Chapter 1

Anonymity Technologies for Privacy-Preserving Data Publishing and Mining

Anna Monreale

Computer Science Dept., University of Pisa, Italy

Dino Pedreschi

Computer Science Dept., University of Pisa, Italy

Ruggero G. Pensa

Computer Science Dept., University of Torino, Italy

1.1 Introduction

Data mining is gaining momentum in society, due to the ever increasing availability of large amounts of data, easily gathered by a variety of collection technologies and stored via computer systems. Data mining is the key step in the process of Knowledge Discovery in Databases, the so-called KDD process. The knowledge discovered in data by means of sophisticated data mining techniques is leading to a new generation of personalized intelligent services. The dark side of this story is that the very same collection technologies gather personal, often sensitive, data, so that the opportunities of discovering knowledge increase hand in hand with the risks of privacy violation. When personal, possibly sensitive data are published and/or analyzed, one important question to take into account is whether this may violate the right of individuals whose data is referred to — the *data subjects* — to have full control of their personal data. Some examples of data collection containing personal sensitive

information include:

- Retail market basket data: the analysis of purchase transaction data can reveal customer preferences, not only strategic and competitive information for a company;
- Social networking data: may disclose personal data, such as phone numbers and e-mail address, but more importantly may reveal relationships among people;
- E-Mails: the contents of e-mails can reveal secrets and interests of a person, besides the identity of the correspondents;
- Phone calls: the list of user's phone calls may reveal the contacts of each user;
- Mobility and location data: a collection of space-time tracks left by a mobile device may reveal the movements of a user, and the place visited.

Clearly, each of the above forms of data may potentially reveal many facets of the private life of the data subjects: but the danger is brought to a limit if the various forms of data can be linked together, painting a precise portrait even of a supposedly unknown person, whose name, or other indirect identifier, has been removed from the data. Quoting Robert O'Harrow Jr. in *No Place to Hide* (Free Press, 2005): "Most of privacy violations are not caused by the revelation of big personal secrets, but by the disclosure of many small facts in a row. Like killer bees, one is just a nuisance, but a swarm can be lethal."

Protecting private information is an important problem in our society: despite the commonsense belief that it is impossible to protect privacy in the digital era, the lack of trustable privacy safeguards in many current services and devices is at the basis of a diffusion that is often more limited than expected; also, people feel reluctant to provide true personal data, if not absolutely necessary. In several countries, many laws have been enacted, that regulate the right to the protection of personal data. In general, these laws regulate the type of information that may be collected and how this information may be stored and used. Privacy is not limited to the individuals: companies and public organizations, such as hospitals, have the necessity to protect strategic information that provides competitive assets, or the privacy of their patients. The term *corporate privacy* is used in these cases, to make a distinction from *individual privacy*.

Many recent research works have focused on privacy-preserving data mining and data publishing in order to provide some forms of data protection, and possibly to adhere to the existing legislation. Usually, these works propose techniques that allow to publish data and/or to extract knowledge while trying to protect the privacy of users and customers (or respondents) represented in the data. Some of these approaches aim at individual privacy, while others aim at corporate privacy. Unfortunately, some realistic examples show

that transforming the data in such a way to guarantee anonymity is a very hard task in the general case. Indeed, in some cases supposedly anonymous datasets can leave open unforeseen doors to a malicious attacker, that can link the personal data of an individual and the identity of the individual itself (the so-called linking attack). Therefore, many issues remain open and require further investigation. Despite an increasing interest in privacy, there exists a lack of technology in privacy-preserving data publishing and mining. This problem is reflected in the lack of communication and collaboration between the law researchers and professionals that study the definitions of the privacy regulations and the scientists that try to develop technical privacy-preserving solutions. The existing regulations pose challenges to the development of the technical solutions for the privacy issue, but we agree with [7] that this problem can only be achieved through an alliance of technology, legal regulations and social norms.

1.1.1 Privacy vs. Utility

In general, the data anonymity problem requires finding an optimal trade-off between privacy and utility. From one side, we would like to transform the data in order to avoid the re-identification of individuals whose data is referred to. Thus, we would like to publish safely the data for analysis and/or mining tasks without risks (or with negligible risk) for each data subject. From the other side, we would like to minimize the loss of information that reduces the effectiveness of the underlying data when it is given as input to data mining methods or algorithms. Therefore, the goal is to maintain the maximum utility of the data. In order to measure the information loss introduced by the anonymization process it is necessary to define measures of utility; analogously, we need to quantify the risks of privacy violation.

