Microdata Protection

V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati

Università degli Studi di Milano, 26013 Crema, Italia
{ciriani, decapita, foresti, samarati}@dti.unimi.it

Governmental, public, and private organizations are more and more frequently required to make data available for external release in a selective and secure fashion. Most data are today released in the form of microdata, reporting information on individual respondents. The protection of microdata against improper disclosure is therefore an issue that has become increasingly important and will continue to be so. This has created an increasing demand on organizations to devote resources for adequate protection of microdata.

In this chapter, we first characterize the microdata protection problem (in contrast to macrodata protection), discussing the disclosure risks at which microdata are exposed. We survey the main techniques that have been proposed to protect microdata from improper disclosure by distinguishing them in masking techniques (which protect data by masking or perturbing their values), and synthetic data generation techniques (which protect data by replacing them with plausible, but made up, values). We conclude the chapter with observations on measures for assessing disclosure risk and information loss brought by the application of protection techniques.

1 Introduction

The increased power and interconnectivity of computer systems available today provide the ability of storing and processing large amounts of data, resulting in networked information accessible from anywhere at any time. This information sharing and dissemination process is clearly selective. Indeed, if on the one hand there is a need to disseminate some data, there is on the other hand an equally strong need to protect those data that, for various reasons, should not be disclosed. Consider, for example, the case of a private organization making available various data regarding its business (products, sales, and so on), but at the same time wanting to protect more sensitive information, such as the identity of its customers or plans for future products. As another example, government agencies, when releasing historical data, may

require a sanitization process to "blank out" information considered sensitive, either directly or because of the sensitive information it would allow the recipient to infer. Effective information sharing and dissemination can take place only if the data holder has some assurance that, while releasing information, disclosure of sensitive information is not a risk.

Many techniques have been developed for protecting data released publicly or semi-publicly from improper disclosure. These techniques depend on the method in which such data are released. In the past, data were principally released in tabular form (*macrodata*) and through *statistical databases* [1]. Macrodata are aggregate information (statistics) on users or organizations usually presented as two-dimensional tables while a statistical database is a database whose users may retrieve only aggregate statistics. Macrodata protection techniques are based on the *selective obfuscation of sensitive cells*. Techniques for protecting statistical databases follow two main approaches. The first approach restricts the statistical queries that can be made (e.g., queries that identify a small/large number of tuples) or the data that can be published. The second approach provides protection by returning to the user a modified result. The modification can be enforced directly on the stored data or run time in the process of computing the result to be returned to the user.

However, many situations require today that the specific stored data themselves, called *microdata*, be released. The advantage of releasing microdata instead of specific pre-computed statistics is an increased flexibility and availability of information for the users. To protect the anonymity of the entities, called *respondents*, to which information refers, data holders often remove or encrypt explicit identifiers such as names, addresses, and phone numbers. De-identifying data, however, provides no guarantee of anonymity. Released information often contains other data, such as race, birth date, sex, and ZIP code, that can be linked to publicly available information to reidentify respondents and inferring information that was not intended for disclosure [7, 20, 22]. Disclosure can be categorized as: *identity disclosure*, *attribute disclosure*, and *inferential disclosure*. Identity disclosure occurs when using a combination of identifying attributes (e.g., social security number, name, and address), an individual's identity can be reconstructed. Attribute disclosure occurs when using a combination of indirect identifying attributes, a given attribute value (or restricted set thereof) can be associated with an individual. Inferential disclosure occurs when information can be inferred with high probability from statistical properties of the released data. A first step in protecting the privacy of the *respondents* (individuals, organizations, associations, business establishments, and so on) to which the data refer, consists in releasing data that are generally "sanitized" by removing all explicit identifiers such as names, addresses, and phone numbers. Although apparently anonymous, the de-identified data may contain other data, such as race, birth date, sex, and ZIP code, which uniquely or almost uniquely pertain to specific respondents and make them

stand out from others [22]. By linking these identifying characteristics with publicly available databases (e.g., databases maintained and released by the Department of Motor Vehicles, Health Maintenance Organizations, insurance companies, public offices, commercial organizations, and so on) associating these characteristics to the respondent's identity, the data recipients can determine to which respondent some pieces of released data refer, or restrict their uncertainty to a specific subset of individuals. This has created an increasing demand to devote resources for an adequate protection of sensitive data. As we will see, the microdata protection techniques usually applied to protect sensitive data follow two main strategies. The first strategy consists in reducing the information content of the data provided to the data recipients. The second strategy consists in changing the data before their release in such a way that the information content is maintained as much as possible.

In this chapter, we survey the main microdata disclosure protection techniques. Section 2 provides a brief overview of the difference between macrodata and microdata (this latter being the focus of this chapter). Section 3 provides a characterization of the main microdata disclosure protection techniques. Sections 4 and 5 describe masking techniques and synthetic data generation techniques, respectively. Section 6 provides a discussion on possible measures to evaluate how much the released microdata are protected and, at the same time, informative. Finally, Sect. 7 gives our conclusions.

2 Macrodata Versus Microdata

Data are collected and shared in many different forms. A broad classification can distinguish release in two main classes: *macrodata* and *microdata*. Macrodata consist of data that have been aggregated (e.g., the population of a county is an aggregate of the populations of the cities), while microdata are the base information reporting data on single *respondents*. In this section, we briefly discuss the major characteristics of macrodata versus microdata.

2.1 Macrodata

Macrodata represent estimated values of *statistical characteristics* concerning a given population. A statistical characteristic is a measure that summarizes the values of one or more *properties*/*attributes* (*variables*, in statistical terminology) of respondents. An example of a statistical characteristic can be the average age of people living in each continent. Macrodata can be represented as tables, where each cell of the table is the aggregate value of a quantity over the considered properties. For instance, Figs. 1(a)-(c) illustrates macrodata tables that contain measures computed over properties `Sex` (M,F) and `Disease` (hypertension, obesity, chest pain, and short breath). Macrodata tables can be classified into the following three groups (types of tables).

	Hypertension	Obesity	Chest Pain	Short Breath	Tot
M	1	2	2	1	6
F	1	2	0	2	5
Tot	2	4	2	3	11

(**a**) number of respondents with a disease

	Hypertension	Obesity	Chest Pain	Short Breath	Tot
M	9.1	18.2	18.2	9.1	54.6
F	9.1	18.2	0	18.2	45.4
Tot	18.2	36.4	18.2	27.2	100

(**b**) percentage of respondents with a disease

	Hypertension	Obesity	Chest Pain	Short Breath	Tot
M	2	8.5	23.5	3	37
F	3	30.5	0	5	38.5
Tot	5	39	23.5	8	75.5

(**c**) average number of days spent in the hospital by respondents with a disease

Fig. 1. An example of count (**a**), frequency (**b**), and magnitude (**c**) macrodata tables

- *Count.* Each cell of the table contains the *number of respondents* that have the same value over all attributes of analysis associated with the table. For instance, the table in Fig. 1(a) contains the number of males and females for each given disease.
- *Frequency.* Each cell of the table contains the *percentage of respondents*, evaluated over the total population, that have the same value over all the attributes of analysis associated with the table. For instance, the macrodata table in Fig. 1(b) contains the percentage of males and females for each given disease.
- *Magnitude.* Each cell of the table contains *an aggregate value of a quantity of interest* over all attributes of analysis associated with the table. For instance, the macrodata table in Fig. 1(c) contains the average number of days that males and females have spent in the hospital for each given disease.

Several macrodata protection techniques have been developed to guarantee the *confidentiality* of the data, that is, the assurance that information about single respondents cannot be derived from macrodata. The first step in protecting a macrodata table consists in discovering *sensitive cells*, that is, cells that can be easily associated with a specific respondent. The strategies for discovering and consequently protecting sensitive cells vary depending on the type of macrodata (count and frequency tables versus magnitude tables). For count and frequency tables, the most important strategy used to detect sensitive cells is the *threshold rule*, according to which a cell is sensitive if the number of respondents is less than a given threshold. As an example, consider

the macrodata table in Fig. 1(a) and suppose that the threshold is 2. The first cell and the last cell in the first tuple, and the first cell and the third cell in the second tuple are sensitive because their value is below the threshold. Some of the most important strategies for protecting sensitive cells are *cell suppression*, *rounding*, *roll up categories*, *sampling*, and the *controlled tabular adjustment function* (CTA) [9, 22, 28]. Cell suppression is a well-known technique that consists in protecting sensitive cells by removing their values. These suppressions are called *primary suppressions*. However, a problem can arise when also the marginal totals of the table are published. In this case, even if it is not possible to exactly recalculate the suppressed cell, it can be possible to calculate an interval that contains the suppressed cell. If the size of such an interval is small, then the suppressed cell can be estimated rather precisely. To block such inferences, additional cells may need to be suppressed (*secondary suppression*) to guarantee that the intervals are sufficiently large. To minimize the number of cells to be suppressed, linear programming techniques have been proposed. Such techniques are suitable for small tables, although they are usually not applicable to more complex structures [6, 8, 13, 22, 28]. Rounding consists in choosing a *base number* and in modifying the original value of sensitive cells by rounding it up or down to a near multiple of the base number. Roll up categories reduces the size of the table: instead of releasing a table with N tuples and M columns, a less detailed table (e.g., a table with $N-1$ tuples and $M-1$ columns) is released. Sampling means that the table is obtained with a sample survey rather than a census. The CTA technique is based on the selective adjustment of cell values. In other words, the value of sensitive cells is replaced by a *safe value*, that is, a value that satisfies the rule chosen to detect sensitive cells, and then uses linear programming to adjust the values of the nonsensitive cells to restore the additivity property.

