Statistical Databases

Josep Domingo-Ferrer, Rovira i Virgili University of Tarragona

November 16, 2007

Introduction

Statistical databases are databases containing statistical information. Such databases are normally released by national statistical institutes but, on occasion, they can also be released by healthcare authorities (epidemiology) or by private organizations (*e.g.* consumer surveys). Statistical databases typically come in three formats:

- *Tabular data*, that is, tables with counts or magnitudes, which are the classical output of official statistics;
- *Queryable databases*, that is, on-line databases to which the user can submit statistical queries (sums, averages, etc.);
- *Microdata*, that is, files where each record contains information on an individual (a citizen or a company).

The peculiarity of statistical databases is that they should provide useful *statistical* information, but they should not reveal private information on the individuals they refer to (respondents). Indeed, supplying data to national statistical institutes is compulsory in most countries but, in return, those institutes commit to preserving the privacy of respondents. Inference control in statistical databases, also known as Statistical Disclosure Control (SDC), is a discipline that seeks to protect data in statistical databases so that they can be published without revealing confidential information that can be linked to specific individuals among those to which the data correspond. SDC is applied to protect *respondent privacy* in 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 [1], a distinction is made between SDC and other technologies for database privacy, like privacy-preserving data mining (PPDM) or private information retrieval (PIR): what makes the difference between those technologies is whose privacy they seek. While SDC is aimed at respondent privacy, the primary goal of PPDM is to protect owner privacy when several database owners wish to co-operate in joint analyses across their databases without giving away their original data to each other. On its side, the primary goal of PIR is user privacy, that is, to allow the user of a database to retrieve some information item without the database exactly knowing which item was recovered.

The literature on SDC started in the 1970s, with the seminal contribution by Dalenius [2] in the statistical community and the works by Schlörer and Denning [3, 4] in the database community. The 1980s saw moderate activity in this field. An excellent survey of the state of the art at the end of the 1980s is [5]. In the 1990s, there was renewed interest in the statistical community and the discipline was further developed under the names of statistical disclosure control in Europe and statistical disclosure limitation in America. Subsequent evolution has resulted in at least three clearly differentiated subdisciplines:

- *Tabular data protection.* 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 [6] for a conceptual survey.

- *Queryable databases.* 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 late 70s, this has been known to be a difficult problem, subject to the tracker attack [4]. SDC strategies here include perturbation, query restriction and camouflage (providing interval answers rather than exact answers).

- *Microdata protection.* 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. Its purpose is to mask the original microdata so that the masked microdata are still analytically useful but cannot be linked to the original respondents.

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) 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) 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.

Theory

Formal definition of data formats

A *microdata* file $\mathbf{X}$ with s respondents and t attributes is an $s \times t$ matrix where X_{ij} is the value of attribute j for respondent i. Attributes can be numerical (*e.g.* age, salary) or categorical (*e.g.* gender, job).

The attributes in a microdata set 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 key attributes 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.* Other attributes which contain non-sensitive information on the respondent.

From microdata, *tabular data* can be generated by crossing one or more categorical attributes. Formally, a table is a function

$$T : D(X_{i1}) \times D(X_{i2}) \times \cdots \times D(X_{il}) \rightarrow \mathbb{R} \text{ or } \mathbb{N}$$

where $l \leq t$ is the number of crossed categorical attributes and $D(X_{ij})$ is the domain where attribute X_{ij} takes its values.

There are two kinds of tables: *frequency tables* that display the count of respondents at the crossing of the categorical attributes (in $\mathbb{N}$) and *magnitude tables* that display information on a numerical attribute at the crossing of the categorical attributes (in $\mathbb{R}$). For example, given some census microdata containing attributes "Job" and "Town", one can generate a *frequency table* displaying the count of respondents doing each job type in each town. If the census microdata also contain the "Salary" attribute, one can generate a magnitude table displaying the average salary for each job type in each town. The number n of cells in a table is normally much less than the number s of respondent records in a microdata file. However, tables must satisfy several linear constraints: marginal row and column totals. Additionally, a set of tables is called *linked* if they share some of the crossed categorical attributes: for example "Job" × "Town" is linked to "Job" × "Gender".

