2. Database Privacy

Josep Domingo-Ferrer, David Sánchez, and Sara Hajian

Abstract Open data is a growing demand by data analysts, companies, and the general public. Yet, when databases to be publicly released contain information on individual respondents (e.g., responses to polls, census information, healthcare records, etc.), they must be released in a way that preserves the privacy of these respondents: it should be *de facto* impossible to relate the published data to specific individuals. To achieve this goal, the Statistical Disclosure Control (SDC) discipline has proposed a plethora of privacy protection methods, known under a variety of names such as SDC methods, anonymization methods, or sanitization methods. This chapter provides an overview of the issues in database privacy, a survey of the best-known SDC methods, a discussion on the related data privacy/utility trade-offs, and a description of privacy models proposed by the computer science community in recent years. Some relevant freeware packages are also identified.

1 Introduction

There is a growing social and economic demand for open data to improve planning, scientific research, market research, and so on. In particular, the public sector is under pressure to release as much information as it can in the name of transparency. Organizations releasing data include national statistical institutes, healthcare authorities (epidemiology), or even private organizations (e.g., consumer surveys).

When published data refer to individual respondents, care must be taken that the privacy of the latter is not violated. It should be *de facto* impossible to relate the published data to specific individuals. Indeed, supplying data to national statistical institutes is compulsory in most countries but, in return, those institutes commit to preserving the privacy of respondents. Hence, rather than publishing exactly accu-

Universitat Rovira i Virgili. Dept. of Computer Engineering and Mathematics. UNESCO Chair in Data Privacy. Av. Països Catalans 26. E-43007 Tarragona, Catalonia. e-mail: \{josep.domingo,david.sanchez,sara.hajian\}@urv.cat

rate information for each individual, the aim should be to provide useful *statistical* information, that is, to preserve as much as possible in the released data the statistical properties of the original data. This is why privacy-preserving databases on individuals are called *statistical databases*.

Statistical databases come in three main formats:

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

Inference control in statistical databases, also known as *Statistical Disclosure Control (SDC)*, *Statistical Disclosure Limitation (SDL)*, *database anonymization* or *database sanitization*, 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 whom 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 [16], 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 [12] in the statistical community and the works by Schlörer and Denning [62, 14] 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 [1]. 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. Towards the turn of the century, with the flourish of data mining, there was renewed activity in the database community, where the field was called data anonymization or data sanitization and was often confused with privacy-preserving data mining. Subsequent evolution has resulted in at least three clearly differentiated subdisciplines:

- **Tabular data protection.** The goal here is to publish *static* aggregate information, that is, tables, in such a way that no confidential information on specific individuals among those to whom the table refers can be inferred. See [72] for a conceptual survey.

- **Queryable databases.** The aggregate information obtained by a user as a result of successive queries should not allow him or her to infer information on specific individuals. Since the late 1970s, this has been known to be a difficult problem, subject to the tracker attack [14]. 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.

The rest of this chapter is organized as follows. Section 2 introduces the basic concepts used throughout the chapter. In Section 3, we detail algorithms and mechanisms for sanitizing (i.e., anonymizing) the records in a database. These algorithms seek to output a sanitized version of data that satisfies a privacy definition (prevents disclosure risks) and has high utility. Section 4 is devoted to ways of measuring disclosure risk and the utility of sanitized data while the formal definitions of privacy models are presented in Section 5. Section 6 explores outstanding challenges that must be addressed in the future and opportunities for new research directions. The final section concludes the chapter and lists relevant software.

2 Background

In this section, we introduce some basic definitions and concepts that are used throughout this chapter related to data formats (Section 2.1) and sanitization of each format (Section 2.2).

2.1 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 that are not necessarily disjoint:

- **Identifiers**. These are attributes that *unambiguously* identify the respondent. Examples are the passport number, social security number, name-surname, and so on.
- **Quasi-identifiers or key attributes**. These are attributes that 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, and so on.

- **Confidential (a.k.a. sensitive) attributes**. These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, and so on.
- **Non-confidential (a.k.a. non-sensitive) 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

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

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”.

2.2 Basic Sanitization Concepts

We will review sanitization/anonymization concepts used for each data format: tabular data, queryable databases, and microdata.

2.2.1 Sanitization of 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, the dominant respondent knows that his or her colleagues in the town are not doing very well).

Sanitization 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).

2.2.2 Sanitization of 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 [24, 55, 71].
- **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 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 [11].
- **Camouflage**. If deterministically correct non-exact answers (i.e., small interval answers) suffice, confidentiality via camouflage (CVC, [30]) 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.

2.2.3 Sanitization of 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 [72]:

- **Perturbative**. The microdata set is distorted before publication. In this way, unique combinations of scores in the original data set may disappear and new unique combinations may appear in the perturbed data set; such confusion is beneficial for preserving statistical confidentiality. The perturbation method used should be such that statistics computed on the perturbed data set do not differ significantly from the statistics that would be obtained on the original data set. *Noise addition, microaggregation, data/rank swapping, microdata rounding, resampling and PRAM* are examples of perturbative masking methods (see the next Section and [42] for details).
- **Non-perturbative**. Non-perturbative methods do not alter data; rather, they produce partial suppressions or reductions of detail in the original data set. Sampling, global recoding, top and bottom coding, and local suppression are examples of non-perturbative masking methods.