For magnitude macrodata, there are many rules that can be used to detect sensitive cells. For instance, the *(n,k)-rule* states that a cell is sensitive if less than n respondents contribute to more than $k\%$ of the total cell value. As an example, consider the macrodata table in Fig. 1(c) and suppose to apply the (1,50)-rule. A cell is therefore sensitive if one respondent contributes to more than 50% of its value. The first cell and the last cell in the first tuple as well as the first cell in the second tuple are sensitive because, according to the macrodata table in Fig. 1(a), there is only one male and one female with hypertension and one male with short breath and therefore their contribution to these cells is 100%. Other similar rules are the *p-percentage* rule and the *pq-rule* [22]. The *p-percentage* states that a cell is sensitive if the total value t of the cell minus the largest reported value v_1 minus the second largest reported value v_2 is less than $(p/100)\cdot v_1$. Intuitively, this rule means that a user can estimate the reported value of some respondent too accurately. In the *pq-rule*, q represents how accurately respondents can estimate another respondent's value ($p < q < 100$). Note that some of the techniques used for protecting count and frequency tables can also be used for protecting magnitude tables (e.g., cell suppression, roll up categories, and CTA).

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Disease	DH	Chol	Temp
		Asian	64/09/27	F	94139	Divorced	Hypertension	3	260	35.2
		Asian	64/09/30	F	94139	Divorced	Obesity	1	170	37.7
		Asian	64/04/18	M	94139	Married	Chest pain	40	200	38.1
		Asian	64/04/15	M	94139	Married	Obesity	7	280	37.4
		Black	63/03/13	M	94138	Married	Hypertension	2	190	35.3
		Black	63/03/18	M	94138	Married	Short breath	3	185	38.2
		Black	64/09/13	F	94141	Married	Short breath	5	200	36.5
		Black	64/09/07	F	94141	Married	Obesity	60	290	39.8
		White	61/05/14	M	94138	Single	Chest pain	7	170	37.6
		White	61/05/08	M	94138	Single	Obesity	10	300	40.1
		White	61/09/15	F	94142	Widow	Short breath	5	200	36.9

Fig. 2. An example of de-identified medical microdata table

2.2 Microdata

Microdata contain a set of attributes relating to single respondents in a sample or in a population. Microdata can be represented as tables composed of tuples (records) with values from a set of attributes. Figure 2 illustrates an example of microdata table with 11 tuples and with attributes `SSN` (social security number), `Name`, `Race`, `DoB` (date of birth), `Sex`, `ZIP` code, `MarStat` (marital status), `Disease`, `DH` (days in hospital), `Chol` (cholesterol), and `Temp` (temperature).[1] In the remainder of this chapter, we refer our examples to this microdata table.

The attributes in an initial microdata table are usually classified as follows.

- *Identifiers.* Attributes that uniquely identify a microdata respondent. For instance, attribute `SSN` uniquely identifies the person with which is associated.
- *Quasi-identifiers.* Attributes that, in combination, can be linked with external information to reidentify, all or some of the respondents to whom information refers or reduce the uncertainty over their identities. For instance, attributes `DoB`, `ZIP`, and `Sex` are quasi-identifiers: they can be linked to external public information to reveal the name and address of the corresponding respondents or to reduce the uncertainty to a specific set of respondents.
- *Confidential attributes.* Attributes of the microdata table that contain sensitive information. For instance, attribute `Disease` can be considered sensitive.

[1] Note that in this table data have been de-identified by suppressing names and social security numbers so not to directly disclose the identities of the respondents to whom the data refer (see Sect. 3 for more details).

- *Non confidential attributes.* Attributes that the respondents do not consider sensitive and whose release do not cause disclosure. For instance, attribute `Race` can be considered non confidential.

In general, protecting microdata from *reidentification* of respondents is a more difficult task than protecting macrodata from disclosure because each tuple of the microdata table contains actual data of single respondents. In the remainder of this chapter, we will focus on the microdata disclosure protection techniques (data protection techniques, for short).

3 Classification of Microdata Disclosure Protection Techniques

Disclosure control of microdata is an important practical issue in the private as well as in the public and governmental sectors. Microdata protection techniques have two apparently contrasting objectives. On the one side, they should avoid *reidentification* that happens whenever the information of a respondent appearing in a microdata table is identified, that is, is associated with the identity of the corresponding respondent. On the other side, the application of such techniques should preserve the *key statistical properties* of the original data that data recipients have indicated as important. More precisely, given a microdata table T, a data protection technique should transform this original table into another microdata table T' in a way that: *i)* the risk that a malicious user can use T' to determine confidential information or to identify a respondent should be low; *ii)* the statistical analysis over T and over T' should produce similar results.

In general, the following main factors contribute to disclosure risks [22].

- The existence of high visibility tuples (i.e., tuples with unique characteristics such as a high income).
- The possibility of matching the microdata table with external information. For instance, suppose that a public voter list includes names, social security numbers, sex, birth dates, and addresses. Attributes `DoB`, `ZIP`, and `Sex` in Fig. 2 can then be linked to the voter list to reveal the names and social security numbers.
- The existence of a high number of common attributes between the microdata table and the external sources, which may increase the possibility of linking or make it more precise.

By contrast, the main factors that decrease the disclosure risks can be summarized as follows.

- A microdata table often contains a subset of the whole population. This implies that the information of a specific respondent, which a malicious user may want to know, may not be included in the microdata table.

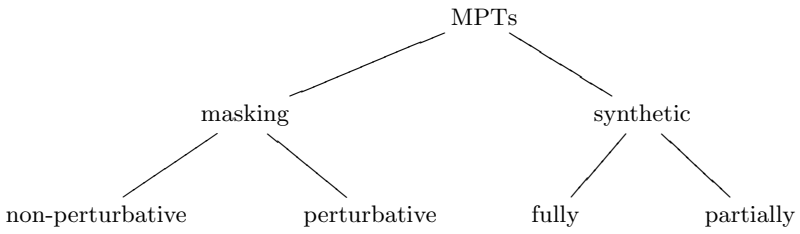

Fig. 3. Classification of microdata protection techniques (MPTs)

- The information specified in microdata tables released to the public are not always up-to-date (often at least one or two-year old). This means that the values of the attributes of the corresponding respondents may have been changed in the meanwhile. In addition, the age of the external sources of information used for linking may be different from the age of the information contained in the microdata table.
- A microdata table and the external sources of information naturally contain noise that decreases the ability to link the information.
- A microdata table and the external sources of information can contain data expressed in different forms thus decreasing the ability to link information.

In general, to limit the disclosure risk of a microdata table it is first necessary to suppress explicit and implicit identifiers (e.g., `SSN` and `Name` in Fig. 2). This process is also known as *de-identification*. Note that de-identification does not necessarily make a tuple anonymous [44, 47], as it may be possible to reidentify the tuple using external information. For instance, consider the microdata in Fig. 2, where all the identifiers have been removed and suppose to link the information in this table with the voter list that is a public non-anonymous dataset. The microdata table contains, for the last tuple, a unique combination of values for attributes `DoB`, `Sex`, `ZIP`, and `MarStat`. This combination, if unique in the voter list as well, uniquely identifies the corresponding tuple in the microdata table as pertaining to a specific respondent. In addition, it is necessary to limit geographical details as well as the number of attributes in microdata tables to reduce the probability of reidentification of respondents.

Several microdata disclosure protection techniques have been proposed in the literature. Basically, these techniques are based on the principle that reidentification can be counteracted by reducing the amount of released information, masking the data (e.g., by not releasing or by perturbing their values), or by releasing plausible but made up values instead of the real ones. According to this principle, the microdata protection techniques can be classified into two main categories: *masking techniques*, and *synthetic data generation techniques* (see Fig. 3).

- *Masking techniques.* The original data are transformed to produce new data that are valid for statistical analysis and such that they preserve the confidentiality of respondents. Masking techniques can be classified as:
  - *non-perturbative*, the original data are not modified, but some data are suppressed and/or some details are removed;
  - *perturbative*, the original data are modified.
- *Synthetic data generation techniques.* The original set of tuples in a microdata table is replaced with a new set of tuples generated in such a way to preserve the key statistical properties of the original data. The generation process is usually based on a statistical model and the key statistical properties that are not included in the model will not be necessarily respected by the synthetic data. Since the released microdata table contains synthetic data, the reidentification risks is reduced. Note that the released microdata table can be entirely synthetic (i.e., *fully* synthetic) or mixed with the original data (i.e., *partially* synthetic).