Overview of methods

Statistical disclosure control will be first reviewed for tabular data, then for queryable databases and finally for microdata.

Methods for tabular data

In spite of tables displaying aggregate information, there is risk of disclosure in tabular data release. Several attacks are conceivable:

- *External attack.* For example, let a frequency table "Job" × "Town" be released where there is a single respondent for job J_i and town T_j. Then if a magnitude table is released with the average salary for each job type and each town, the exact salary of the only respondent with job J_i working in town T_j is publicly disclosed.
- *Internal attack.* Even if there are two respondents for job J_i and town T_j, the salary of each of them is disclosed to each other.
- *Dominance attack.* If one (or a few) respondents dominate in the contribution to a cell of a magnitude table, the dominant respondent(s) can upper-bound the contributions of the rest (*e.g.* if the table displays the total salary for each job type and town and one individual contributes 90% of that salary, he knows that his colleagues in the town are not doing very well).

SDC methods for tables fall into two classes: non-perturbative and perturbative. *Non-perturbative methods* do not modify the values in the tables; the best known method in this class is *cell suppression* (CS). *Perturbative methods* output a table with some modified values; well-known methods in this class include *controlled rounding* (CR) and the recent *controlled tabular adjustment* (CTA).

The idea of CS is to suppress those cells that are identified as sensitive, *i.e.* from which the above attacks can extract sensitive information, by the so-called sensitivity rules (*e.g.* the dominance rule, which identifies a cell as sensitive if it is vulnerable to a dominance attack). Sensitive cells are the primary suppressions. Then additional suppressions (secondary suppressions) are performed to prevent primary suppressions from being computed or even inferred within a prescribed protection interval using the row and column constraints (marginal row and column totals). Usually, one attempts to minimize either the number of secondary suppressions or their pooled magnitude, which results in complex optimization problems. Most optimization methods used are heuristic, based on mixed integer linear programming or network flows [8], most of them implemented in the τ-Argus free software package [9]. CR rounds values in the table to multiples of a rounding base. This may entail rounding the marginal totals as well. On its side, CTA modifies the values in the table to prevent the inference of values of sensitive cells within a prescribed protection interval. The idea of CTA is to find the *closest* table to the original one that ensures such a protection for all sensitive cells. This requires optimization methods, which are typically based on mixed linear integer programming. Usually CTA entails less information loss than CS.

Methods for queryable databases

In SDC of queryable databases, there are three main approaches to protect a confidential vector of numerical data from disclosure through answers to user queries:

- *Data perturbation.* Perturbing the data is a simple and effective approach whenever the users do not require deterministically correct answers to queries that are functions of the confidential vector. Perturbation can be applied to the records on which queries are computed (input perturbation) or to the query result after computing it on the original data (output perturbation). Perturbation methods can be found in [10, 11, 12].
- *Query restriction.* This is the right approach if the user does require deterministically correct answers and these answers have to be exact (*i.e.* a number). Since exact answers to queries provide the user with a very powerful information, it may become necessary to

refuse to answer certain queries at some stage to avoid disclosure of a confidential datum. There are several criteria to decide whether a query can be answered; one of them is query set size control, that is, to refuse answers to queries which affect a set of records which is too small. An example of the query restriction approach can be found in [13].

- *Camouflage.* If deterministically correct non-exact answers (*i.e.* small interval answers) suffice, confidentiality via camouflage (CVC, [14]) is a good option. With this approach, unlimited answers to any conceivable query types are allowed. The idea of CVC is to "camouflage" the confidential vector a by making it part of the relative interior of a compact set Π of vectors. Then each query $q = f(a)$ is answered with an inverval $[q^-, q^+]$ containing $[f^-, f^+]$, where f^- and f^+ are, respectively, the minimum and the maximum of f over Π.