3 Database Sanitization Methods

Data publishing organizations usually face a fundamental trade-off between privacy and utility.

The two extreme policies are the following:

- To release no data in order to maintain total privacy.
- To release original data without any modification to maximize data utility, without regard to privacy protection.

In this section, we detail methods based on the concepts introduced in Section 2 that offer good trade-offs between the two above extreme policies. We focus on *microdata sanitization methods*, because microdata are the most detailed type of data. In fact, based on protected microdata, one can also obtain protected tables and protected query answers: just build tables and compute query answers based on the protected microdata records.

Each sanitization method consists of an algorithm instantiating in a specific way a generic sanitization mechanism. We first discuss methods based on deterministic sanitization mechanisms, and then methods based on randomized sanitization mechanisms.

3.1 Deterministic Sanitization Mechanisms

3.1.1 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

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

where x_{ij} indicates the microaggregated version of the attribute value. 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 [13, 40, 21] vs variable group size [19, 44]; ii) exact optimal (only for the univariate case, [37]) vs heuristic microaggregation (the rest of the microaggregation literature); iii) categorical [21] 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[21, 18]) for multivariate fixed group size microaggregation on unprojected continuous data. We designed and implemented MDAV for the μ-Argus package [40]. In the algorithm below we assume $n \geq k$.

1. Compute the average record $\bar{x}$ of all records in the data set. 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 data set the previous data set 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 data set.

3.1.2 Bucketization

Like microaggregation, bucketization (also known as Anatomy, [75]) partitions the input data into non-overlapping buckets. However, rather than summarizing records in each bucket into one average record, the bucketization approach simply breaks the connection between quasi-identifier and confidential attributes. The bucketization mechanism produces a sanitized data set by first partitioning the original data set into non-overlapping groups (or buckets) and then, for each group, releasing its projection on the quasi-identifier attributes and also its projection on the confidential attributes. The idea is that, after bucketization, the confidential attribute values of an individual are indistinguishable from those of any other individual in the same bucket.

3.1.3 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 [32], 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.

3.1.4 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 [40]. 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'.

3.1.5 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 [40].

3.1.6 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 [40].

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.

3.2 Randomized Sanitization Mechanisms

3.2.1 Additive noise

Additive noise is a family of perturbative masking methods. The noise addition algorithms in the literature are:

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

  $$z_j = x_j + \varepsilon_j$$

  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 [43], 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 [42] for more details.

3.2.2 PRAM

The Post-RAndomization Method (PRAM, [31]) 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.

3.2.3 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 [72]. 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.

3.2.4 Synthetic microdata generation

Publication of synthetic, that is, 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 data set should be preserved. More than 20 years ago, Rubin suggested in [57] to create an entirely synthetic data set based on the original survey data and multiple imputation. A simulation study of this approach was given in [56].

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 of 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 or her 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, that is, similar results were obtained in a number of subsets of the original data set and the corresponding subsets of the synthetic data set. 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) data set is independent from the number of records in the original data set.

4 Evaluation

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

4.1 Measuring Data Utility

Defining what a generic utility loss measure is can be a tricky issue [39]. 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 data set is analytically valid and interesting according to the following definitions by [73]:

- 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 a significant number of attributes (say half a dozen) 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 data set. 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 data set and records in the protected data set. For masking methods, each record in the protected data set is naturally paired to the record in the original data set it originates from. For synthetic protected data sets, pairing is less obvious.
- Compare some statistics computed on the original and the protected data sets. 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 data set 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 data set:

- 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 [20, 66] 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 data set. 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.

Information loss measures for categorical data. These have been usually based on direct comparison of categorical values, comparison of contingency tables, or on Shannon's entropy [20]. More recently, the importance of the semantics underlying categorical data for data utility has been realized [51]. As a result, semantically-grounded information loss measures have been proposed both to measure the practical utility and guide the sanitization algorithms [23]. Since this is an ongoing research line, it is further discussed in Section 6 on Challenges and Opportunities.

Bounded information loss measures. The information loss measures discussed above are unbounded, that is, they do not take values in a predefined interval. On the other hand, as discussed in Section 4.2, 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 [52], probabilistic information loss measures bounded between 0 and 1 are proposed for continuous data.

4.2 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 data set $\mathbf{X}'$ to derive confidential information on an individual among those in the original data set $\mathbf{X}$ [15]. 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 data set 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 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 data set, because *that* combination is most probably *not found* in the original data set.

2.2. Record linkage. 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, record linkage can also be applied to synthetic data.

In the specific setting of tabular data protection, Bayesian methods for disclosure risk assessment have been proposed [15].

4.3 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 sanitization method is one achieving a good trade-off 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*.

4.3.1 Score construction

Following this idea, [20] 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 data set obtained after applying method *M* with parameterization *P* to an original data set **X**. In [20] *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.

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.

4.3.2 R-U maps

A tool which may be enlightening when trying to construct a score or, more generally, optimize the trade-off 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 [26].