1.1.2 Attacks and Countermeasures

The techniques for privacy preservation strongly depend on the nature of the data that we want to protect. For example, many proposed methods are suitable for continuous variables but not for categorical variables (or the other way around), while other techniques employed to anonymize sequential data such as clinical data or tabular data are not appropriate for moving object datasets. Clearly, different forms of data have different properties that must be considered during the anonymization process. We believe that it is necessary to adopt a *purpose-oriented* anonymity framework, based on the definition of: (i) specific hypotheses on the form and the nature of the data, and (ii) specific hypotheses on the attack model for privacy violation. First, a valid framework for privacy protection has to define the background knowledge of the adversary, that strongly depends on the context and on the kind of data. Second, an attack model, based on the background knowledge of the attacker, has to be formalized. Third, a specific countermeasure associated to that attack

model has to be defined in terms of the properties of the data to be protected. The definition of a suitable attack model is very important in this context. Different assumptions on the background knowledge of an attacker entail different defense strategies. Indeed, it is clear that when the assumption on the background knowledge changes, the anonymity approach to be adopted also changes significantly. Consider, for example, that an attacker gains access to a spatio-temporal dataset and that he/she knows some spatio-temporal points belonging to some trajectory of an individual. Two cases are possible: (a) the attacker knows the exact points or (b) the attacker knows these points with a given uncertainty threshold. The attacker can try to re-identify the respondent by using his/her knowledge and by observing the protected database. Specifically, he/she should generate all the possible candidate trajectories by using the background knowledge as constraints. Clearly, the defense strategy that it is necessary to use in the case (b) might be unsuitable for the case (a), because the assumption (b) is weaker than the assumption (a). This does not mean that assumption (b) is not valid, as it can be adequate for particular situations where (a) is unrealistically strong.

1.1.3 Privacy-Preserving Data Mining and Statistical Disclosure Control

The problem of protecting privacy when disclosing information is not trivial and this makes the problem scientifically attractive. It has been studied extensively in two different communities: in data mining, under the general umbrella of *privacy-preserving data mining*, and in statistics, under the general umbrella of *statistical disclosure control*. Often, the different communities have investigated lines of work which are quite similar, sometimes with little awareness of this strong tie. In this chapter, we will provide a survey of the main anonymity techniques proposed by the two different communities, analyzing them from the two perspectives. The Figure 1.1 shows a taxonomy tree that describes our classification of the privacy-preserving techniques.

1.1.4 Anonymity in Data Protection Laws

We also provide a brief discussion on how the concept of anonymous data is defined in the most influential privacy regulations enacted internationally. The baseline of this discussion is that no satisfactorily precise definition of data anonymity is up to date available: given that, it is unrealistic to define anonymity as the theoretical impossibility to re-identify the data subject(s) by looking at their data, the legal definitions adopt concepts like reasonableness, or disproportion of efforts — agreeable yet vague concepts, weakly actionable to the purpose of assessing the degree of anonymity protection supported by a given technology or organizational procedure. This observation brings evidence that more work is needed to make anonymity definitions operational in the juridical practice: in this precise sense, there's a strong need for quantifi-