Methods for microdata

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

- either by *masking original data*, *i.e.* generating $\mathbf{X}'$ a modified version of the original microdata set $\mathbf{X}$;
- or by *generating synthetic data* $\mathbf{X}'$ that preserve some statistical properties of the original data $\mathbf{X}$.

Masking methods can in turn be divided in two categories depending on their effect on the original data [6]:

- *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. *Noise addition, microaggregation, data/rank swapping, microdata rounding, resampling and PRAM* are examples of perturbative masking methods (see [8] for details).

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

While a reasonable overview of the methods for protecting tables or queryable databases has been given above, microdata protection methods are more diverse, so that a description of some of them is needed.

Some microdata protection methods

Additive noise

Additive noise is a family of perturbative masking methods. 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 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 neither 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 $\Sigma_\varepsilon = \alpha\Sigma$ with Σ being the covariance matrix of the original data.

- *Masking by noise addition and linear transformation.* In [15], 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.* Combining simple additive noise and nonlinear transformation has also been proposed, in such a way that application to discrete attributes is possible and univariate distributions are preserved. Unfortunately, the application of this method is very time-consuming and requires expert knowledge on the data set and the algorithm. See [8] for more details.

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.

In general, 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 X_i (which may be infinite).

- The noise being added is typically continuous and with mean zero, which suits well continuous original data.

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.

Given 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 [16, 17, 18] vs variable group size [19, 20, 21]; ii) exact optimal (only for the univariate case, [22]) vs heuristic microaggregation (the rest of the microaggregation literature); iii) categorical [18, 23] vs continuous (the rest of references cited in this paragraph).

To illustrate, we next give a heuristic algorithm called MDAV (Maximum Distance to Average Vector,[18, 24]) for multivariate fixed group size microaggregation on unprojected continuous data. We designed and implemented MDAV for the μ-Argus package [17].

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 a perturbative SDC method for databases containing only categorical attributes. 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, its basic idea had a clear influence in subsequent methods. A variant of data swapping for microdata is *rank swapping*, which will be described next in some detail.

Although originally described only for ordinal attributes [25], rank swapping can also be used for any numerical attribute. 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.

PRAM

The Post-RAndomization Method (PRAM, [26]) 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 called the PRAM matrix. The Markov approach makes PRAM very general, because it encompasses noise addition, data suppression and data recoding.

Since the PRAM matrix must contain a row for each possible value of each attribute to be protected, PRAM cannot be used for continuous data.

Sampling

This is a non-perturbative masking method. Instead of publishing the original microdata file, what is published is a sample S of the original set of records [6].

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 X_i unperturbed for all records in S. Thus, if attribute X_i is present in an external administrative public file, unique matches with the published sample are very likely: indeed, given a continuous attribute X_i and two respondents o_1 and o_2, it is highly unlikely that X_i will take the same value for both o_1 and o_2 unless $o_1 = o_2$ (this is true even if X_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, 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 is a non-perturbative masking method, also known sometimes as generalization. For a categorical attribute X_i, several categories are combined to form new (less specific) categories, thus resulting in a new X_i' with $|D(X_i')| < |D(X_i)|$ where $|\cdot|$ is the cardinality operator. For a continuous attribute, global recoding means replacing X_i by another attribute X_i' which is a discretized version of X_i. In other words, a potentially infinite range $D(X_i)$ is mapped onto a finite range $D(X_i')$. This is the technique used in the μ-Argus SDC package [17].

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. 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 X_i are no longer easy or intuitive on the discretized X_i'.

Top and bottom coding

Top and bottom coding are special cases 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 [17].

Local suppression

This is a non-perturbative masking method in which 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 implemented in the μ-Argus SDC package [17].

If a continuous attribute X_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 X_i, we conclude that local suppression is rather oriented to categorical attributes.

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.

More than ten years ago, Rubin suggested in [27] to create an entirely synthetic dataset based on the original survey data and multiple imputation. A simulation study of this approach was given in [28].