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.

5 Privacy Models

The computer science community has also contributed to sanitization for disclosure control under the names *Privacy Preserving Data Publishing* (PPDP) [2, 3, 9] and *Privacy Preserving Data Mining* (PPDM) [28, 29]. The former focuses on privacy-preserving publication of microdata, whereas the latter focuses on bringing privacy protection to traditional data mining tasks (for example, data classification or clustering).

There is a substantial difference between the sanitization approaches by the statistical and the computer science communities:

- ***A posteriori* disclosure risk control**. The statistical community is mainly concerned with analytical validity, so it first applies a sanitization method that incurs tolerable information loss and then measures the disclosure risk that publishing the sanitized data would incur (*a posteriori* control). If the extant disclosure risk is too high, then sanitization is re-applied to the original data with higher information loss. The process is iterated until tolerable disclosure risk is obtained.

- ***A priori* disclosure risk control**. In the computer science community, the primary focus in on disclosure risk. A *privacy model* is used to select the tolerable disclosure risk level from the outset (*a priori* control). Then a sanitization method is applied which guarantees by design that the selected disclosure risk level is not exceeded. The incurred information loss is measured after sanitization has been completed.

  We next review the two main privacy models used in the literature.

5.1 k-Anonymity

A common approach to prevent disclosure via record linkage attacks is to hide each individual record within a group. This is the approach that k-anonymity [59, 58, 68] takes:

Definition 1. (k-Anonymity) A data set is said to satisfy k-anonymity for an integer $k > 1$ if, for each combination of values of quasi-identifier attributes, at least k records exist in the data set sharing that combination.

To achieve k-anonymity, identifying attributes are removed and quasi-identifiers are masked so that they become indistinguishable within each group of k records. Confidential attributes remain in clear form so that they preserve their analytical utility. In this way, an intruder with access to an external non-anonymous data set that contains the quasi-identifiers in the related data set will be unable to perform an exact re-identification.

Table 1 shows a sample medical data set containing one identifying attribute (SS number), three quasi-identifier attributes (age, zip code and nationality) and one confidential attribute (condition). Table 2 shows a possible sanitized version of the data set after 4-anonymization.

Table 1 Sample input data set

	Identifier	*Quasi-identifiers*			*Confidential*
	SS number	Age	Zip code	Nationality	Condition
1	1234-12-1234	25	23053	Russian	Heart Disease
2	2345-23-2345	26	23068	Catalan	Heart Disease
3	3456-34-3456	21	23068	French	Viral Infection
4	4567-45-4567	27	23053	Italian	Viral Infection
5	5678-56-5678	49	44853	Indian	AIDS
6	6789-67-6789	43	44853	Chinese	Heart Disease
7	7890-78-7890	47	44850	Japanese	Viral Infection
8	8901-89-8901	49	44850	Indian	Viral Infection
9	9012-90-9012	32	33153	Spanish	AIDS
10	0123-12-0123	38	33153	French	AIDS
11	4321-43-4321	34	33168	Greek	AIDS
12	5432-54-5432	35	33168	French	AIDS

Table 2 4-anonymous output

	Identifier	*Quasi-identifiers*			*Confidential*
	SS number	Age	Zip code	Nationality	Condition
1	*	[20-30)	230**	European	Heart Disease
2	*	[20-30)	230**	European	Heart Disease
3	*	[20-30)	230**	European	Viral Infection
4	*	[20-30)	230**	European	Viral Infection
5	*	[40-50)	448**	Asian	AIDS
6	*	[40-50)	448**	Asian	Heart Disease
7	*	[40-50)	448**	Asian	Viral Infection
8	*	[40-50)	448**	Asian	Viral Infection
9	*	[30-40)	331**	European	AIDS
10	*	[30-40)	331**	European	AIDS
11	*	[30-40)	331**	European	AIDS
12	*	[30-40)	331**	European	AIDS

The sanitization method originally proposed to generate a k-anonymous data set was based on generalization and suppression [60].

Generalization reduces the granularity of the information contained in the quasi-identifier attributes, thereby increasing the chance of several records sharing the values of the attributes. A generalization hierarchy should be defined for each attribute. On the other hand, suppression removes records from the original data set that present outlying values. Suppression is usually performed prior to generalization to reduce the amount of generalization required to generate the k-anonymous data set.

An important goal of k-anonymity sanitization proposals is to obtain a protected data set where the information loss is as small as possible. To render k-anonymity practical, a large number of heuristic generalization algorithms have been proposed [59, 7, 46, 47, 4] that reduce the search space or look for sub-optimal solutions.

A different approach towards k-anonymity is based on the microaggregation method discussed in Section 3.1.1. k-Anonymity via microaggregation was introduced in [21]. First, records are clustered so that each cluster contains at least k records and then these records are replaced by a representative value from the cluster to which they belong (typically the centroid record), thus producing a k-anonymous data set. Different heuristics and comparison functions have been proposed to group similar records together, so that the information loss resulting from the replacement by the representative record can be minimized (see [22, 45]).