Another important feature of microdata protection techniques is that they can operate on different data types. In particular, data types can be categorized as follows.

- *Continuous.* An attribute is said to be continuous if it is numerical and arithmetic operations are defined on it. For instance, attributes date of birth and temperature are continuous attributes.
- *Categorical.* An attribute is said to be categorical if it can assume a limited and specified set of values and arithmetic operations do not have sense on it. Note that an order relationship can be defined over a categorical attribute. For instance, attributes marital status and race are categorical attributes.

In the following, we describe the principal microdata protection techniques indicating also whether they are applicable to continuous data, categorical data, or both.

4 Masking Techniques

We present some of the most popular non-perturbative and perturbative masking techniques. Figure 4 and Fig. 5 lists the techniques indicating whether they are applicable (yes) or not (no) to continuous or categorical data types.

4.1 Non-Perturbative Techniques

Non-perturbative techniques produce protected microdata by eliminating details from the original microdata. We discuss these techniques illustrating as examples their application to the protection of the table in Fig. 2. The result of the application of the techniques is illustrated in Fig. 7.

Technique	Continuous	Categorical
Sampling	yes	yes
Local suppression	yes	yes
Global recoding	yes	yes
Top-coding	yes	yes
Bottom-coding	yes	yes
Generalization	yes	yes

Fig. 4. Applicability of non-perturbative masking techniques to the different data types

Technique	Continuous	Categorical
Resampling	yes	no
Lossy compression	yes	no
Rounding	yes	no
PRAM	no	yes
MASSC	no	yes
Random noise	yes	yes
Swapping	yes	yes
Rank swapping	yes	yes
Micro-aggregation	yes	yes

Fig. 5. Applicability of perturbative masking techniques to the different data types

Sampling [22]

The protected microdata table is obtained as a sample of the original microdata table. In other words, the protected microdata table includes only the data (tuples) of a sample of the whole population. Since there is an uncertainty about whether or not a specific respondent is in the sample, the risk of reidentification in the released microdata decreases. For instance, we can decide to publish only the even tuples of the original microdata table. This technique operates on categorical attributes only.

Local Suppression [5, 44]

It suppresses the value of an attribute (i.e., it i replaces it with a missing value) thus limiting the possibilities of analysis. Basically, this technique blanks out some attribute values (sensitive cells) that are likely to contribute significantly to the disclosure risk of the tuple involved. For instance, we can suppress attributes `ZIP` and `MarStat` in the last tuple.

Global Recoding (or Recoding into Intervals) [17, 18, 49]

The domain of an attribute is partitioned into disjoint intervals, usually of the same width, and each interval is associated with a label. The protected microdata table is obtained by replacing the values of the attribute with the label associated with the corresponding interval. Intuitively, global recoding

decreases the details in the microdata table and therefore it should reduce the risk of reidentification. For instance, suppose that the values of attribute `Temp` are partitioned into three intervals: [35.0,36.9] with label *no fever* (nf); [37.0,38.9] with label *fever* (f); and [39.0,40.9] with label *high fever* (hf). The value in the first tuple is then replaced by label "nf"; the second, third, and fourth value are replaced by label "f"; and so on. Note that if the original domain of the considered attribute is continuous, it becomes discrete after the application of this technique.

Two particular global recoding techniques are the top-coding and the bottom-coding described in the following.

Top-Coding [17, 18]

It is based on the definition of an upper limit, called *top-code*, for each attribute to be protected. Any value greater than this value is replaced with the top-code. For instance, consider attribute `DH` and suppose that the top-code is 30. In this case, rather than publishing the third and eighth tuple showing a number of days in a hospital equal to 40 and 60, respectively, these two tuples may only show that the number of days is "> 30". The idea is that long periods in the hospital can be easily associated with specific respondents. Top-coding can be applied to categorical attributes that can be linearly ordered as well as to continuous attributes.

Bottom-Coding [17, 18]

It is similar to top-coding. It consists in defining a lower limit, called *bottom-code*, for each attribute to be protected. Therefore, any value lower than this limit is not published and is replaced with the bottom-code. For instance, consider attribute `Chol` and suppose that the bottom-code is 195. The second, fifth, sixth, and ninth tuples are modified in such a way that the value published for attribute `Chol` is "< 195". Basically, since low cholesterol values for people having obesity or hypertension problems are uncommon, they have to be obfuscated to avoid a possible reidentification. Like for top-coding, this technique can be applied to categorical attributes that can be linearly ordered as well as to continuous attributes.

Generalization [44]

It consists in representing the values of a given attribute by using more general values. This technique is based on the definition of a *generalization hierarchy*, where the most general value is at the root of the hierarchy and the leaves correspond to the most specific values. A generalization process therefore proceeds by replacing the values represented by the leaf nodes with one of their ancestor nodes at a higher level. Different generalized microdata tables can be built, depending on the number of generalization steps applied on the considered attribute. For instance, consider attribute `ZIP` and the corresponding

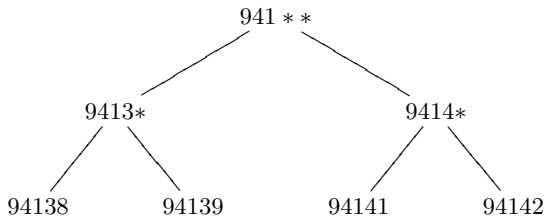

Fig. 6. Generalization hierarchy for attribute ZIP

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Disease	DH	Chol	Temp
		Asian	64/09/27	F	9413*	Divorced	Hypertension	3	260	nf
		Asian	64/09/30	F	9413*	Divorced	Obesity	1	<195	f
		Asian	64/04/18	M	9413*	Married	Chest pain	>30	200	f
		Asian	64/04/15	M	9413*	Married	Obesity	7	280	f
		Black	63/03/13	M	9413*	Married	Hypertension	2	<195	nf
		Black	63/03/18	M	9413*	Married	Short breath	3	<195	f
		Black	64/09/13	F	9414*	Married	Short breath	5	200	nf
		Black	64/09/07	F	9414*	Married	Obesity	>30	290	hf
		White	61/05/14	M	9413*	Single	Chest pain	7	<195	f
		White	61/05/08	M	9413*	Single	Obesity	10	300	hf
		White	61/09/15	F			Short breath	5	200	nf

Fig. 7. Microdata table of Fig. 2 obtained by applying the non-perturbative techniques listed in Fig. 4

generalization hierarchy in Fig. 6. Each generalization step consists in suppressing the least significant digit in the ZIP code. In this case, if we choose to apply one generalization step, values 94138, 94139, 94141, and 94142 are generalized to 9413* and 9414*. This technique is applicable on both continuous and categorical attributes. Note also that the global recoding technique can be seen as a particular case of generalization.

Figure 7 contains the protected microdata table obtained from the microdata table of Fig. 2 by applying, as discussed, the top-coding technique on attribute DH, the bottom-coding technique on attribute Chol, the global recoding technique on attribute Temp, the local suppression technique on the last tuple, and one generalization step on attribute ZIP.

4.2 Perturbative Techniques

With perturbative techniques, the microdata table is modified for publication. Modifications can make unique combinations of values in the original table disappear as well as introduce new combinations.

S_1	S_2	S_3	S_4
260	220	170	210
170	280	290	190
200	210	220	230
280	310	270	200
190	290	185	185
185	180	300	260
200	285	250	220
290	265	260	290
170	150	190	230
300	270	270	310
200	298	200	170

(**a**) Initial samples

S_1	S_2	S_3	S_4	Average
170	150	170	170	165
170	180	185	185	180
185	190	190	190	188.75
190	210	200	200	200
200	220	220	210	212.5
200	265	250	220	233.75
200	270	260	230	240
260	280	270	230	260
280	285	270	260	273.75
290	290	290	290	290
300	310	300	310	305

(**b**) Ordered samples

Original value (S_1)	Released value
260	260
170	165
200	212.5
280	273.75
190	200
185	188.75
200	233.75
290	290
170	180
300	305
200	240

(**c**) Released data

Fig. 8. An example of resampling over attribute `Chol`

Resampling [14, 17]

This technique consists in replacing the values of a sensitive continuous attribute with the average value computed over a given number of samples taken from the original population. More precisely, let N be the number of tuples in a microdata table and $S_1, \ldots, S_t$ be t samples of size N. Each sample is independently ranked (using the same ranking criterion for all samples) and the average of the j-th ranked values in $S_1, \ldots, S_t$ is computed. The obtained averages are then re-ordered by taking into consideration the order of original values; the first average value then replaces the first original value, the second average replaces the second original value, and so on. For instance, suppose that attribute `Chol` is protected by applying this technique and that we choose $t = 4$ samples. Figure 8 illustrates the different steps in protecting attribute `Chol`. Note that the first sample (column S_1) corresponds to the `Chol` values in the original microdata table.