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 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.

Evaluation

Evaluation of SDC methods must be carried out in terms of data utility and disclosure risk.

Measuring data utility

Defining what a generic utility loss measure is can be a tricky issue. Roughly speaking, such a definition should capture the amount of information loss for a reasonable range of data uses.

We will attempt a definition on the data with maximum granularity, that is, microdata. Similar definitions apply to rounded tabular data; for tables with cell suppressions, utility is normally measured as the reciprocal of the number of suppressed cells or their pooled magnitude. As to queryable databases, they can be logically viewed as tables as far as data utility is concerned: a denied query answer is equivalent to a cell suppression and a perturbed answer is equivalent to a perturbed cell.

We will say there is little information loss if the protected dataset is analytically valid and interesting according to the following definitions by [29]:

- A protected microdata set is *analytically valid* if it approximately preserves the following with respect to the original data (some conditions apply only to continuous attributes):

1. Means and covariances on a small set of subdomains (subsets of records and/or attributes)
2. Marginal values for a few tabulations of the data
3. At least one distributional characteristic

- A microdata set is *analytically interesting* if six attributes on important subdomains are provided that can be validly analyzed.

More precise conditions of analytical validity and analytical interest cannot be stated without taking specific data uses into account. As imprecise as they may be, the above definitions suggest some possible measures:

- Compare raw records in the original and the protected dataset. The more similar the SDC method to the identity function, the less the impact (but the higher the disclosure risk!). This requires pairing records in the original dataset and records in the protected dataset. For masking methods, each record in the protected dataset is naturally paired to the record in the original dataset it originates from. For synthetic protected datasets, pairing is less obvious.
- Compare some statistics computed on the original and the protected datasets. The above definitions list some statistics which should be preserved as much as possible by an SDC method.

A strict evaluation of information loss must be based on the data uses to be supported by the protected data. The greater the differences between the results obtained on original and protected data for those uses, the higher the loss of information. However, very often microdata protection cannot be performed in a data use specific manner, for the following reasons:

- Potential data uses are very diverse and it may be even hard to identify them all at the moment of data release by the data protector.

- Even if all data uses could be identified, releasing several versions of the same original dataset so that the i-th version has an information loss optimized for the i-th data use may result in unexpected disclosure.

Since that data often must be protected with no specific data use in mind, generic information loss measures are desirable to guide the data protector in assessing how much harm is being inflicted to the data by a particular SDC technique.

Information loss measures for numerical data. Assume a microdata set with n individuals (records) $I_1, I_2, \cdots, I_n$ and p continuous attributes $Z_1, Z_2, \cdots, Z_p$. Let X be the matrix representing the original microdata set (rows are records and columns are attributes). Let X' be the matrix representing the protected microdata set. The following tools are useful to characterize the information contained in the dataset:

- Covariance matrices V (on X) and V' (on X').

- Correlation matrices R and R'.

- Correlation matrices RF and RF' between the p attributes and the p factors $PC_1, \cdots, PC_p$ obtained through principal components analysis.

- Communality between each of the p attributes and the first principal component PC_1 (or other principal components PC_i's). Communality is the percent of each attribute that is explained by PC_1 (or PC_i). Let C be the vector of communalities for X and C' the corresponding vector for X'.

- Factor score coefficient matrices F and F'. Matrix F contains the factors that should multiply each attribute in X to obtain its projection on each principal component. F' is the corresponding matrix for X'.

There does not seem to be a single quantitative measure which completely reflects those structural differences. Therefore, we proposed in [30, 31] to measure information loss through the discrepancies between matrices X, V, R, RF, C and F obtained on the original data and the corresponding X', V', R', RF', C' and F' obtained on the protected dataset. In particular, discrepancy between correlations is related to the information loss for data uses such as regressions and cross tabulations.

Matrix discrepancy can be measured in at least three ways:

Mean square error Sum of squared componentwise differences between pairs of matrices, divided by the number of cells in either matrix.