Despite being one of the most commonly used privacy models, k-anonymity suffers from certain limitations. The most common criticism refers to the lack of protection against attribute disclosure [49, 74, 48, 17]: if all the individuals within a group of k-indistinguishable records share the same value for a confidential attribute, then the intruder will learn that value for all the members of the group without requiring an unequivocal re-identification.

For example, take the 4-anonymized output from Table 2. The last group of four records sharing a combination of quasi-identifier attribute values also shares the confidential attribute value condition (AIDS). In this case, if the intruder can establish that her target respondent's record is within that group (because it is the only group with compatible age, zipcode and nationality), the intruder learns that the target respondent suffers from AIDS without requiring an unequivocal re-identification.

To tackle this problem, some refinements to the basic k-anonymity model have been proposed. First l-diversity [49] requires the presence of l different well-represented values for the confidential attribute in every group of records sharing the same quasi-identifier attribute values. The stricter t-closeness [48] defines a tighter requirement, stating that the distribution of the confidential attributes within any group of records sharing the same quasi-identifier values should be close to (at distance no more than t from) the distribution of the confidential attributes in the whole data set.

5.2 ε-Differential Privacy

Disclosure limitation via k-anonymity is based on guessing the information that is available to potential intruders, that is, which attributes in the data set should be considered as quasi-identifiers. As long as this guessing is accurate, the disclosure limitation method accomplishes its duty, but a privacy breach may happen if more information is available to intruders.

A different approach to anonymization is ε-differential privacy [27]. This approach was designed for sanitization in queryable databases and it makes no assumptions on the intruder's knowledge. The goal is to transform the answers to queries so that the effect of the presence or absence of any single individual record on the returned answers is minimized.

To achieve this goal, the influence of each individual on the query answer needs to be limited. More concretely, the model imposes that the presence or absence of any single individual changes the query answer by at most a factor depending on ε. The smaller ε, the more difficult it is for an intruder to use the query answer to infer the contribution of any specific individual. A formal definition of the ε-differential privacy model follows:

Definition 2. (ε-Differential privacy) A randomized function κ gives ε-differential privacy if, for all data sets X_1, X_2 such that one can be obtained from the other by modifying a single record, and all $S \subset Range(\kappa)$, it holds

$$P(\kappa(X_1) \in S) \leq \exp(\varepsilon) \times P(\kappa(X_2) \in S). \tag{1}$$

Differential privacy was introduced as an interactive mechanism, where the data set is held by a trusted party that provides masked answers to queries made by data users. To do so, the trusted party computes the real response $f(X)$ to the user query f (e.g., the average of an attribute value, the number of records with a specific attribute value, etc.), perturbs the result and sends the output to the user. The usual way to compute the perturbed result is to add a random amount of noise, say $Y(X)$, to the answer $f(X)$ that depends on the ε and the variability of the query response; thus the perturbed response can be obtained as $\kappa(X) = f(X) + Y(X)$.

To generate $Y(X)$ according to ε-differential privacy, a common choice is to use a Laplace distribution with zero mean and $\Delta(f)/\varepsilon$ scale parameter, where:

- ε is the differential privacy parameter;
- $\Delta(f)$ is the L_1-sensitivity of f, that is, the maximum variation of the query function between neighbor data sets, that is, sets differing in at most one record.

Specifically, the density function of the Laplace noise is

$$p(x) = \frac{\varepsilon}{2\Delta(f)} e^{-|x|\varepsilon/\Delta(f)}.$$

Notice that, for fixed ε, the higher the sensitivity $\Delta(f)$ of the query function f, the more Laplace noise is added: indeed, satisfying the ε-differential privacy definition (Definition 2) requires more noise when the query function f can vary strongly between neighbor data sets. Also, for fixed $\Delta(f)$, the smaller ε, the more Laplace noise is added: when ε is very small, Definition 2 almost requires that the probabilities on both sides of Equation (1) be equal, which requires the randomized function $\kappa(\cdot) = f(\cdot) + Y(\cdot)$ to yield very similar results for all pairs of neighbor data sets; adding a lot of noise is a way to achieve this. In the interactive setting, however, the type and number of queries that can be performed over the data are limited, in order to avoid an attacker to reconstruct the original data by performing consecutive queries.

Differential privacy was also extended for the non-interactive setting, that is, for sanitization of microdata sets [8, 25, 38, 10]. Even though a non-interactive data release can be used to answer an arbitrarily large number of queries, in all these proposals, this is obtained at the cost of offering utility guarantees only for a restricted class of queries [8], typically count queries. This contrasts with the general-purpose utility-preserving data release offered by the k-anonymity model.

In fact, it must be said that, while ε-differential privacy offers very high disclosure protection, it causes a huge information loss unless ε is quite high. But taking a high ε somehow seems to contradict the basic requirement of the model, namely that the influence of any single record on the returned output must be small.

6 Research Challenges and Opportunities

Most privacy-preserving methods have been designed to deal with numerical attributes. Numbers are easy to treat because arithmetical functions can be applied on them to perform the comparison and transformation operations required for data anonymization. However, categorical attributes (such as diagnoses, preferences, etc.), which take values from a finite set of categories and for which arithmetical operations do not make sense, are very common in available data sets.