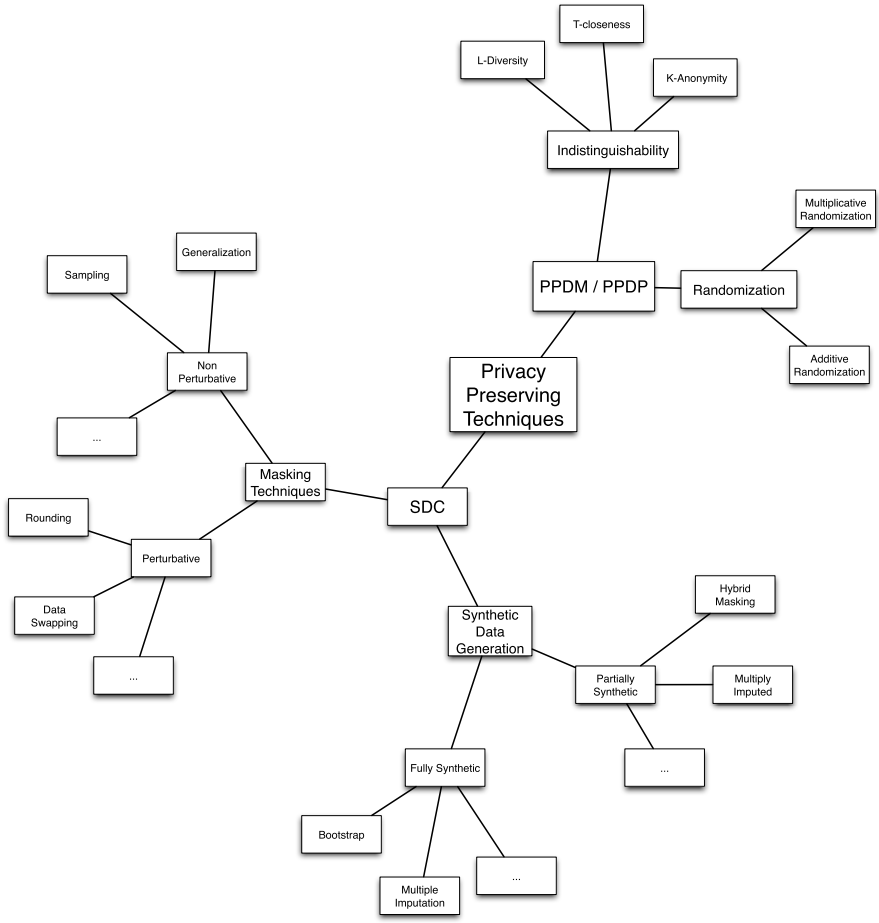

FIGURE 1.1: Taxonomy of privacy-preserving techniques

able notions of privacy and anonymity that can help privacy laws and privacy technologies to have a stronger impact on people's life.

1.1.5 Anonymity in Complex Data

Also, a discussion about anonymity in complex data domain is provided, in order to underline that the techniques proposed for tabular data are not always suitable for data of more complex nature, where specific semantics may offer to the attackers more means to link the data with external knowledge. Specifically, we focus our attention on spatio-temporal data and a brief overview on the recent approaches for the anonymity of this particular form of data is provided, while other chapters in this book have a similar aim toward

other forms of data, including query logs, (social) networking data, etc.

1.1.6 Plan of the Chapter

The chapter is organized as follows. Section 1.2 provides an overview on the main privacy-preserving data publishing and mining techniques proposed by the data mining community. Section 1.3 presents the main techniques for anonymity of microdata proposed by the statistical disclosure control community. An overview on how privacy has been considered in the legal frameworks of different countries is presented in Section 1.4. Section 1.5 discusses the privacy issues in complex domains, focusing the attention on the context of spatio-temporal data and describes some approaches proposed for anonymity of this type of data. Finally, Section 1.6 concludes.

1.2 Anonymity for Data Publishing and Mining

We have discussed how the importance of privacy-preserving data publishing and mining is growing. In this section, we provide an overview of the anonymity techniques proposed in the literature.

1.2.1 Anonymity by Randomization

Randomization methods are used to modify data with the aim of preserving the privacy of sensitive information. They were traditionally used for statistical disclosure control [2] and later have been extended to privacy-preserving data mining problems [8]. Randomization is a technique for privacy-preserving data mining using a noise quantity in order to perturb the data. The algorithms belonging to this group of techniques first of all modify the data by using randomization techniques. Then, from the perturbed data it is still possible to extract patterns and models.

In literature, there exist two types of random perturbation techniques:

- additive random perturbation
- multiplicative random perturbation.

1.2.1.1 Additive Random Perturbation

In this section, we will discuss the method of *additive random perturbation* and its applications in the data mining problem. This method can be described as follows. Denote by $X = \{x_1 \dots x_m\}$ the original dataset. The new distorted dataset, denoted by $Z = \{z_1 \dots z_m\}$, is obtained drawing independently from the probability distribution a noise quantity n_i and adding it to each record

$x_i \in X$. The set of noise components is denoted by $N = \{n_1, \ldots, n_m\}$. The original record values cannot be easily guessed from the distorted data as the variance of the noise is assumed large enough. Instead, the distribution of the dataset can be easily recovered. Indeed, if X is the random variable representing the data distribution for the original dataset, N is the random variable denoting the noise distribution, and Z is the random variable describing the perturbed dataset, we have:

$$Z = X + N$$
$$X = Z - N$$

Notice that, both m instantiations of the probability distribution Z and the distribution N are known. In particular, the distribution N is known publicly. Therefore, by using one of the methods discussed in [8, 6], we can compute a good approximation of the distribution Z, by using a large enough number of values of m. Then, by subtracting N from the approximated distribution of Z, we can compute N approximation of X. At the end of this process individual records are not available, while we obtain a distribution only along individual dimensions describing the behavior of the original dataset X.

The additive perturbation method has been extended to several data mining problems. But, it is evident that traditional data mining algorithms are not adequate as based on statistics extracted from individual records or multivariate distributions. Therefore, new data mining approaches have to be devised to work with aggregate distributions of the data in order to obtain mining results. This can sometimes be a challenge. In the works presented in [8, 70, 71] authors propose new techniques based on the randomization approach in order to perturb data, and then we build classification models over randomized data. In particular, the work in [8] is based on the fact that the probability distribution is sufficient in order to construct data mining models as classifiers. Authors show that the data distribution can be reconstructed with an iterative algorithm. Later, in [6] Agrawal and Aggarwal show that the choice of the reconstruction algorithm affects the accuracy of the original probability distribution. Furthermore, they propose a method that converges to the maximum likelihood estimate of the data distribution. Authors in [70, 71] introduce methods to build a Naive Bayesian classifier over perturbed data. Randomization approaches are also applied to solve the privacy-preserving association rules mining problem as in [57, 28]. In particular, the paper [57] presents a scheme attempting to maximize the privacy to the user and to maintain a high accuracy in the results obtained with the association rule mining. While in [28], authors present a framework for mining association rules from randomized data. They propose a class of randomization operators more effective than uniform distribution and a data mining approach to recover itemset supports from distorted data.