Lossy Compression [16, 17]

It is a recent technique that exploits image compression algorithms. A continuous microdata table is interpreted as an image, and a lossy compression algorithm (e.g., jpeg) is applied on it. The result is the protected microdata table. Depending on the lossy compression algorithm used, it is necessary to detect an appropriate correspondence between attribute ranges and color scales. This technique can only be applied on continuous data and the compression rate coincides with the obfuscation parameter: the higher the compression rate, the more protected the data.

Rounding [13]

It is similar to the homonymous technique used for protecting macrodata and is applicable only on continuous attributes. It replaces the original values of the considered attribute with rounded values. Rounded values are chosen among a set of *rounding points* p_i each of which defines a *rounding set*. As an example, the rounding points could be chosen as multiples of a base value b, that is, $p_{i+1} - p_i = b$, and the rounding sets could be defined as $[p_i - b/2, p_i + b/2)$, $i = 2 \ldots r-1$, $[0, p_1 + b/2)$, and $[p_r - b/2, X_{max}]$ (X_{max} is the largest possible value for attribute X) for p_1 and p_r, respectively. An original value v of X is then replaced by the rounding point corresponding to the rounding set where v lies. For instance, consider attribute `Temp`, $b = 1$, and the rounding points 36, 37, 38, and 39. The corresponding rounding sets are: $[0, 36.5)$; $[36.5, 37.5)$; $[37.5, 38.5)$; and $[38.5, 40.1]$, respectively. The value 35.2 in the first tuple is replaced by 36, the second value 37.7 is replaced by 38, and so on. Note that this rounding technique is usually performed on one attribute at a time (univariate rounding); although multivariate rounding operating on whole tuples is also possible [53].

PRAM (Post RAndomized Method) [18, 29, 35]

It consists in replacing the categorical value for one or more attributes in each tuple with another categorical value based on some probability mechanism. For instance, a *Markov matrix* $P = [p_{ij}]$ (i.e., a real $n \times n$ matrix, where all elements p_{ij} are greater than or equal to 0 and $\sum_{j=1}^{n} p_{ij} = 1, i = 1, \ldots, n$) can contain the probability to replace categories in the original microdata table with other categories. In other words, p_{ij} is the probability that category c_i in the original microdata is substituted by category c_j in the protected microdata.

MASSC (Micro-Agglomeration, Substitution, Sub-sampling and Calibration) [46]

It is a technique that consists of four steps that work as follows.

- *Micro-agglomeration.* Tuples in the original microdata table are partitioned into different groups characterized by a similar risk of disclosure. Each group is formed on the basis of their quasi-identifier. Intuitively, tuples with rare combinations of values for quasi-identifier attributes are at a higher risk and should be in the same group.
- *Substitution.* Original data are perturbed by following an optimal probabilistic strategy.
- *Sub-sampling.* Some cells or whole tuples are suppressed according to an optimal probabilistic subsampling strategy.
- Optimal *calibration.* The sampling weights, used in the previous step, are calibrated to preserve a certain statistical property. In particular, this calibration involves attributes that are to be used by data recipients for surveys.

This technique has been originally proposed for reducing the disclosure risk due to the linkage of categorical attributes with external sources. It is therefore not suitable for tables that contain continuous attributes.

Random Noise [22]

It perturbs a sensitive attribute by adding or by multiplying it with a random variable with a given distribution. The *additive noise* [3, 17] is more frequently used than *multiplicative noise* and can be formally expressed as follows. Let X_j be the j-th column of the original microdata table corresponding to a sensitive attribute and suppose that there are N tuples. Each value $x_{ij}, i = 1, \ldots, N$, is replaced by $x_{ij} + \varepsilon_{ij}$, where ε_j is a vector of normally distributed errors drawn from a random variable with mean equals to zero and, in general, with a variance that is proportional to those of the original attributes (i.e., $\varepsilon_j \approx N(0, \sigma^2_{\varepsilon_j})$ and $\sigma^2_{\varepsilon_j} = \alpha \cdot \sigma^2_{X_j}$, where α is the proportional coefficient). This method, also called *uncorrelated additive noise*, preserves the mean and the covariance of the original data while variances and correlation coefficients are not preserved. *Correlated additive noise* is another technique that preserves the mean and can allow preservation of correlation coefficients. The difference with the previous method is that the co-variance matrix of the errors is proportional to the co-variance matrix of the original data.

In general, masking by correlated additive noise produces masked data with higher analytical validity than masking by uncorrelated additive noise. However, additive noise is seldomly used by itself because of the low level of protection it provides [50, 51]. Rather, it is often combined with *linear* (for continuous attributes [34]) or *non linear* (for categorical attributes [48]) transformations. This means that the microdata obtained after the application

Original value	Error	Released value
3	+2	5
1	+1	2
40	−10	30
7	+3	10
2	+5	7
3	+8	11
5	+4	9
60	−11	49
7	−2	5
10	−3	7
5	+3	8

Fig. 9. An example of uncorrelated additive noise over attribute DH

of the additive noise technique are then linearly (or non linearly) transformed before release. Such an additional transformation must preserve mean and covariance. Note that the parameters used in the linear transformation should not be revealed because their knowledge allows the inversion of the function used: the released microdata would have the same degree of protection as if they were protected only by additive noise. The additive noise technique is suitable to protect continuous data since no assumption on the possible values of sensitive attributes can be made, and because no exact matching with external sources of information is possible. Additive noise is usually not suitable to protect categorical data.

To illustrate, consider attribute DH and suppose to protect such an attribute by applying uncorrelated additive noise. We first need to compute the mean and the variance of the original attribute: $\sigma^2_{\texttt{DH}} = 328.36$ and $\mu_{\texttt{DH}} = 13$. We then set α to 0.1 and obtain that $\sigma^2_{\varepsilon_j} \cong 33$. We now draw the error vector (ε) from the normal distribution $N = (0, 33)$, that is, a distribution with mean equals to zero and variance equals to 33. Figure 9 illustrates the original values, the error, and the released values. Note that the mean of the released data is equal to the mean of the original one, while the variance is not preserved.

Swapping [10, 13, 33]

It consists in modifying a subset of the tuples in a microdata table by swapping the values of a set of sensitive attributes, called *swapped attributes*, between selected pairs of tuples (the pairs are selected according to a well-defined criteria). Intuitively, this technique reduces the risk of reidentification because it introduces uncertainty about the true value of a respondent's data. As an example, suppose that the swapped attributes are Disease, DH, Chol, and Temp and that the selected pairs of tuples must have a matching on attributes Sex and MarStat. Figure 10 illustrates the table obtained by swapping tuple t_3 with t_5, t_7 with t_8, and t_9 with t_{10} (the swapped values are reported in

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Disease	DH	Chol	Temp
		Asian	64/09/27	F	94139	Divorced	Hypertension	3	260	35.2
		Asian	64/09/30	F	94139	Divorced	Obesity	1	170	37.7
		Asian	64/04/18	M	94139	Married	*Hypertension*	*2*	*190*	*35.3*
		Asian	64/04/15	M	94139	Married	Obesity	7	280	37.4
		Black	63/03/13	M	94138	Married	*Chest pain*	*40*	*200*	*38.1*
		Black	63/03/18	M	94138	Married	Short breath	3	185	38.2
		Black	64/09/13	F	94141	Married	*Obesity*	*60*	*290*	*39.8*
		Black	64/09/07	F	94141	Married	*Short breath*	*5*	*200*	*36.5*
		White	61/05/14	M	94138	Single	*Obesity*	*10*	*300*	*40.1*
		White	61/05/08	M	94138	Single	*Chest pain*	*7*	*170*	*37.6*
		White	61/09/15	F	94142	Widow	Short breath	5	200	36.9

Fig. 10. Microdata table of Fig. 2 protected through swapping over attributes `Disease`, `DH`, `Chol`, and `Temp`

italic). Although this technique is easy to apply, in general it has the disadvantage of not preserving statistical properties on subdomains. The original technique has been presented for categorical attributes only. However, in [42] data swapping has been extended to continuous data.

Rank Swapping [17, 30, 55]

It is a variation of swapping that can be applied to continuous and categorical attributes with an order relationship. Basically, the values of an attribute X are ranked in ascending order, and each value is swapped with another value in such a way that the swapped tuples are guaranteed to be within a specified *rank-distance* of one another (i.e., the swapped values should be in a range of p% of the total range). For instance, suppose to apply this technique on attribute `Temp` and assume $p = 10\%$. The range of this attribute is [35.2,40.1] and therefore the difference between the swapped values should be equal to o lesser than $((40.1-35.2)\cdot 10)/100 = 0.49$. We can then, for instance, swap the value in the first tuple with the value in the fifth tuple; the value in the second tuple with the value in the fourth tuple; and so on. Figure 11 illustrates the resulting microdata table.

Micro-Aggregation (or Blurring) [12, 17]

It consists in grouping individual tuples into small aggregates of a fixed dimension k: the average over each aggregate is published instead of individual values. Groups are formed by using maximal similarity criteria. Although different functions can be defined to measure the similarity, it can be difficult to find an optimal grouping solution [39] and recently some heuristic algorithms have been proposed to maximize similarity [12].