Applying existing data anonymization methods to categorical attributes is not straightforward:

- Several anonymization techniques replace each categorical attribute by as many binary 0-1 attributes as the number of possible attribute categories; such is the case of multiply-imputed synthetic data [57] and data shuffling [54]. This approach soon yields unmanageable data sets.
- PRAM [31] is an anonymization technique designed for nominal attributes. It certainly does not need binary attributes, but it requires as a control parameter a Markov transition matrix, whose size grows quadratically in the number of nominal categories.
- In [70] and [21], extensions of microaggregation algorithms for categorical attributes were proposed: the former paper addressed only categorical ordinal attributes and proposed the median as an aggregation operator; the latter paper also considered categorical attributes using the equality/inequality predicate and proposed the modal value as an aggregation operator for them. However, the modal value is a very coarse aggregation operator which may not even be uniquely defined, especially over a small group of values.

In summary, the above-mentioned methods incur a high complexity for anonymizing categorical data or they are coarse and cause substantial information loss. This is because they treat categorical data as flat categorical values, for which the only possible operator is the binary comparison for equality [21]. This simplistic approach omits data semantics. Overlooking semantics decreases the utility of the anonymized data set since it fails to preserve the meaning of the original data. Semantically-grounded analyses would be desirable to better preserve data utility.

Since categorical attributes are usually words or noun phrases referring to concepts (e.g., disease names) which capture their semantics, and semantics is a human-inherited feature, a semantic analysis requires a human-tailored knowledge base that captures and structures the conceptualization of nominal attributes. For this purpose, structured thesauri, taxonomies, or ontologies [33] can be used.

Recently, some ongoing works have been proposed exploiting available knowledge bases to anonymize categorical data sets. In [23] a knowledge-based numerical mapping for categorical attributes that captures and quantifies their underlying semantics is presented. By means of this mapping, the authors show that it is possible to compute semantically and mathematically coherent *mean*, *variance* and *covariance* functions for nominal data, which can be used to compare and manage categorical data sets in existing anonymization methods. In [51], the notion of semantic similarity [61], that is, the semantic resemblance between categorical attribute values, is extensively exploited to define *comparison*, *aggregation* and *sorting* operators. Those are then used to create semantically-grounded versions of existing anonymization methods based on recoding, microaggregation and resampling. In [6], similar principles are applied to anonymize multi-valued categorical attributes (i.e., set-valued data like query-logs) by defining a set of aggregation functions that allows comparing multi-valued attributes with different cardinalities. Most of the above semantic methods have been applied to microaggregation algorithms.

Future lines of research may apply semantic technologies to other anonymity mechanisms such as those based on noise addition.

Regarding privacy models, it is not uncommon in the data anonymization literature to oppose the “old” k-anonymity model to the “new” differential privacy model, which offers more robust privacy guarantees. However, compared to the general-purpose data publication offered by k-anonymity, which makes no assumptions on the uses of published data, ε-differential privacy offers quite limited utility. Combining the strengths of k-anonymity (flexible utility) and ε-differential privacy (strong privacy) remains a challenge.

The usual approach to release differentially private microdata sets is based on histogram queries [76, 77]; that is, on approximating the data distribution by partitioning the data domain and counting the number of records in each partition set. To prevent the counts from leaking too much information they are computed in a differentially private manner. Apart from the counts, partitioning can also reveal information. One way to prevent partitioning from leaking information consists in using a predefined partition that is independent of the actual data under consideration (e.g., by using a grid [50]). The accuracy of the approximation obtained via histogram queries depends on the number of records contained in each of the histogram bins: the more records, the less relative error. For data sets with sparsely populated regions, using a predefined partition may be problematic.

In a recent approach [67] the authors show that a synergy between k-anonymity and ε-differential privacy can be found in order to achieve more accurate and general-purpose ε-differential privacy. Specifically, they show that the amount of noise required to fulfill ε-differential privacy can be greatly reduced if the query is run on a k-anonymous version of the data set obtained through microaggregation of all attributes (instead of running it on the raw input data). The rationale is that the microaggregation performed to achieve k-anonymity helps reduce the sensitivity of the input versus modifications of individual records; hence, it helps reduce the amount of noise to be added to achieve ε-differential privacy.

In any case, there is still room for improvement because, as it has been empirically shown in [67], the practical utility of general-purpose differentially private data sets is still significantly lower than the one of k-anonymous data sets.

On the legal side, parallel to the development of privacy legislation, anti-discrimination legislation has undergone a remarkable expansion, and in some countries it now prohibits discrimination against protected groups on the grounds of race, color, religion, nationality, sex, marital status, age, and pregnancy, and in a number of settings, like credit and insurance, personnel selection and wages, and access to public services. On the technology side, efforts at fighting discrimination have led to developing anti-discrimination techniques in data mining. Some proposals are oriented to the discovery and measurement of discrimination, while others aim at discrimination-protected data mining (DPDM), that is, at data mining which does not become itself a source of discrimination, due to automated decision making based on discriminatory models extracted from inherently biased data sets.