1.2.1.2 Multiplicative Random Perturbation

For privacy-preserving data mining, *multiplicative random perturbation* techniques can also be used. There exist two types of multiplicative noise. The first one applies a logarithmic transformation to the data, and generates a random noise that follows a multivariate normal distribution with mean equal to zero and constant variance. Then, this noise is added to each element of the transformed data. Finally, the antilog of the noise-added data is taken. The second approach generates random noise by truncated normal distribution with mean equal to 1 and small variance, and then multiplies this noise by the original data. This method preserves the inter-record distances approximately. Therefore, in this case it is possible to reconstruct both aggregate distributions and some record-specific information as distance. This means that the multiplicative random perturbation method is suitable for many data mining applications. For example, in the work presented in [18] authors showed that this technique can be applied to the problem of classification. Moreover, the technique is suitable for the problem of privacy-preserving clustering [53, 54]. The work in [53] introduces a family of geometric data transformation methods (GDTMs) that distort confidential numerical attributes in order to meet privacy protection in clustering analysis. Oliveira et al. in [54] address the problem of guaranteeing privacy requirements while preserving valid clustering results. To achieve this dual goal, the authors introduce a novel spatial data transformation method called Rotation-Based Transformation (RBT) and for distributed privacy-preserving data mining as shown in [45]. The main techniques of multiplicative perturbation are based on the work presented in [37].

1.2.1.3 Strengths and Weakness of Randomization

The main advantage of the randomization method is that it can be implemented at data-collection time, because it is very simple and does not require knowledge of the distribution of other records in the data for the data transformation. This means that the anonymization process does not need a trusted server containing all the original records.

The problem of the randomization is that it does not consider the local density of the records and thus, all records are handled equally. Outlier records can be compared to records in denser regions in the data and thus, this can make an attack easier. Another weakness of the randomization framework is that it does not provide guarantees in case of re-identification attack done by using public information. Specifically, if an attacker has no background knowledge of the data, then the privacy can be difficult to compromise. Instead, in [3], authors showed that the randomization method is unable to effectively guarantee privacy in high-dimensional cases. Moreover, they provide an analysis revealing that the use of public information makes this method vulnerable. In [38] Kargupta et al. challenged the effectiveness of randomization methods, showing that the original data matrix can be obtained from the randomized data matrix using a random matrix-based spectral filtering technique.

1.2.2 Anonymity by Indistinguishability

As said in the previous section, the randomization method has some weakness. The main problem is that it is not safe in case of attacks with prior knowledge. When the process of data transformation for privacy-preserving is not to be performed at data-collection time, it is better to apply methods that reduce the probability of record identification by public information. In literature three techniques have been proposed: *k-anonymity*, *l-diversity* and *t-closeness*. These techniques differ from the randomization methods as they are not data-independent.

1.2.2.1 *k*-Anonymity

One approach to privacy-preserving data publishing is *suppression* of some of the data values, while releasing the remaining data values exactly. However, suppressing just the identifying attributes is not enough to protect privacy, because other kinds of attributes, that are available in public such as age, zipcode and sex can be used in order to accurately identify the records. These kinds of attributes are known as *quasi-identifiers* [63]. In [62] it has been observed that for 87% of the population in the United States, the combination of Zip Code, Gender and Date of Birth corresponded to a unique person. This is called *record linkage*. In this work, authors proposed **k-anonymity** in order to avoid the record linkage. This approach became popular in privacy-preserving data publishing. The goal of k-anonymity is to guarantee that every individual object is hidden in a crowd of size k. A dataset satisfies the property of k-anonymity if each released record has at least $(k-1)$ other records also visible in the release whose values are indistinct over the quasi-identifiers. In k-anonymity techniques, methods such as *generalization* and *suppression* are usually employed to reduce the granularity of representation of quasi-identifiers. The method of *generalization* generalizes the attribute values to a range in order to reduce the granularity of representation. For instance, the city could be generalized to the region. Instead, the method of *suppression*, removes the value of an attribute. It is evident that these methods guarantee privacy but also reduce the accuracy of applications on the transformed data.