There are different variations of micro-aggregation. For instance, the average can substitute the original value only for a tuple in the group or for

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Disease	DH	Chol	Temp
		Asian	64/09/27	F	94139	Divorced	Hypertension	3	*225*	*35.3*
		Asian	64/09/30	F	94139	Divorced	Obesity	1	*260*	*37.4*
		Asian	64/04/18	M	94139	Married	Chest pain	40	*185*	*38.2*
		Asian	64/04/15	M	94139	Married	Obesity	7	*260*	*37.7*
		Black	63/03/13	M	94138	Married	Hypertension	2	*225*	*35.2*
		Black	63/03/18	M	94138	Married	Short breath	3	*195*	*38.1*
		Black	64/09/13	F	94141	Married	Short breath	5	*195*	*36.9*
		Black	64/09/07	F	94141	Married	Obesity	60	*260*	*40.1*
		White	61/05/14	M	94138	Single	Chest pain	7	*185*	37.6
		White	61/05/08	M	94138	Single	Obesity	10	*260*	*39.8*
		White	61/09/15	F	94142	Widow	Short breath	5	*195*	*36.5*

Fig. 11. Microdata table of Fig. 2 protected through rank-swapping over attribute `Temp` and micro-aggregation over attribute `Chol`

all of them; different attributes can be protected through micro-aggregation using the same or different grouping; and the group size may be fixed or variable with a fixed minimum size. As an example, consider attribute `Chol`, and suppose to group tuples according to their value over attribute `Disease` and that the size of each group is variable (while the minimum size is set to 2). The groups are: $\{t_1, t_5\}$, $\{t_2, t_4, t_8, t_{10}\}$, $\{t_3, t_9\}$, and $\{t_6, t_7, t_{11}\}$. Figure 11 illustrates the resulting table. Note that micro-aggregation was first proposed only to protect continuous attributes, but recently some variants for categorical data have been studied. These solutions are based on existing clustering and aggregation definitions such as the c-means [52].

5 Synthetic Data Generation Techniques

The generation of synthetic data is an alternative option for protecting microdata. The basic principle on which such techniques are based is that since the statistical content of the data is not related with the information provided by each respondent, a model well representing the data could in principle replace the data themselves [4]. An important requirement for the generation of synthetic data, which makes the generation process a complicate issue, is that the synthetic and original data should present the same quality of statistical analysis. The main advantage of this class of techniques is that the released synthetic data are not referred to any respondent and therefore their release cannot lead to reidentification. These techniques allow the data holders to pose their attention on the quality of the released data instead of posing attention on the reidentification problem.

In the remainder of this section we describe the main synthetic data generation techniques. Figure 12 and Fig. 13 lists the techniques indicating whether they are applicable (yes) or not (no) to continuous or categorical data types.

Technique	Continuous	Categorical
Bootstrap	yes	no
Cholesky decomposition	yes	no
Multiple imputation	yes	yes
Maximum entropy	yes	yes
Latin Hypercube Sampling	yes	yes

Fig. 12. Applicability of fully synthetic techniques to the different data types

Technique	Continuous	Categorical
IPSO	yes	no
Hybrid masking	yes	no
Random response	no	yes
Blank and impute	yes	yes
SMIKe	yes	yes
Multiply imputed partially synthetic dataset	yes	yes

Fig. 13. Applicability of partially synthetic techniques to the different data types

The techniques are divided into two categories: *fully synthetic* techniques and *partially synthetic* techniques. The first category contains techniques that generate a completely new set of data, while the techniques in the second category merge the original data with synthetic data.

5.1 Fully Synthetic Techniques

We describe some significant *fully synthetic* generation techniques that release only synthetic data.

Bootstrap [24]

Given a microdata table with p attributes, this technique first computes the corresponding *p-variate cumulative distribution function* F. A p-variate cumulative distribution function is a function that completely describes the probability distribution of a set of p real-valued random variables (e.g., the *Gaussian function*). The parameters that characterize F can be determined by using the bootstrap technique. Basically, bootstrap estimates each parameter of the population by using a set of synthetic samples, obtained from the original sample through a resampling with replacement. Once the parameters have been estimated, the corresponding function F on the population is modified to obtain a similar function F'. This new function is then sampled to obtain a set of synthetic data. The modifications on function F should however preserve the statistical properties of the original data. Note that this technique can be applied only on continuous attributes because it is not possible to compute function F on categorical data.

Cholesky Decomposition [38]

This technique, which operates only on continuous attributes and in time linear in the sample size, preserves mean, variance, and co-variance of the original data and is based on the *Cholesky matrix decomposition* method. Given a microdata table T, that can be represented as a matrix of $N \times M$ elements, where rows are tuples and columns are attributes, it is first necessary to compute the co-variance matrix C over T. The next step consists in generating a random matrix, denoted as R, of size $N \times M$, such that the identity matrix I is the co-variance matrix. Then, the Cholesky decomposition U of C is determined, where $C = U^t \times U$. The synthetic microdata matrix is then computed as $R \cdot U$, and it has exactly the same co-variance matrix as T.

Multiple Imputation [41, 43]

Given a microdata table with N tuples (i.e., a sample of N respondents) obtained from a much larger population of M individuals, attributes in the table are partitioned into three sets: a set A of *background* attributes (e.g., age, address), a set B of *non confidential* attributes, and a set C of *confidential* attributes. The values of attributes in A are known for the whole population while the values of attributes in B and C are known for the sample only. The multiple imputation method consists of the following three steps.

- Starting from the sample, a *multiple imputed population* of size M is constructed.[2] Such a population contains the N tuples of the microdata table plus p matrices of (B,C) data (p is the multiply-imputed parameter) for the $M - N$ individuals that do not belong to the sample.
- Starting from the known values in A, a set of couples (B,C) is predicted. In this way, the whole population has a value (original or imputed) for A, B, and C. Couples (B,C) are generated using a prediction model.
- A sample of N tuples on the multiply-imputed population is then drawn. This step is repeated p times to create p replicates of (B,C) values. As a result, we obtain p multiply-imputed synthetic datasets. To avoid the inclusion of the original sample (i.e., the N tuples in the microdata table), the samples can be drawn from the multiply-imputed population excluding the N original tuples from it.

  This technique operates on both continuous and categorical attributes.

[2] Imputation is the practice of filling in missing data with plausible values. Multiple imputation means that the missing values are replaced with p simulated values, where p usually varies between 3 and 10.

Maximum Entropy [4, 40]

It is based on the consideration that by knowing the exact distribution of actual data, it is possible to generate an optimal sample by correctly tuning the parameters of the distribution function and randomly drawing tuples from it. However, the main problem is that since the exact distribution function is not known, it has to be estimated on the basis of the original sample. Therefore, we need to detect the family of distribution functions to which the original data distribution belongs. Then, a specific function is chosen from the family as the one having the *maximum entropy distribution* (such a function exists and is unique). Entropy is defined as the measure of data conformity to a set of constraints. The main task for the data holder is to find out a suitable set of constraints. Typically, constraints are defined on the value of certain statistics, that is, the release synthetic data preserve on average some selected sample characteristics. This technique operates on both continuous and categorical attributes.

Latin Hypercube Sampling (LHS) [25, 32]

It produces a synthetic microdata sample reproducing the univariate (i.e., related to a single attribute) statistics of interest, which usually are mean and variance of the values of an attribute. This technique can be applied to a single attribute or to a set of uncorrelated attributes. Recent refinements of this technique reproduce, on synthetic data, the rank correlation structure of the original sample. To this aim, if the sample is composed of a number of attributes or if there are different observations on the same attributes for the same sample of respondents, it is necessary to iteratively refine the rank correlation matrix used to minimize the difference between the rank correlation of the original and the synthetic data. If the rank correlation matrix is well tuned, the rank correlation between subsets of attributes is better preserved. The main drawback is that it is computationally expensive to produce the synthetic sample and such a complexity depends on the number of statistics to preserve in the synthetic sample and on their value. The technique can be used on both continuous and categorical data.

5.2 Partially Synthetic Techniques

Since it may be difficult to generate plausible synthetic data for all attributes, techniques that generate partially synthetic datasets have also been considered. Basically, these techniques produce a mix of synthetic and original values. We now describe the main partially synthetic techniques.

IPSO (Information Preserving Statistical Obfuscation) [4]

It is based on the distinction of two categories of attributes: *public data* Y and *specific survey data* X. It releases a subset of the original sample after a perturbation operation performed only over public attributes, thus obtaining a new set of values Y'. Since the main purpose of this technique is to release as many values as possible collected in the specific survey, preventing reidentification, only some information in Y is released to preserve the most important statistics S on these data. More precisely, set Y' is generated in such a way to preserve X unaltered and to maintain the set S of statistics over Y. At the end, the new sample (X, Y') is released. This technique operates on continuous attributes only.