Mean absolute error Sum of absolute componentwise differences between pairs of matrices, divided by the number of cells in either matrix.

Mean variation Sum of absolute percent variation of components in the matrix computed on protected data with respect to components in the matrix computed on original data, divided by the number of cells in either matrix. This approach has the advantage of not being affected by scale changes of attributes.

The following table summarizes the measures proposed in the above references. In this table, p is the number of attributes, n the number of records, and components of matrices are represented by the corresponding lowercase letters (*e.g.* x_{ij} is a component of matrix X). Regarding $X - X'$ measures, it makes also sense to compute those on the averages of attributes rather than on all data (call this variant $\bar{X} - \bar{X}'$). Similarly, for $V - V'$ measures, it would also be sensible to use them to compare only the variances of the attributes, *i.e.* to compare the diagonals of the covariance matrices rather than the whole matrices (call this variant $S - S'$).

	Mean square error	Mean abs. error	Mean variation
$X - X'$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
$V - V'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
$R - R'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$
$RF - RF'$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\lvert rf_{ij}-rf'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\frac{\lvert rf_{ij}-rf'_{ij}\rvert}{\lvert rf_{ij}\rvert}}{p^2}$
$C - C'$	$\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$	$\frac{\sum_{i=1}^{p}\lvert c_i-c'_i\rvert}{p}$	$\frac{\sum_{i=1}^{p}\frac{\lvert c_i-c'_i\rvert}{\lvert c_i\rvert}}{p}$
$F - F'$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}(f_{ij}-f'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\lvert f_{ij}-f'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p} w_j \sum_{i=1}^{p}\frac{\lvert f_{ij}-f'_{ij}\rvert}{\lvert f_{ij}\rvert}}{p^2}$

Information loss measures for categorical data. These can be based on direct comparison of categorical values, comparison of contingency tables, on Shannon's entropy. See [30] for more details.

Bounded information loss measures. The information loss measures discussed above are unbounded, *i.e.* they do not take values in a predefined interval. On the other hand, as discussed below, disclosure risk measures are naturally bounded (the risk of disclosure is naturally bounded between 0 and 1). Defining bounded information loss measures may be convenient to enable the data protector to trade off information loss against disclosure risk. In [32], probabilistic information loss measures bounded between 0 and 1 are proposed for continuous data.

Measuring disclosure risk

In the context of statistical disclosure control, disclosure risk can be defined as the risk that a user or an intruder can use the protected dataset $\mathbf{X}'$ to derive confidential information on an individual among those in the original dataset $\mathbf{X}$.

Disclosure risk can be regarded from two different perspectives:

1. **Attribute disclosure.** This approach to disclosure is defined as follows. Disclosure takes place when an attribute of an individual can be determined more accurately with access to the released statistic than it is possible without access to that statistic.

2. **Identity disclosure.** Attribute disclosure does not imply a disclosure of the identity of any individual. Identity disclosure takes place when a record in the protected dataset can be linked with a respondent's identity. Two main approaches are usually employed for measuring identity disclosure risk: uniqueness and re-identification.

   2.1. Uniqueness. Roughly speaking, the risk of identity disclosure is measured as the probability that rare combinations of attribute values in the released protected data are indeed rare in the original population the data come from. This approach is used typically used with non-perturbative statistical disclosure control methods and, more specifically, sampling. The reason that uniqueness is not used with perturbative methods is that, when protected attribute values are perturbed versions of original attribute values, it makes no sense to investigate the probability that a rare combination of protected values is rare in the original dataset, because *that* combination is most probably *not found* in the original dataset.

   2.2. Re-identification. This is an empirical approach to evaluate the risk of disclosure. In this case, record linkage software is constructed to estimate the number of re-identifications that might be obtained by a specialized intruder. Re-identification through record linkage provides a more unified approach than uniqueness methods because the former can be applied to any kind of masking and not just to non-perturbative masking. Moreover, re-identification can also be applied to synthetic data.