The work proposed in [59] is based on the construction of tables that satisfy the k-anonymity property by using domain generalization hierarchies of the quasi-identifiers. The main problem of the k-anonymity is to find the minimum level of generalization that allows us to guarantee high privacy and a good data precision. Indeed, in [50], Meyerson and Williams showed that the problem of optimal k-anonymization is NP-hard. Fortunately, many efforts have been done in this field and many heuristic approaches have been designed as those in [43, 12]. LeFevre et al. in [43] propose a framework to implement a model of k-anonymization, named full-domain generalization. They introduce a set of algorithms, called *Incognito* that allows us to compute a k-minimal generalization. This method generates all possible full-domain generalizations of a given table and thus, uses a bottom-up breadth-first search

of the domain generalization hierarchy. In particular, it begins by checking if the single quasi-identifiers attributes satisfy the k-anonymity property and removing all the generalizations that do not satisfy it. In general, for each iteration i the *Incognito* algorithm performs these operations for the subset of quasi-identifiers of size i. Another algorithm, called *k-Optimize* is presented in [12] by Bayardo and Agrawal. This approach determines an optimal k-anonymization of a given dataset. This means that it perturbs the dataset as little as is necessary in order to obtain a dataset satisfying the k-anonymity property. In particular, the authors try to solve the problem to find the power-set of a special alphabet of domain values. They propose a top-down search strategy, i.e., a search beginning from the most general to the more specific generalization. In order to reduce the search space *k-Optimize* uses pruning strategies. Another interesting work has been proposed in [65], where a bottom-up generalization approach for k-anonymity is presented. Instead, in [33] the authors introduced a method of top-down specialization for providing an anonymous dataset. Both these algorithms provide masked data that are still useful for building classification models.

The problem of k-anonymization can be seen as a search over a space of possible multi-dimensional solutions. Therefore, some work used heuristic search techniques such as genetic algorithms and simulated annealing [36, 67]. Unfortunately, by applying these approaches the quality of the anonymized data is not guaranteed and often they require high computational times.

Aggarwal et al. proposed an approach based on clustering to implement the k-anonymity [4]. The same basic idea is used in [24], where the authors described how to use micro-aggregation for obtaining k-anonymity. Moreover, it has been studied how some approximation algorithms guarantee the quality of the solution of this problem [50, 5]. In particular, in [5] the authors provide an $O(k)$-approximation algorithm for k-anonymity that uses a graph representation. By using a notion of approximation, authors try to minimize the cost of anonymization, due to the number of entries generalized and the degree of anonymization.

In literature, there also exist applications of the k-anonymity framework that preserve privacy while publishing valid mining models. For example, in [9, 10, 11] the authors focused on the notion of individual privacy protection in frequent itemset mining and shift the concept of k-anonymity from source data to the extracted patterns.

Finally, another application of the k-anonymity notion is proposed in [56], where the authors addressed the problem of anonymizing sequence dataset trying to preserve sequential frequent pattern mining results. They reformulated the anonymization problem as the problem of hiding all the sequences occurring less than k times in the original sequence dataset.

Based on the definition of k-anonymity, new notions such as l-diversity [46] and t-closeness [44] have been proposed to provide improved privacy.

1.2.2.2 *l*-Diversity

In literature, there exist many techniques based on the k-anonymity notion. It is due to the fact that k-anonymity is a simple way to reduce the probability of record identification by public information. Unfortunately, the k-anonymity framework in some cases can be vulnerable; in particular, it is not safe against homogeneity attacks and background knowledge attacks, that allow to infer the values of sensitive attributes. Suppose that we have a k-anonymous dataset containing a group of k entries with the same value for the sensitive attributes. In this case, although the data are k-anonymous, the values of the sensitive attributes can be easily inferred (Homogeneity Attack). Another problem happens when an attacker knows information useful to associate some quasi-identifiers with some sensitive attributes. In this case the attacker can reduce the number of possible values of the sensitive attributes (Background Knowledge Attack). In order to eliminate this weakness of the k-anonymity the technique of l-diversity was proposed [46]. The main aim is to maintain the diversity of sensitive attributes. In particular, the main idea of this method is that every group of individuals that can be isolated by an attacker should contain at least l *well-represented* values for a sensitive attribute. A number of different instantiations for the l-diversity definition are discussed in [46, 68].

1.2.2.3 *t*-Closeness

l-diversity is insufficient to prevent an attack when the overall distribution is skewed. The attacker can know the global distribution of the attributes and use it to infer the value of sensitive attributes. In this case, the **t-closeness** method introduced in [44] is safe against this kind of attack. This technique requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. The distance between the two distributions should be no more than a threshold t [44].

1.3 Statistical Disclosure Control

The aim of Statistical Disclosure Control (SDC) is to protect statistical data. In particular, it seeks to modify the data in such a way that they can be published and mined without compromising the privacy of individuals or entities occurring in the database. In other words, SDC seeks to provide safe techniques against linking attacks. Moreover, after the data protection, data analyses have to be possible and the results obtained should be the same or similar to the ones that would be obtained analyzing the data before the protection.

The youngest sub-discipline of SDC is the microdata protection. It aims at protecting static individual data, also called *microdata*. In this section, we provide a survey of SDC methods for microdata, the most common data used for data mining.