Another challenge in this area is the relationship between PPDM and DPDM. Is it sufficient to guarantee data privacy while allowing automated discovery of discriminatory profiles/models? In [34, 36, 35], the authors argue that the answer is no. If there is a chance to create a trustworthy technology for knowledge discovery and deployment, it is with a holistic approach which faces both privacy and discrimination threats (risks).

7 Conclusions and Relevant Software

In this chapter, we have reviewed *a priori* and *a posteriori* approaches to disclosure control in database privacy sanitization. The *a posteriori* approach is adopted in the statistical community, which prioritizes publishing analytically valid data and, once the sanitized data have been obtained, measures disclosure risk. The *a priori* approach is followed in the computer science community, which focuses on selecting the maximum tolerable disclosure risk from the outset via a privacy model; after data are protected according to the privacy model, their extant utility is evaluated.

Common to both approaches is the use of sanitization methods, which we have also reviewed for tabular data, queryable databases and microdata, with a special focus on the latter. Finally, we have identified research challenges and opportunities in the area of statistical disclosure control.

Freeware packages that implement the sanitization methods and the risk estimation needed by the *a posteriori* approach include the following:

- The Argus software: τ-Argus for tabular data [41] and μ-Argus for microdata, see [40].
- The *sdc* software: *sdcTable* [65] for tabular data and *sdcMicro* for microdata [64, 69].

Regarding the *a priori* approach, a software package that implements k-anonymity, l-diversity and t-closeness is ARX [5].

Acknowledgments and Disclaimer

This work was partly supported by the Government of Catalonia under grant 2014 SGR 537, by the Spanish Government through project TIN2011-27076-C03-01 “CO-PRIVACY”, and by the European Commission under H2020 projects “CLARUS”. J. Domingo-Ferrer is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this chapter, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

1. Adam, N. R. and Wortmann, J. C.: Security-control for statistical databases: a comparative study. ACM Computing Surveys, **21**(4), 515–556 (1989)
2. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, SIGMOD'00, pp. 439-450. ACM, New York, USA (2000)
3. Aggarwal, C.C., Yu, P.S. (eds.): Privacy-Preserving Data Mining: Models and Algorithms, volume 34 of Adv. in Database Systems. Springer (2008)
4. Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigraphy, R., Thomas, D., Zhu, A.: Anonymizing tables. In: Proceedings of the 10th International Conference on Database Theory, ICDT 2005, pp. 246-258 (2005)
5. ARX - Powerful Data Anonymization (2014). `http://arx.deidentifier.org`
6. Batet, M., Erola, A., Sánchez, D., Castellá-Roca, J.: Utility preserving query log anonymization via semantic microaggregation. Information Sciences **242**, 110–123 (2013)
7. Bayardo, R.J., Agrawal, R.: Data privacy through optimal *k*-anonymization. In: Proceedings of the 21st International Conference on Data Engineering, ICDE'05, pp. 217-228. IEEE Computer Society. Washington, DC, USA (2005)
8. Blum, A., Ligett, K., Roth, A.: A learning theory approach to non-interactive database privacy. In: Proc. of the 40th Annual Symposium on the Theory of Computing-STOC 2008, pp. 609-618 (2008)
9. Chen, B.-C., Kifer, D., LeFevre, K., and Machanavajjhala, A.: Privacy-preserving data publishing. Foundations and Trends in Databases **2**(1-2), 1–167 (2009)
10. Chen, R., Mohammed, N., Fung, B.C.M., Desai B.C., Xiong, L.: Publishing set-valued data via differential privacy. In: 37th Intl. Conference on Very Large Data Bases-VLDB 2011/Proc. of the VLDB Endowment **4**(11), 1087-1098 (2011)
11. Chin, F. Y. and Ozsoyoglu, G.: Auditing and inference control in statistical databases. IEEE Transactions on Software Engineering **SE-8**, 574–582 (1982)
12. Dalenius, T.: The invasion of privacy problem and statistics production. An overview. Statistik Tidskrift **12**, 213-225 (1974)
13. Defays, D. Nanopoulos, P.: Panels of enterprises and confidentiality: the small aggregates method. In: Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys, pages 195-204, Ottawa, Canada (1993)
14. Denning, D. E., Denning, P. J. and Schwartz, M. D.: The tracker: a threat to statistical database security. ACM Transactions on Database Systems **4**(1), 76–96 (1979).
15. Dobra, A., Fienberg, S.E. and Trottini, M.: Assessing the risk of disclosure of confidential categorical data. In: J. Bernardo et al., editors, Bayesian Statistics 7, Proceedings of the Seventh Valencia International Meeting on Bayesian Statistics, pp. 125-139. Oxford University Press (2003)
16. Domingo-Ferrer, J.: A three-dimensional conceptual framework for database privacy. In: Secure Data Management-4th VLDB Workshop SDM'2007, pp. 193-202. Lecture Notes in Computer Science, vol. 4721 (2007)
17. Domingo-Ferrer, J.: A critique of *k*-anonymity and some of its enhancements. In: Proceedings of ARES/PSAI 2008, pp. 990-993. IEEE Computer Society (2008)
18. Domingo-Ferrer, J, Martínez-Ballesté, A.: Mateo-Sanz, J. M. and Sebé, F.: Efficient multivariate data-oriented microaggregation. VLDB Journal **15**, 355–369 (2006)
19. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering **14**(1), 189–201 (2002)
20. Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata. In: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111–134, Amsterdam, North-Holland (2001)
21. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogenous *k*-anonymity through microaggregation. Data Mining and Knowledge Discovery **11**(2), 195–212 (2005)