Trading off information loss and disclosure risk

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. Several approaches have been proposed to handle this trade-off. We discuss *SDC scores*, *R-U maps* and *k-anonymity*.

Score construction

Following this idea, [30] 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:

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

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

In [30] 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. To illustrate how a score can be constructed, we next describe the particular score used in [30].

Example. 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:

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

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

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 [30], PLD (Probabilistic Linkage Record Disclosure risk) is the percentage of correctly linked records using probabilistic linkage [33], 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, [30] 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 MDAV above with partitioning of the set of attributes). □

Using a score permits regarding the selection of a masking method and its parameters as an optimization problem. A masking method can be applied to the original data file and then a post-masking optimization procedure can be 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 [34]. Here, R stands for disclosure risk and U for data utility. 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 [35]. A protected dataset is said to satisfy k-anonymity for $k > 1$ if, for each combination of key attribute 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 multivariate microaggregation and is a useful concept because key attributes 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 key attributes: sometimes, numerical outcome attributes —which are continuous and often cannot be categorized— give enough clues for re-identification. Microaggregation was suggested in [18] as a possible way to achieve k-anonymity for numerical, ordinal and nominal attributes: the idea is to use multivariate microaggregation on the key attributes of the dataset.

Future

There are many open issues in SDC, 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 probably could 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 whose solution seems less likely in the near future, due to the very nature of SDC methods. 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.

References

[1] J. Domingo-Ferrer. A three-dimensional conceptual framework for database privacy. In *Secure Data Management-4th VLDB Workshop SDM'2007*, *Lecture Notes in Computer Science*, vol. 4721, pp. 193-202, 2007. Springer-Verlag.

[2] T. Dalenius. The invasion of privacy problem and statistics production. An overview. *Statistik Tidskrift*, 12: 213-225, 1974.

[3] J. Schlörer. Identification and retrieval of personal records from a statistical data bank. *Methods Inform. Med.*, 14(1):7-13, 1975.

[4] D. E. Denning, P. J. Denning and M. D. Schwartz. The tracker: a threat to statistical database security. *ACM Transactions on Database Systems*, 4(1): 76-96, 1979.

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

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

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

[8] 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. http://neon.vb.cbs.nl/CENEX/

[9] A. Hundepool, A. van de Wetering, R. Ramaswamy, P.-P. de Wolf, S. Giessing, M. Fischetti, J.-J. Salazar, J. Castro and P. Lowthian, *τ-ARGUS v. 3.2 Software and User's Manual*, CENEX SDC Project Deliverable, Feb. 2007. http://neon.vb.cbs.nl/casc/TAU.html

[10] 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*, 45:720-729, 2000.

[11] 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.

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

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

[22] 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.

[23] V. Torra. Microaggregation for categorical variables: a median based approach. In *Privacy in Statistical Databases-PSD 2004*, *Lecture Notes in Computer Science*, vol. 3050, pp. 162-174, 2004. Springer-Verlag.

[24] J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz and F. Sebé. Efficient multivariate data-oriented microaggregation. *VLDB Journal*, 15:355–369, 2006.

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

[26] 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).

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

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

[29] W. E. Winkler. Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Research in Official Statistics*, 1(2):50-69, 1998.

[30] 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.

[31] F. Sebé, J. Domingo-Ferrer, J. M. Mateo-Sanz and V. Torra. Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets. In *Inference Control in Statistical Databases*, *Lecture Notes in Computer Science*, vol. 2316, pp. 163-171, 2002. Springer-Verlag.

[32] J. M. Mateo-Sanz, J. Domingo-Ferrer and F. Sebé. Probabilistic information loss measures in confidentiality protection of continuous microdata.

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

[34] 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.

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

Reading list

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. http://neon.vb.cbs.nl/CENEX/

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

Cross-references

Statistical disclosure control. See Statistical Databases.

Statistical disclosure limitation. See Statistical Databases.

Inference control. See Statistical Databases.

Privacy in statistical databases. See Statistical Databases.