A microdata set X can be viewed as a table or a file with n records. Each record related to a respondent contains m values associated to m attributes. The attributes can be classified in the following categories: *Identifiers*, *Quasi-identifiers*, *Confidential attributes* and *Non-confidential attributes*.

As stated above, the purpose of SDC is to prevent confidential information from being linked to specific respondents, thus we will assume all the identifiers have been removed from the original microdata sets to be protected.

In the literature, several microdata disclosure protection methods have been proposed. Microdata protection methods can classified as follows: *masking techniques* and *synthetic data generation techniques*.

Masking techniques, usually, generate a modified version of the original microdata set, which is still suitable for statistical analysis although the respondents' privacy is guaranteed and can be divided in two sub-categories [66]: Non-perturbative and Perturbative. Synthetic data generation techniques, instead, produce new data that replace the original data and preserve their key statistical properties. The released synthetic data are not referred to any respondent. Hence, the release of this data cannot lead to re-identification. The techniques can be of two kinds: *fully synthetic techniques* and *partially synthetic techniques*.

1.3.1 Non-Perturbative Masking Techniques

Non-perturbative techniques do not modify the original dataset; rather, these methods produce a protected dataset by using suppressions or reductions of details in the original dataset. Some of these methods are suitable only for categorical data while others are suitable for both continuous and categorical data.

Non-perturbative methods include: *Sampling*, *Generalization*, *Global Recoding* and *Local Suppression*.

1.3.1.1 Sampling

Sampling methods allow us to publish a sample of the original microdata [66]. Thus, the protected microdata contains only the data about a part of the whole population. In this way, the probability of not finding the data about a specific respondent in the protected microdata may be not null; this reduces the risk of re-identification of a respondent. These methods are not suitable for continuous data.

1.3.1.2 Generalization

Generalization provides protected microdata by replacing the values of a given attribute by using more general values [59]. This technique first of all defines a *generalization hierarchy*. The most general value of this hierarchy is at the root of it while the most specific values are represented by the leaves. The generalization method, thus, replaces the values represented by the leaf nodes with one of their predecessor nodes. A particular case of generalization is the global recoding technique. Clearly, it is possible to generate different generalizations of a microdata set.

1.3.1.3 Global Recoding

The Global Recoding method reduces the details in the microdata by substituting the value of some attributes with other values [25, 26]. For a continuous attribute, the method divides in disjoint intervals the domain of that attribute. Then it associates a label to each interval and finally, replaces the real attribute value with the label associated with the corresponding interval. For a categorical attribute, the method combines several categories in order to form new and less specific categories and then the new value is computed. Two particular global recoding techniques are the *Top-coding* and the *Bottom-coding*. The first one [25, 26] is a method based on the definition of *top-code*, that is an upper limit. The idea is that values above a certain threshold are replaced with the top-code. Similarly, the second one [25, 26] is based on a notion of *lower limit*, named *bottom-code*, that is used to replace any value lower than this limit. *Top-coding* and *Bottom-coding* can be applied to both continuous attributes and to categorical attributes that can be linearly ordered.

1.3.1.4 Local Suppression

The Local Suppression method [59] suppresses the value of some individual or sensitive attributes, by replacing them with a missing value. In this way the possibility of analysis is limited. DeWaal et al. in [22] discussed the combinations of local suppression and global recoding techniques.

1.3.2 Perturbative Masking Techniques

Perturbative techniques alter the microdata set before publication for preserving statistical confidentiality. The statistics computed on the dataset protected by perturbation do not differ significantly from the ones computed on the original microdata set. In general, a perturbative approach modifies the microdata set by introducing new combinations of values and making unique combinations of values in the original microdata set. In the following, we describe the main approaches belonging to this group of techniques.

1.3.2.1 Random Noise

These methods perturb microdata set by adding random noise following a given distribution [55]. In general, the *additive noise*, given X_j the j-th column of the original microdata, replaces each x_{ij} $(i = 1 \ldots n)$ with $x_{ij} + e_{ij}$, where e_j is an error vector. Two kinds of additive noise exist in the literature: *uncorrelated* and *correlated*. In the former, e_{ij} is a vector of normally distributed errors drawn from a random variable with mean equal to zero and with a variance that is proportional to those of the original attributes. This does not preserve variances and correlation coefficients, but preserves mean and covariance. In the latter, the co-variance matrix of the errors is proportional to the co-variance matrix of the original data; therefore, this technique also preserves correlation coefficients. Additive noise is often combined with *linear* or *non linear* transformations [41, 61]. This means that before publishing the microdata, linearly /non linearly transformation has to be applied on the data after the process of noise addition. Notice that additive noise is usually not suitable to protect categorical data. As stated in Section 1.2.1 the Randomization techniques introduced by the data mining community come from the methods traditionally used in statistical disclose control described now.