Hybrid Masking [11]

This class of techniques combines original data with synthetic data. In particular, after the generation of a simulated sample, each tuple in the original microdata table is matched with a tuple in the simulated one. Then, all the paired tuples are linearly combined, by adding or multiplying their values, and the values obtained are published. These techniques have the advantage of preserving data analytical validity. They operate on continuous attributes only.

Random Response [2, 13]

It is used in situations where sensitive data are collected from a population and there is the possibility that individuals do not respond truthfully. For instance, if an individual has to respond to the question: "Have you ever taken drugs for depression?," the individual may lie and may respond "NO". To avoid this problem, a set of questions is prepared, where some of them are sensitive and some others are not. An individual is requested to respond to one of these questions without indicating what question has been chosen. In this way, if the distribution of the answers to the non sensitive questions is known, the percentage of positive responses on the sensitive question can be deducted from the number of positive answers. Since this technique can only be applied if the distribution of answers is known and if the questions have the same set of possible answers, it is usually adopted for boolean attributes only.

Blank and Impute [22]

It is a technique also used for protecting macrodata and consists in randomly choosing a set of tuples, either sensitive or not, deleting their original values for a given pre-determined set of attributes, and replacing them with a value computed using a suitable function (e.g., the average). For instance, suppose we choose to blank and impute attributes `DH`, `Chol`, and `Temp`, and that the

SSN	Name	Race	DoB	Sex	ZIP	MarStat	Disease	DH	Chol	Temp
		Asian	64/09/27	F	94139	Divorced	Hypertension	3	260	35.2
		Asian	64/09/30	F	94139	Divorced	Obesity	1	170	37.7
		Asian	64/04/18	M	94139	Married	Chest pain	*24*	*228*	*37.7*
		Asian	64/04/15	M	94139	Married	Obesity	7	280	37.4
		Black	63/03/13	M	94138	Married	Hypertension	*2*	*216*	*36.7*
		Black	63/03/18	M	94138	Married	Short breath	3	185	38.2
		Black	64/09/13	F	94141	Married	Short breath	5	200	36.5
		Black	64/09/07	F	94141	Married	Obesity	*20*	*216*	*37.2*
		White	61/05/14	M	94138	Single	Chest pain	7	170	37.6
		White	61/05/08	M	94138	Single	Obesity	10	300	40.1
		White	61/09/15	F	94142	Widow	Short breath	*4*	*223*	*37.2*

Fig. 14. Microdata table of Fig. 2 protected through blank and impute over attributes `DH`, `Chol`, and `Temp`

randomly selected tuples are the third, the fifth, the eighth, and the eleventh. The new values for attribute `DH` are computed as the average over patients having the same health problem; the new values for `Chol` are computed as the average over patients of the same race; the new values for `Temp` are computed as the average over patients that were born in the same month and year. Figure 14 illustrates the resulting microdata table; the new values are reported in *italic*. This technique operates on both continuous and categorical attributes.

SMIKe (Selective Multiple Imputation of Keys) [36]

It releases multiple sets of modified data rather than just one set. Let X be the set of quasi-identifiers in the original microdata table, and Y be the set of the other attributes (either sensitive or not). First, it is necessary to introduce the concept of *sensitive case*. If the number of tuples with a specific combination on attributes X is lesser that a predefined sensitive threshold, that combination is a sensitive case. SMIKe executes the following four steps before data publication.

- Each sensitive tuple t is associated with the non sensitive tuples closest to it, where the distance is computed on the basis of the values of attributes in Y. These tuples are inserted in the i-th mixing set M_i, where i is the i-th sensitive case of tuple t. The mixing sets for different sensitive cases may overlap.
- Let M be the union of sensitive cases and selected non sensitive cases. A completely *random imputation model* (i.e., a model that generates imputations for the missing values) for X is built. X' is the value imputed to X.
- A randomly set of tuples is chosen, where attribute values in X will be imputed synthetically, and randomly draws the new values from the distribution X' just defined.

- The quality of the synthetic sample is evaluated and, if it is too low, the process restarts from the beginning, trying to better tune the size of M_i.

This technique imputes only quasi-identifiers and substitutes a subset of the sensitive tuples with simulated tuples. It operates on both continuous and categorical attributes.

Multiply Imputed Partially Synthetic Dataset [27]

This class of techniques is based on the assumption that only sensitive attributes are to be protected through simulation, while other attributes can be published as in the original microdata table. The sensitive attributes can be simulated by using the multiple imputation technique above-mentioned.

In addition to the techniques here described for data generation, which can be applied to any kind of data, there are also some techniques for the protection of specific categories of data. For instance, specific regression models have been studied for the correct release of business microdata collected by census agencies [4]. Also, since the microdata release problem has become of great importance, different software solutions have been developed to protect microdata. For instance, μ-ARGUS is a software that exploits global recoding, local suppression, PRAM, additive noise, and micro-aggregation [31].

6 Measures for Assessing Microdata Confidentiality and Utility

As discussed in the previous sections, there is a broad choice of techniques for protecting microdata. A microdata protection technique has to be chosen in such a way to balance two contrasting needs: the need for data and the need for confidentiality protection. To this purpose, the performance of any protection technique is usually measured in terms of *information loss* and *disclosure risk*. Information loss is the amount of information that exists in the original microdata and because of the protection technique does not occur in protected microdata. Disclosure risk is the risk that a disclosure will be encountered if protected microdata are released. Two extreme solutions for releasing microdata are:

- the encryption of the original data (no disclosure risk and maximal information loss);
- the release of the original data (maximal disclosure risk and no information loss).

On the other hand, the application of any of the techniques presented in this chapter can provide means to balance the two. In the following, we describe some of the most important methods used for quantifying disclosure risk and information loss.

6.1 Disclosure Risk

In general, there are two types of disclosure: *identity* disclosure and *attribute* disclosure [21]. Identity disclosure means that a specific identity can be linked to a tuple in the microdata table. Attribute disclosure means that information has been disclosed about an attribute of an individual. In general, two factors may have an impact on identity disclosure:

- *population uniqueness* means that the probability of identifying a respondent who is the unique respondent with a specific combination of attributes is high if those attributes are present in the microdata table;
- *reidentification* means that the released microdata is linked to another published table, where the identifiers have not been removed.

Different methods have been proposed to measure the disclosure risk of released microdata. For instance, the *minimum unsafe combination of attributes* [49] returns the number of attributes with a unique combination in a specific microdata tuple. This method can be adopted only with non-perturbative masking techniques and the higher such a value, the lower the disclosure risk. Other specific methods have been proposed in [4, 55]. In the remainder of this section we focus on the main methods for measuring the risk of identity disclosure, which are *uniqueness* and *record linkage*, and the main method to measure attribute disclosure, which is *interval disclosure* [15, 17].

Uniqueness

Whenever a sample unique is also a population unique, identity disclosure becomes much more likely. There are different methods for evaluating the uniqueness risk and all the methods rely on probability evaluations.

The first method measures the probability of *population uniqueness* (PU), that is, the probability that there is only an individual in the population having a certain combination of values over a certain set of attributes. This probability is measured as: $Pr(PU) = \sum_j I(F_j = 1)/N$, where N is the population size, F_j is the number of individuals in the population with the j-th combination over the considered attributes, and $I()$ is a function where $I(A)$ is equal to 1 if A is true; 0 otherwise.

The second method measures the probability that *a sample unique* (SU) is also a *population unique* (PU). This probability is measured as: $Pr(PU|SU)=\sum_j I(f_j = 1, F_j = 1)/\sum_j I(f_j = 1)$, where f_j is the number of individuals in the sample with the j-th combination over the considered attributes. These two methods are called *file-level* measures because assign the same risk to all tuples [47]. *Tuple-level* disclosure risk measure is the probability that the identity of a specific individual is disclosed [26]. This measure has been introduced because the risk of reidentification is not homogeneous over the whole microdata table. Suppose that there are K different combinations of quasi-identifier values in a population. These combinations produce a

partition both on the population and on the sample. Let F_k be the frequency of the k-th partition, the disclosure risk for a tuple in the sample with the k-th combination is $1/F_k$. The problem of this method is that F_k is generally not known for the population. Since the sample distribution frequencies f_k are known, the distribution of frequencies F_k, given f_k, is considered ($F_k|f_k$ can be modeled as a negative binomial).

Note that uniqueness can be used as a measure of disclosure risk only if the microdata have been protected through a non-perturbative masking technique. Perturbative techniques change data values and therefore it is not possible to establish correctly the frequency of a value in the released sample because new unique combinations may be introduced and original unique combinations may disappear.

Record Linkage

Record linkage consists in finding a matching between a tuple in the protected microdata table and a tuple in a public and non anonymous external source of information (e.g., a voter list that contains the registry of all the electors of a region or a town). Since it is not possible to know a priori all the external sources of information that can be used by a possible malicious user, a probabilistic check on the protected microdata is performed. Different record linkage methods have to be adopted depending on whether or not the microdata table and the external information have common attributes. If there are common attributes, it is first necessary to adopt a unique representation for the common attributes. For instance, different abbreviations in the name of a person would lead to the conclusion that two tuples are not related, while actually they refer to the same respondent. It is then possible to adopt a strategy for record linkage [17, 18, 23]. Record linkage methods can be partitioned into three broad categories: *deterministic*, *probabilistic*, and *distance-based*.