22. Domingo-Ferrer, J., Sebé, F., Solanas, A.: A polynomial-time approximation to optimal multivariate microaggregation. Computers & Mathematics with Applications **55**(4), 714–732 (2008).
23. Domingo-Ferrer, J., Sánchez, D., Rufian-Torrell, G.: Anonymization of nominal data based on semantic marginality. Information Sciences **242**, 35–48 (2013)
24. Duncan, G. T. and Mukherjee, S.: Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise. Journal of the American Statistical Association **45**, 720–729 (2000)
25. Dwork, C., Naor, M., Reingold, O., Rothblum G.N., Vadhan, S.: On the complexity of differentially private data release: efficient algorithms and hardness results. In: Proc. of the 41st Annual Symposium on the Theory of Computing-STOC 2009, pp. 381-390 (2009)
26. Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., Roehrig., S. F.: Disclosure limitation methods and information loss for tabular data. In: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 135-166, Amsterdam, North-Holland (2001)
27. Dwork, C.: Differential privacy. In: Proceedings of 33rd International Colloquium on Automata, Languages and Programming, ICALP 2006, pp. 1-12. Springer (2006)
28. Fung, B.C.M., Wang, K., Yu, P.S.: Top-down specialization for information and privacy preservation. In: Proceedings of the 21st International Conference on Data Engineering, ICDE'05, pp. 205-216. IEEE Computer Society, Washington, DC, USA (2005)
29. Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S.: Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys **42**(4) (2010)
30. Gopal, R., Garfinkel, R. and Goes, P.: Confidentiality via camouflage: the CVC approach to disclosure limitation when answering queries to databases. Operations Research **50**, 501–516 (2002)
31. Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J. and DeWolf. P.-P.: Post randomisation for statistical disclosure control: Theory and implementation. Research paper no. 9731, Voorburg: Statistics Netherlands (1997)
32. Greenberg, B.: Rank swapping for ordinal data. Washington, DC: U. S. Bureau of the Census (unpublished manuscript) (1987)
33. Guarino, N.: Formal ontology in information systems, In: Proceedings of the 1st International Conference on Formal Ontology in Information Systems, pp. 3-15. Trento, Italy (1998)
34. Hajian, S. and Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Transactions on Knowledge and Data Engineering, **25**(7), 1445–1459 (2013)
35. Hajian, S., Monreale, A., Pedreschi, D., Domingo-Ferrer, J., and Giannotti, F.: Injecting discrimination and privacy awareness into pattern discovery. In: Proceedings of the IEEE 12th International Conference on Data Mining Workshops, pp. 360-369. IEEE Computer Society (2012)
36. Hajian, S. and Domingo-Ferrer, J. and Farràs O.: Generalization-based privacy preservation and discrimination prevention in data publishing and mining. In Data Mining and Knowledge Discovery (to appear)
37. Hansen, S. L. and Mukherjee, S.: Polynomial algorithm for optimal univariate microaggregation. IEEE Transactions on Knowledge and Data Engineering **15**(4), 1043–1044 (2003)
38. Hardt, M., Ligett, K., McSherry, F.: A simple and practical algorithm for differentially private data release. Preprint arXiv:1012.4763v1 (2010)
39. Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A framework for evaluating the utility of data altered to protect confidentiality. The American Statistician **60**(3) (2006)
40. Hundepool, A., Van de Wetering A., Ramaswamy, R., Franconi, L., Polettini, S., Capobianchi, A., DeWolf, P.-P., Domingo-Ferrer, J., Torra, V., Brand, R., and Giessing, S.: μ-ARGUS version 4.2 Software and User's Manual. Statistics Netherlands, Voorburg NL (Dec. 22, 2008), `http://neon.vb.cbs.nl/casc/mu.htm`
41. Hundepool, A., Van de Wetering, A., Ramaswamy, R., de Wolf, P.-P., Giessing, S., Fischetti, M., Salazar, J.-J., Castro, J., and Lowthian, P.: τ-ARGUS v. 3.5 Software and User's Manual, CENEX SDC Project Deliverable (Oct. 2011), `http://neon.vb.cbs.nl/casc/tau.htm`

42. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte-Nordholt, E., Spicer, K., De Wolf, P.P.: Statistical Disclosure Control. Wiley (2012)
43. Kim, J. J.: A method for limiting disclosure in microdata based on random noise and transformation. In: Proceedings of the Section on Survey Research Methods, pp. 303-308, Alexandria VA, American Statistical Association (1986)
44. Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. IEEE Transactions on Knowledge and Data Engineering, **17**(7), 902–911 (2005)
45. Laszlo, M., Mukherjee, S.: Approximation bounds for minimum information loss microaggregation. IEEE Transactions on Knowledge and Data Engineering, **21**(11), 1643–1647 (2009)
46. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: Efficient full-domain *k*-anonymity. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of Data, SIGMOD'05, pp. 49-60. ACM, New York, USA (2005)
47. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional *k*-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering, ICDE'06, pp. 25. IEEE Computer Society, Washington, DC, USA (2006)
48. Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the IEEE International Conference on Data Engineering, ICDE 2007, pp. 106-115 (2007)
49. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-Diversity: privacy beyond k-anonymity. In: Proceedings of the IEEE International Conference on Data Engineering, ICDE 2006, pp. 24 (2006)
50. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: Proceedings of the IEEE International Conference on Data Engineering, ICDE 2008, pp. 277-286 (2008)
51. Martínez, S., Sánchez, D., Valls, A.: A semantic framework to protect the privacy of electronic health records with non-numerical attributes. Journal of Biomedical Informatics **46**(2), 294–303 (2013)
52. Mateo-Sanz, J. M., Domingo-Ferrer, J. and Sebé, F.: Probabilistic information loss measures in confidentiality protection of continuous microdata. Data Mining and Knowledge Discovery **11**(2), 181–193 (2005)
53. Meyerson, A., Williams, R.: On the complexity of optimal *k*-anonymity. In: Proceedings of the 23th ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems, PODS'04, pp. 223-228. ACM, New York, USA (2004)
54. Muralidhar, D., Sarathy, R.: Data shuffling - a new masking approach for numerical data. Management Science **52**(5), 658–670 (2006)
55. Muralidhar, K., Batra, D. and Kirs, P. J.: Accessibility, security and accuracy in statistical databases: the case for the multiplicative fixed data perturbation approach. Management Science, **41**, 1549-1564 (1995)
56. Reiter, J. P.: Satisfying disclosure restrictions with synthetic data sets. Journal of Official Statistics, **18**(4), 531–544 (2002)
57. Rubin, D. B.: Discussion of statistical disclosure limitation. Journal of Official Statistics, **9**(2), 461–468 (1993)
58. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transasctions on Knowledge and Data Engineering **13**(6), 1010–1027 (2001)
59. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. Technical report, SRI International (1998)
60. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information. In: Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of Database Systems, PODS'98, pp. 188. ACM, New York, USA (1998)
61. Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: a new feature-based approach. Expert Systems with Applications **39**(9), 7718–7728 (2012)
62. Schlörer, J.: Identification and retrieval of personal records from a statistical data bank. Methods of Information in Medicine, **14**(1), 7-13 (1975)

63. Schlörer, J.: Disclosure from statistical databases: quantitative aspects of trackers. ACM Transactions on Database Systems, **5**, 467–492 (1980)
64. sdcMicro: Statistical Disclosure Control methods for anonymization of microdata and risk estimation, v. 4.2.0 (Jan. 10, 2014). `http://cran.r-project.org/web/packages/sdcMicro/index.html`
65. sdcTable: Methods for statistical disclosure control in tabular data, v. 0.10.3 (Nov. 4, 2013). `http://cran.r-project.org/package=sdcTable`
66. Sebé, F., Domingo-Ferrer, J., Mateo-Sanz, J. M. and Torra, V.: 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, Springer-Verlag (2002)
67. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., Martínez, S.: Enhancing Data Utility in Differential Privacy via Microaggregation-based k-Anonymity, VLDB Journal (to appear)
68. Sweeney, L.: *k*-Anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, **10**(5), 557–570 (2002)
69. Templ, M.: Statistical disclosure control for microdata using the R-package sdcMicro, Transactions on Data Privacy, **1**(2), 67–85 (2008)
70. Torra, V.: Microaggregation for categorical variables: a median based approach. In: Privacy in Statistical Databases-PSD 2004, LNCS 3050, pp. 162-174, Springer (2004)
71. Traub, J. F., Yemini, Y. Wozniakowski, H.: The statistical security of a statistical database. ACM Transactions on Database Systems, **9**, 672–679 (1984)
72. Willenborg, L. and DeWaal, T.: Elements of Statistical Disclosure Control. Springer-Verlag, New York (2001)
73. Winkler, W. E.: Re-identification methods for evaluating the confidentiality of analytically valid microdata. Research in Official Statistics, **1**(2), 50-69 (1998)
74. Wong, R., Li, J., Fu, A., and Wang, K.: (α, k)-Anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 754-759 (2006)
75. Xiao, X., Tao, Y.: Anatomy: simple and effective privacy preservation. In: Proceedings of the 32nd Intl. Conference on Very Large Data Bases-VLDB 2006, pp. 139-150 (2006)
76. Xiao, Y., Xiong, L., Yuan, C.: Differentially private data release through multidimensional partitioning. In: Proceedings of the 7th VLDB conference on Secure data management, SDM'10, pp. 150-168 (2010)
77. Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G.: Differentially Private Histogram Publication. In: Proceedings of the IEEE International Conference on Data Engineering, ICDE 2012, pp. 32-43 (2012)