1.3.2.2 Data Swapping

In order to perturb the data another technique can be used, i.e., so-called *Data Swapping*. This technique does not change the aggregate statistical information of the original data, although the confidentiality of individual sensitive information is preserved. The basic idea is to switch a subset of attributes between selected pairs of records in the original database [29]. In this way, the data confidentiality is not compromised and the lower order frequency counts or marginals are preserved. Therefore, certain kinds of aggregate computations can be performed without compromising the privacy of the data.

1.3.2.3 Rank Swapping

Rank Swapping [25] is seen as a variation of the swapping method. It is possible to apply this technique to both continuous and categorical attributes, which can be sorted by using an order relationship. The idea is to rank the values of an attribute according to their ascending order. Then, each value is swapped with another value, guaranteeing that the swapped records are within a specified rank-distance of one another.

1.3.2.4 Resampling

The Resampling technique [25, 23] replaces the values of a sensitive continuous attribute with the average value computed over a given number of samples of the original population in the microdata set. Specifically, if we consider h independent samples $S_1, \ldots S_h$ of the values of T_i (an original attribute). This

method sorts the samples considering the order of original values. Then, it computes the set $\bar{t_1}, \ldots, \bar{t_n}$ where each $\bar{t_j}$ is the average of the j-th ranked values in $S_1, \ldots S_h$ and n is the number of records. Finally, the masked attribute is generated replacing each original value of the attribute with the correspondent average value.

1.3.2.5 Rounding

Rounding replaces original values of attributes with rounded values. In order to replace the value of an attribute, the technique defines a *rounding set*, that for example contains the multiples of a given base value. Then, it selects rounded values in this set. Usually, this method is suitable for continuous data. In case of multivariate original datasets, univariate rounding is usually performed, i.e., the rounding is applied on one attribute at a time. However, in [66, 20] authors show that it is possible to perform multivariate rounding.

1.3.2.6 PRAM

PRAM (Post RAndomized Method) [42, 26] allows one to perturb categorical value for one or more attributes by using a probabilistic mechanism, namely a Markov matrix. Each row of this matrix contains the possible values of each attribute. It is important to notice that the choice of the Markov matrix affects the risk of disclosure and information loss.

1.3.2.7 MASSC

MASSC (Micro-Agglomeration, Substitution, Sub-sampling and Calibration) [60] is a perturbative technique that consists of four steps:

- *Micro-agglomeration*: Records in the original microdata set are partitioned into different groups. Each group contains records with a similar risk of disclosure. Moreover, each group is formed using the quasi-identifiers in the records. Intuitively, records with rare combinations of values for quasi-identifiers are considered to be at a higher risk and thus, they should be in the same group.

- *Substitution*: An optimal probabilistic strategy is used to perturb the data.

- *Sub-sampling*: Some attributes or whole records are suppressed by using an optimal probabilistic subsampling strategy.

- *Optimal calibration*: In order to preserve a specific statistical property, the sampling weights, used in the previous step, are calibrated.

This technique is not suitable for datasets containing continuous attributes.

1.3.2.8 Micro-Aggregation

The Micro-Aggregation technique, described in [25], groups individual record into aggregates of dimension k. Next, given a group, its average value is computed and then it is published instead of individual values. In this kind of method an important notion is the *maximal similarity function*, which is used in order to form the groups. Finding an optimal grouping solution is a difficult problem [52], so some heuristic algorithms have been proposed to maximize similarity. There are different variations of micro-aggregation approaches. For example, some of them use different grouping strategies for the perturbation of different attributes. Other methods, instead, use the same grouping.

Another strategy consists in substituting the original value of all tuples (or part of them) in a group with the mean. Micro-aggregation is proposed for protecting both continuous attributes and categorical data.

1.3.3 Fully Synthetic Techniques

Fully synthetic techniques generate a set of data that is completely new. This means that the released data are referred to any respondent. Hence, no respondent can be re-identified. Different techniques exist that can be applied only on categorical or continuous data, or on both of them. Some methods belonging to this category are: *Cholesky decomposition* [49], *Bootstrap* [30], *Multiple imputation* [58], *Latin Hypercube Sampling* [31].

1.3.3.1 Cholesky Decomposition

This technique is based on the Cholesky matrix decomposition method and consists of five steps [49]:

1. Represent the original microdata set X as a matrix with N rows (tuples) and M columns (attributes)
2. Compute the co-variance matrix C over X
3. Generate a random matrix of $N \times M$ elements, named R, such that its co-variance matrix is the identity matrix I
4. Compute the Cholesky decomposition D of C, such that $C = D^t \times D$
5. Generate the synthetic data X' by matrix product $R \times D$.

Notice that X' and X have the same co-variance matrix. Indeed, this approach preserves variance, co-variance and mean of the original microdata set. This method is suitable for continuous attributes.

