Ozsoyoglu. Auditing and inference control in statistical databases. *IEEE Transactions on Software Engineering*, SE-8:574–582, 1982.

[10] L. H. Cox and J. J. Kim. Effects of rounding on the quality and confidentiality of statistical data. In J. Domingo-Ferrer and L. Franconi, editors, *Privacy in Statistical Databases-PSD 2006*, volume 4302 of *Lecture Notes in Computer Science*, pages 48–56, Berlin Heidelberg, 2006.

[11] T. Dalenius and S. P. Reiss. Data-swapping: a technique for disclosure control (extended abstract). In *Proc. of the ASA Section on Survey Research Methods*, pages 191–194, Washington DC, 1978. American Statistical Association.

[12] R. Dandekar, M. Cohen, and N. Kirkendall. Sensitive micro data protection using Latin hypercube sampling technique. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 245–253, Berlin Heidelberg, 2002. Springer.

[13] R. Dandekar, J. Domingo-Ferrer, and F. Sebé. LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 153–162, Berlin Heidelberg, 2002. Springer.

[14] P.-P. de Wolf. Risk, utility and PRAM. In J. Domingo-Ferrer and L. Franconi, editors, *Privacy in Statistical Databases-PSD 2006*, volume 4302 of *Lecture Notes in Computer Science*, pages 189–204, Berlin Heidelberg, 2006.

[15] D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In *Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204, Ottawa, 1993. Statistics Canada.

[16] A. G. DeWaal and L. C. R. J. Willenborg. Global recodings and local suppressions in microdata sets. In *Proceedings of Statistics Canada Symposium'95*, pages 121–132, Ottawa, 1995. Statistics Canada.

[17] J. Domingo-Ferrer and J. M. Mateo-Sanz. On resampling for statistical confidentiality in contingency tables. *Computers & Mathematics with Applications*, 38:13–32, 1999.

[18] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[19] 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 *Pre-proceedings of ETK-NTTS'2001 (vol. 2)*, pages 807–826, Luxemburg, 2001. Eurostat.

[20] J. Domingo-Ferrer, F. Sebé, and A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Computers & Mathematics with Applications*, 2007. (To appear).

[21] J. Domingo-Ferrer and V. Torra. A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J. I. Lane, J. J. M. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 111–134, Amsterdam, 2001. North-Holland. http://vneumann.etse.urv.es/publications/bcpi.

[22] J. Domingo-Ferrer and V. Torra. Algorithmic data mining against privacy protection methods for statistical databases. *manuscript*, 2004.

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

[24] G. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman, and S. F. Roehrig. Disclosure limitation methods and information loss for tabular data. In P. Doyle, J. I. Lane, J. J. Theeuwes, and L. V. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 135–166, Amsterdam, 2001. North-Holland.

[25] G. T. Duncan, S. A. Keller-McNulty, and S. L Stokes. Disclosure risk vs. data utility: The R-U confidentiality map, 2001.

[26] G. T. Duncan and S. Mukherjee. Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise. *Journal of the American Statistical Association*, 95:720–729, 2000.

[27] G. T. Duncan and R. W. Pearson. Enhancing access to microdata while protecting confidentiality: prospects for the future. *Statistical Science*, 6:219–239, 1991.

[28] E.U.Privacy. European privacy regulations, 2004. http://europa.eu.int/comm/internal_market/privacy/law_en.htm.

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

[30] S. E. Fienberg. A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical Report 611, Carnegie Mellon University Department of Statistics, 1994.

[31] S. E. Fienberg, U. E. Makov, and R. J. Steele. Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics*, 14(4):485–502, 1998.

[32] S. E. Fienberg and J. McIntyre. Data swapping: variations on a theme by Dalenius and Reiss. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 14–29, Berlin Heidelberg, 2004. Springer.

[33] A. Florian. An efficient sampling scheme: updated Latin hypercube sampling. *Probabilistic Engineering Mechanics*, 7(2):123–130, 1992.

[34] L. Franconi and J. Stander. A model based method for disclosure limitation of business microdata. *Journal of the Royal Statistical Society D - Statistician*, 51:1–11, 2002.

[35] R. Garfinkel, R. Gopal, and D. Rice. New approaches to disclosure limitation while answering queries to a database: protecting numerical confidential data against insider threat based on data and algorithms, 2004. Manuscript. Available at http://www-eio.upc.es/seminar/04/garfinkel.pdf.

[36] S. Giessing. Survey on methods for tabular data protection in Argus. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 1–13, Berlin Heidelberg, 2004. Springer.

[37] R. Gopal, R. Garfinkel, and P. Goes. Confidentiality via camouflage: the CVC approach to disclosure limitation when answering queries to databases. *Operations Research*, 50:501–516, 2002.

[38] R. Gopal, P. Goes, and R. Garfinkel. Interval protection of confidential information in a database. *INFORMS Journal on Computing*, 10:309–322, 1998.

[39] J. M. Gouweleeuw, P. Kooiman, L. C. R. J. Willenborg, and P.-P. DeWolf. Post randomisation for statistical disclosure control: Theory and implementation, 1997. Research paper no. 9731 (Voorburg: Statistics Netherlands).

[40] B. Greenberg. Rank swapping for ordinal data, 1987. Washington, DC: U. S. Bureau of the Census (unpublished manuscript).

[41] S. L. Hansen and S. Mukherjee. A polynomial algorithm for optimal univariate microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):1043–1044, 2003.

[42] G. R. Heer. A bootstrap procedure to preserve statistical confidentiality in contingency tables. In D. Lievesley, editor, *Proc. of the International Seminar on Statistical Confidentiality*, pages 261–271, Luxemburg, 1993. Office for Official Publications of the European Communities.