- *Deterministic.* It looks for an exact match on one or more attributes between tuples in different datasets. The main disadvantage of this method is that it does not take into consideration the attribute relevance in finding a link.
- *Probabilistic.* Given two datasets, D_1 and D_2, the set of all possible pairs of tuples (d_{1i}, d_{2j}) is computed, where $d_{1i} \in D_1$ and $d_{2j} \in D_2$. Each pair is associated with a probability that represents whether the pair is a real match. If the probability is lower than a fixed threshold T_1, the pair is discarded because the tuples are considered not linked; if the probability is greater than a second fixed threshold T_2 the pair is considered a real match; if the probability is between T_1 and T_2, it is needed a human evaluation to verify whether it represents a match or not. Such a probability is computed considering different weights for different attributes and the agreement or partial agreement over the attribute values. The weights associated with the attributes and the two thresholds T_1 and T_2 are established by the data holder.

- *Distance-based.* Given two datasets, D_1 and D_2, each tuple $d_{1i} \in D_1$ is matched to the nearest tuple $d_{2j} \in D_2$. This method requires the definition of a distance function f between couples of tuples. For instance, the definition of f can exploit distance functions defined on attributes and may assign different weights to each attribute, depending on its importance in the linking process. An example of distance function is the *Euclidean Distance* that considers each tuple as a vector and assigns the same weight to each attribute. This record linkage method is not suitable for categorical attributes, because it is difficult to define the distance between two categories, in particular if their domain is not ordered.

Other methods are used when there are datasets without common attributes. In these cases, the reidentification is more difficult. One method recently proposed is based on *clustering* [19]. Basically, a clustering method is applied on the considered datasets. The result is a set of clusters of tuples and each cluster within a dataset is mapped onto a cluster within the other dataset. Such a mapping is performed by using a *similarity function*.

Note that although record linkage is considered a threat, there are many situations where it can be useful. Record linkage can be used in the management of large databases to extract important information about the same subject. This is particularly useful when data are distributed on different servers (e.g., the medical information of the population is usually distributed on different systems and a record linkage technique can be exploited for reconstructing the information associated with a given individual) [45].

Interval Disclosure

The interval disclosure measure is computed in different ways, depending on the data type of the attribute (continuous or categorical). In case of a categorical attribute, for each tuple in the microdata table, *ranked intervals* are constructed as follows. Each attribute is independently ranked and a rank interval is defined around the value assumed by the attribute in each tuple t. The ranks of values within the interval constructed around tuple t should differ less than $p\%$, of the total number of tuples. Also, the rank in the center of the interval should correspond to the value assumed by the considered attribute in tuple t. The disclosure risk is then the proportion of the original values that fall into the interval centered around the corresponding protected value. If such a proportion is equal to 100%, a potential attacker is sure that the original value lies in the interval around the protected value. In case of continuous data, the method is similar to the previous one. The main difference is how ranked intervals are constructed: it is not possible to exploit ranking and the construction is based on the standard deviation of the attribute.

6.2 Information Loss

The information loss measure is strictly connected to the *purpose* for which the information will be used. Since the purposes may be different and not known a priori, it is not possible to establish a general information loss measure based on purpose. The methods used are therefore based on the concepts of *analytically valid* and *analytically interesting*, which are defined as follows [54]:

- a protected microdata table is *analytically valid* if it approximately preserves statistical analyzes (e.g., mean and co-variance) that can be produced with the original microdata;
- a protected microdata table is *analytically interesting* if it contains a sufficient number of attributes that can be validly analyzed.

In general, there are two strategies for computing information loss: *i)* directly comparing the tuples of the protected microdata with the tuples in the original microdata; *ii)* comparing the statistics computed on the protected microdata with the same statistics evaluated on the original microdata. We now describe the basic idea of some of the most common information loss measures that are partitioned into two categories according to the data type of the attributes. Other methods have been proposed, both for specific microdata protection techniques and for generic cases [4, 55].

Continuous Data

To measure information loss, the statistic of interest (e.g., co-variance matrices, correlation matrices, or variants of them) is evaluated on both the original and protected data, the difference between the two values is computed. The discrepancies between the two statistics can be evaluated in three different ways: *mean square error*, *mean absolute error*, and *mean variation*. In addition to statistical measures, data can be compared, before and after the application of a microdata protection technique, by computing again the difference using one of the three methods above-mentioned.

It is important to note that the value of information loss should have a maximal value (e.g., 100 if a percentage notation is used) to compare different methods having the same scale for information loss computation [15, 16, 17, 37].

Categorical Data

The information loss measures briefly introduced for continuous attributes are not directly applicable for categorical attributes. In this case, there are three main measures [16]: *direct comparison*, *contingency tables comparison*, and *entropy measure*. The direct comparison of the values of categorical attributes requires the definition of a *distance function* between the categories. In case of non ordered categories, the distance between category c_1 in the original microdata and the corresponding category c_2 in the protected microdata is

equal to 0, if the two categories are the same; 1, otherwise. By contrast, if there is an ordering between the categories, the distance between categories c_1 and c_2 is equal to the number of categories between c_1 and c_2 divided by the total number of categories. The contingency tables comparison measure consists in comparing the corresponding contingency tables. An entropy-based measure [35, 53] can be used whenever a microdata table has been protected by applying the local suppression, global recoding, or PRAM techniques. The idea is that the information loss can be measured using the *Shannon Entropy* because the masking process is modeled as the noise added to the original microdata when transmitted through a noisy channel. The information loss measure uses the conditional probability (the probability of a value in the original microdata, once the value in the protected microdata is given).

6.3 Disclosure Risk and Information Loss Combination

The microdata protection techniques described in this chapter have a different impact on data utility and disclosure risk. To be able to assess alternative microdata protection techniques, we first need a framework for assessing how good a protection technique is. Disclosure risk and information loss therefore need to be combined. A simple method consists in computing the average of the 2 values and choosing the technique (and the parameter setting) that has the highest score value [17]. Another method is the *R-U confidentiality maps* [20], which is a graph where the measure of data utility (the inverse of information loss) is reported on the x axis, and the disclosure risk is reported on the y axis. For each microdata protection technique, a line is drawn on the Cartesian plane with a point for each parameter setting. On the basis of the graphic obtained, it is possible to compare the various protection techniques and choose the most suitable. Once a protection technique has been chosen, the R-U confidentiality maps can also be used for selecting the parameters. It is important to note that a R-U map is only a method for correlating disclosure risk and information loss and such measures have to be computed using one of the methods above-mentioned.

Another approach for balancing data utility and disclosure risk is represented by the concept of k-minimal table with the k-anonymity (see Chap. "k-anonymity" and [44]). k-anonymity establishes a lower bound threshold of disclosure risk for a table, by ensuring that every tuple in the table cannot be related to fewer than k respondents. The k-anonymity approach aims at finding (by applying generalization and suppression techniques) a k-minimal table, that is, one that does not generalize more than it is needed to reach the threshold k. In other words, a k-minimal table is one that minimizes information loss.

The measures described should be used before releasing the data to verify whether the protection is adequate to the respondents' requests of confidentiality and to the data recipients' needs of information. After the application

of a protection technique, the protected microdata can be checked and released only if they present a certain degree of protection. These measures can also be used by the data recipient to evaluate respondents' identity protection and data utility.

7 Conclusions

Today's globally networked society places great demand on the dissemination and sharing of information. While in the past released information was mostly in tabular and statistical form, many situations call today for the release of specific microdata. To address this issue, a wide variety of protection techniques have been proposed. In this chapter, we have described the basic microdata disclosure protection techniques, classifying them as masking techniques and synthetic data generation techniques. Masking techniques protect data by transforming their values. Synthetic data generation techniques protect data by replacing them with new data that preserve the original statistical properties. We have also illustrated the main measures usually adopted for assessing data confidentiality and data utility of the protected microdata.

8 Acknowledgments

This work was supported in part by the European Union within the PRIME Project in the FP6/IST Programme under contract IST-2002-507591 and by the Italian MIUR within the KIWI and MAPS projects.

References

1. Adam NR, Wortman JC (1989). Security-control methods for statistical databases: A comparative study. ACM Computing Surveys, 21(4):515–556.
2. Bourke PD, Dalenius T (1975). Some new ideas in the realm of randomized inquiries. Technical Report 5, Detpartment of Statistics, University of Stockholm, Stockholm, Sweden.
3. Brand R (2002). Microdata protection through noise addition. In Domingo-Ferrer J, editor, Inference Control in Statistical Databases, vol. 2316 of LNCS, pp. 97–116. Springer, Berlin Heidelberg.
4. Burridge J, Franconi L, Polettini S, Stander J (2002). A methodological framework for statistical disclosure limitation of business microdata. Technical Report 1.1-D4, CASC Project.
5. Cox LH (1980). Suppression methodology and statistical disclosure analysis. Journal of the American Statistical Association, 75(370):377–385.
6. Cox LH (1981). Linear sensitivity measures in statistical disclosure control. Journal of Statistical Planning and Inference, 5(2):153–164.

7. Cox LH (1987). A constructive procedure for unbiased controlled rounding. Journal of the American Statistical Association, 82(398):520–524.
8. Cox LH (1995). Network models for complementary cell suppression. Journal of the American Statistical Association, 90(432):1453–1462.
9. Cox LH, Dandekar RA (2002). Synthetic tabular data – An alternative to complementary cell suppression. Unpublished manuscript.
10. Dalenius T, Reiss SP (1978). Data-swapping: a technique for disclosure control (extended abstract). In Proc. of the ASA Section on Survey Research Methods, pp. 191–194, Washington DC.
11. Dandekar R, Domingo-Ferrer J, Sebé F (2002). LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In Domingo-Ferrer J, editor, Inference Control in Statistical Databases, vol. 2316 of LNCS, pp. 153–162. Springer, Berlin Heidelberg.
12. Defays D, Nanopoulos P (1993). Panels of enterprises and confidentiality: the small aggregates method. In Proc. of the 92nd Symposium on Design and Analysis of Longitudinal Surveys, pp. 195–204, Ottawa.
13. Denning DE (1982). Inference controls. In Cryptography and Data Security, pp. 331–392. Addison-Wesley Publishing Company, Reading, Massachusetts; Menlo Park, California; London; Amsterdam; Don Mills, Ontario; Sydney.
14. Domingo-Ferrer J, Mateo-Sanz JM (1999). On resampling for statistical confidentiality in contingency tables. Computers & Mathematics with Applications, 38(11-12):13–32.
15. Domingo-Ferrer J, Mateo-Sanz JM, Torra V (2001). Comparing SDC methods for microdata on the basis of information loss and disclosure risk. In Preproceedings of ETK-NTTS'2001, vol. 2, pp. 807–826, Luxemburg. Eurostat.
16. Domingo-Ferrer J, Torra V (2001). Disclosure protection methods and information loss for microdata. In Doyle P, Lane JI, Theeuwes J, Zayatz L, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam.
17. Domingo-Ferrer J, Torra V (2001). A quantitative comparison of disclosure control methods for microdata. In Doyle P, Lane JI, Theeuwes J, and Zayatz L, editors, Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam.
18. Domingo-Ferrer J, Torra V (2002). Distance-based and probabilistic record linkage for re-identification of records with categorical variables. Butlleti de l'Associacio Catalana d'Intelligencia Artificial, 27.
19. Domingo-Ferrer J, Torra V (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. Statistics and Computing, 13(4):343–354. Kluwer Academic Publishers.
20. Duncan GT, Keller-McNulty SA, Stokes SL (2001). Disclosure risk vs. data utility: The R-U confidentiality map. Technical report, Los Alamos National Laboratory. LA-UR-01-6428.
21. Duncan GT, Lambert D (1989). The risk of disclosure for microdata. Journal of Business and Economic Statistics, 7:207–217.
22. Federal Committee on Statistical Methodology (1994). Statistical policy working paper 22. USA. Report on Statistical Disclosure Limitation Methodology.
23. Fellegi IP, Sunter AB (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328):1183–1210.

24. Fienberg SE (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical Report 611, Carnegie Mellon University Department of Statistics.
25. Florian A (1992). An efficient sampling scheme: updated latin hypercube sampling. Probabilistic Engineering Mechanics, 7(2):123–130.
26. Franconi L, Polettini S (2004). Individual risk estimation in μ-ARGUS: a review. In Domingo-Ferrer J, Torra V, editors, Privacy in Statistical Databases, vol. 3050 of LNCS, pp. 262–372. Springer, Berlin Heidelberg.
27. Franconi L, Stander J (2002). A model based method for disclosure limitation of business microdata. Journal of the Royal Statistical Society D-Statistician, 51(1):1–11.
28. Gonzalez JF, Cox LH (2005). Software for tabular data protection. Statistics in Medicine, 24(4):65–669.
29. Gouweleeuw JM, Kooiman P, Willenborg RCLJ, DeWolf PP (1997). Post randomization for statistical disclosure control: Theory and implementation. Technical Report 9731, Voorburg: Statistics Netherlands, Netherlands.
30. Greenberg B (1987). Rank swapping for ordinal data. Technical report, U. S. Bureau of the Census (unpublished manuscript), Washington, DC.
31. Hundepool A, Van deWetering A, Ramaswamy R, Franconi L, Capobianchi A, DeWolf PP, Domingo-Ferrer J, Torra V, Brand R, Giessing S (2003). μ-ARGUS version 3.2 software and user's manual. Statistics Netherlands. http://neon.vb.cbs.nl/casc.
32. Huntington DE, Lyrintzis CS (1998). Improvements to and limitations of latin hypercube sampling. Probabilistic Engineering Mechanics, 13(4):245–253.
33. Karr AF, Sanil AP (2004). Data quality and data confidentiality for microdata: Implications and strategies. Technical Report 149, National Institute of Statistical Sciences, Research Triangle Park, NC 27709-4006 USA.
34. Kim JJ (1986). A method for limiting disclosure in microdata based on random noise and transformation. In Proc. of the Section on Survey Research Methods, pp. 303–308, Alexandria VA.
35. Kooiman PL, Willenborg L, Gouweleeuw J (1998). PRAM: A method for disclosure limitation of microdata. Technical report, Statistics Netherlands, Voorburg, NL.
36. Little RJA, Liu F (2002). Selective multiple imputation of keys for statistical disclosure control in microdata. In Proc. of the Section on Survey Research Methods.
37. Mateo-Sanz JM, Domingo-Ferrer J, Sebé F (2004). Probabilistic information loss measures for continuous microdata. Technical report, University of Tarragona, Department of Computer Engineering and Mathematics, Research Triangle Park, NC 27709-4006 USA.
38. Mateo-Sanz JM, Martìnez-Ballesté A, Domingo-Ferrer J (2004). Fast generation of accurate synthetic microdata. In Domingo-Ferrer J, Torra V, editors, Privacy in Statistical Databases, vol. 3050 of LNCS, pp. 298–306. Springer, Berlin Heidelberg.
39. Oganian A, Domingo-Ferrer J (2001). On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the UNECE, 18(4):345–354.
40. Polettini S, Franconi L (2002) Simulation methods in data protection: an approach based on maximum entropy. In Proc. of the International Conference of the Royal Statistical Society, Plymouth.

41. Raughnathan TE, Reiter JP, Rubin DB (2003). Multiple imputation for statistical disclosure limitation. Journal of Official Statitsics, 19(1):1–16.
42. Reiss S (1982). Non-reversible privacy transform. In Proc. of the ACM Symposium on Principles of Database Systems, Los Angeles, CA, USA.
43. Rubin DB (1993). Discussion of statistical disclosure limitation. Journal of Official Statistics, 9(2):461–468.
44. Samarati P (2001). Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027.
45. Computer Science and Telecommunications Board National Research Council, editors (1997). For the record protecting electronic health information. National Accademy Press, Washington, D.C., USA.
46. Singh AC, Yu F, Dunteman GH (2004). MASSC: A new data mask for limiting statistical information loss and disclosure. In Linden H, Riecan J, Belsby L, editors, Work Session on Statistical Data Confidentiality 2003, pp. 373–394. Eurostat, Luxemburg. Monographs in Official Statistics.
47. Skinner CJ, Elliot MA (2001). A measure of disclosure risk for microdata. Journal of the Royal Statistical Society, 64(4):855–867.
48. Sullivan GR (1989). The use of added error to avoid disclosure in microdata releases. Master's thesis, Iowa State University.
49. Takemura A (2001). On recent developments in statistical disclosure control techniques. In Proc. of the IAOS Satellite Meeting on Statistics for the Information Society, Tokyo, Japan.
50. Tendick P (1991). Optimal noise addition for preserving confidentiality in multivariate data. Journal of Statistical Planning and Inference, 27(3):341–353.
51. Tendick P, Matloff N (1994). A modified random perturbation method for database security. ACM Transactions on Database Systems, 19(1):47–63.
52. Torra V (2004). Microaggregation for categorical variables: a median based approach. In Domingo-Ferrer J, Torra V, editors, Privacy in Statistical Databases, vol. 3050 of LNCS, pp. 162–174. Springer, Berlin Heidelberg.
53. Willenborg L, DeWaal T (2001). Elements of Statistical Disclosure Control. Springer-Verlag, New York, USA.
54. Winkler WE (1999). Re-identification methods for evaluating the confidentiality of analytically valid microdata. In Domingo-Ferrer J, editor, Statistical Data Protection. Office for Official Publications of the European Communities, Luxemburg.
55. Winkler WE (2004). Masking and re-identification methods for public-use microdata: Overview and research problems. In Domingo-Ferrer J, editor, Privacy in Statistical Databases 2004. Springer, New York.

Index