[43] HIPAA. Health insurance portability and accountability act, 2004. http://www.hhs.gov/ocr/hipaa/.

[44] A. Hundepool, A. Van de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing. *μ-ARGUS version 4.0 Software and User's Manual*. Statistics Netherlands, Voorburg NL, may 2005. http://neon.vb.cbs.nl/casc.

[45] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte-Nordholt, G. Seri, and P.-P. DeWolf. *Handbook on Statistical Disclosure Control (version 1.0)*. Eurostat (CENEX SDC Project Deliverable), 2006.

[46] D. E. Huntington and C. S. Lyrintzis. Improvements to and limitations of Latin hypercube sampling. *Probabilistic Engineering Mechanics*, 13(4):245–253, 1998.

[47] A. B. Kennickell. Multiple imputation and disclosure control: the case of the 1995 Survey of Consumer Finances. In *Record Linkage Techniques*, pages 248–267, Washington DC, 1999. National Academy Press.

[48] A. B. Kennickell. Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. In J. Domingo-Ferrer, editor, *Statistical Data Protection*, pages 248–267, Luxemburg, 1999. Office for Official Publications of the European Communities.

[49] J. J. Kim. A method for limiting disclosure in microdata based on random noise and transformation. In *Proceedings of the Section on Survey Research Methods*, pages 303–308, Alexandria VA, 1986. American Statistical Association.

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

[51] J. M. Mateo-Sanz and J. Domingo-Ferrer. A method for data-oriented multivariate microaggregation. In J. Domingo-Ferrer, editor, *Statistical Data Protection*, pages 89–99, Luxemburg, 1999. Office for Official Publications of the European Communities.

[52] A. Meyerson and R. Williams. General k-anonymization is hard. Technical Report 03-113, Carnegie Mellon School of Computer Science (USA), 2003.

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

[54] K. Muralidhar, D. Batra, and P. J. Kirs. Accessibility, security and accuracy in statistical databases: the case for the multiplicative fixed data perturbation approach. *Management Science*, 41:1549–1564, 1995.

[55] A. Oganian and J. Domingo-Ferrer. On the complexity of optimal microaggregation for statistical disclosure control. *Statistical Journal of the United Nations Economic Comission for Europe*, 18(4):345–354, 2001.

[56] S. Polettini, L. Franconi, and J. Stander. Model based disclosure protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 83–96, Berlin Heidelberg, 2002. Springer.

[57] T. J. Raghunathan, J. P. Reiter, and D. Rubin. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19(1):1–16, 2003.

[58] S. P. Reiss. Practical data-swapping: the first steps. *ACM Transactions on Database Systems*, 9:20–37, 1984.

[59] S. P. Reiss, M. J. Post, and T. Dalenius. Non-reversible privacy transformations. In *Proceedings of the ACM Symposium on Principles of Database Systems*, pages 139–146, Los Angeles, CA, 1982. ACM.

[60] J. P. Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531–544, 2002.

[61] J. P. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29:181–188, 2003.

[62] J. P. Reiter. Using CART to generate partially synthetic public use microdata, 2003. Duke University working paper.

[63] J. P. Reiter. Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A*, 168:185–205, 2005.

[64] J. P. Reiter. Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131(2):365–377, 2005.

[65] D. B. Rubin. Discussion of statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.

[66] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[67] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, 1998.

[68] G. Sande. Exact and approximate methods for data directed microaggregation in one or more dimensions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):459–476, 2002.

[69] J. Schlörer. Disclosure from statistical databases: quantitative aspects of trackers. *ACM Transactions on Database Systems*, 5:467–492, 1980.

[70] F. Sebé, J. Domingo-Ferrer, J. M. Mateo-Sanz, and V. Torra. Postmasking optimization of the tradeoff between information loss and

disclosure risk in masked microdata sets. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 163–171, Berlin Heidelberg, 2002. Springer.

[71] A. C. Singh, F. Yu, and G. H. Dunteman. MASSC: A new data mask for limiting statistical information loss and disclosure. In H. Linden, J. Riecan, and L. Belsby, editors, *Work Session on Statistical Data Confidentiality 2003*, Monographs in Official Statistics, pages 373–394, Luxemburg, 2004. Eurostat.

[72] G. R. Sullivan. *The Use of Added Error to Avoid Disclosure in Microdata Releases*. PhD thesis, Iowa State University, 1989.

[73] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):571–588, 2002.

[74] L. Sweeney. k-anonimity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):557–570, 2002.

[75] V. Torra. Microaggregation for categorical variables: a median based approach. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 162–174, Berlin Heidelberg, 2004. Springer.

[76] J. F. Traub, Y. Yemini, and H. Wozniakowski. The statistical security of a statistical database. *ACM Transactions on Database Systems*, 9:672–679, 1984.

[77] U.S.Privacy. U. S. Privacy Regulations, 2004. http://www.media-awareness.ca/english/issues/privacy/us_legislation_privacy.cfm.

[78] L. Willenborg and T. DeWaal. *Statistical Disclosure Control in Practice*. Springer-Verlag, New York, 1996.

[79] L. Willenborg and T. DeWaal. *Elements of Statistical Disclosure Control*. Springer-Verlag, New York, 2001.

[80] W. E. Winkler. Re-identification methods for masked microdata. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 216–230, Berlin Heidelberg, 2004. Springer.

[81] W. E. Yancey, W. E. Winkler, and R. H. Creecy. Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 135–152, Berlin Heidelberg, 2002. Springer.