1.3.3.2 Bootstrap

This technique generates synthetic data by using Bootstrap methods [30]. In particular, it computes the p-variate cumulative distribution function F

of the original dataset X with p attributes. Then, it alters the function F in order to transform it into another similar function F'. Finally, this last function is sampled to generate a synthetic dataset X'.

Notice that this method is particularly suitable for continuous data.

1.3.3.3 Multiple Imputation

The multiple imputation technique [58] considers a dataset X of N tuples corresponding to a sample of N respondents belonging to a larger population of M individuals. Moreover, it divides the attributes into: background attributes (A), non-confidential attributes (B) and confidential attributes (C). The first ones are available for the whole population, while the second ones and the last ones are only known for the N individuals.

This method is performed in three steps:

a. Construct a multiply imputed population of M individuals starting from X. In this population there are N tuples of X and k matrices (B, C) for the $M - N$ remaining individuals. Notice that k is the number of multiple imputations.

b. Predict a set of couples (B, C) starting from the attributes A using a prediction model. In this way, the whole population has a value for each kind of attribute, some values will be imputed while others will be original.

c. Draw a sample X' of N records from the multiply imputed population. This is can be repeated k times in order to produce k replicates of (B, C) values. At the end, k multiply imputed synthetic datasets are obtained and in order to assure that no original data are contained in synthetic datasets, when the sample is drawn the N original tuples are excluded.

Multiple imputation method works on both categorical and continuous data.

1.3.3.4 Latin Hypercube Sampling

Latin Hypercube Sampling [31] is a technique that provides both the univariate and the multivariate structure of the original dataset. Usually, univariate structures are the mean and covariance of an attribute, while a multivariate structure is the rank correlation structure of the data. The main problem of this technique is that it is time-intensive and its complexity depends on the number of statistics to preserve on synthetic data and on the values to be reproduced.

The Latin Hypercube Sampling method can be used on both categorical and continuous data.

1.3.4 Partially Synthetic Techniques

Partially synthetic techniques produce a dataset, where the original data and synthetic data are mixed. In literature, several techniques belonging to this category have been proposed and in the following we describe the main ones.

1.3.4.1 Hybrid Masking

The Hybrid Masking method [21] is based on the idea of combining original and synthetic data in order to mask the data. Usually, this technique generates synthetic records. Using a distance function, each original record is matched with synthetic record. The masked data are obtained by combining the paired records. For the combination the values in these records can be added or multiplied.

1.3.4.2 Information Preserving Statistical Obfuscation

This technique, proposed in [17], explicitly preserves certain information contained in the data. The data are assumed to be composed of two kind of information for each respondent: *public data* and *specific survey data*. Before releasing the data, this approach alters the public data by a perturbation operation, while it discloses the specific survey data without alteration. Usually, both sets of data are released for a subset of respondents. The new set of data maintains a certain set of statistics over the original public data.

1.3.4.3 Multiply Imputed Partially Synthetic Data

This technique alters confidential attributes by using the multiple imputation method to simulate them, while releasing the others attributes without alteration [32]. The basic idea is that only the confidential attributes should be protected. This method can be applied to both categorical and continuous data.

1.3.4.4 Blank and Impute

The Blank and Impute technique [55] is suitable for both categorical and continuous data. First of all, it selects some records randomly and then deletes the original values of a set of attributes in these records. The deleted values are replaced by a sort of imputation method.

1.4 Anonymity in Privacy Regulations

In recent years, privacy has been one of the most discussed jurisdictional issues in many countries. Citizens are increasingly concerned about what companies and institutions do with their data, and ask for clear positions and policies from both the governments and the data owners. Despite this increasing need, there is not a unified view on privacy laws across countries. In some of them (e.g., Spain, Portugal) the right to privacy has been established as a constitutional principle; in other countries (e.g., USA) there exist multiple law articles that deal with special cases. In this section we will present an overview on how privacy has been considered in the jurisdiction of Canada, United States and European Union.

1.4.1 Privacy Laws in Canada

Canada's first response at the government level to the call for protection of personal information — or data protection, as it is frequently called in Europe — was to introduce data protection provisions into the Canadian Human Rights Act. Subsequently, the Canadian Charter of Rights and Freedoms outlined that everyone has "the right to life, liberty and security of the person" and "the right to be free from unreasonable search or seizure," but never mentioned directly the concept of privacy. However, in 1982, Parliament enacted purpose-specific legislation — the federal Privacy Act. This act puts limits and obligations on over 150 federal government departments and agencies, on the collection, use and disclosure of personal information. It also gives Canadians the right to find out what personal information the federal government has about them by making a formal request under the Privacy Act. The Office of the Privacy Commissioner of Canada has the authority to investigate complaints. The Act came into force the following year.

The governments of all provinces and territories in Canada, except for Newfoundland and Labrador, also have legislation governing the collection, use and disclosure of personal information. The legislation varies from province to province, but the general right to access and correct personal information exists in all, and each has a commissioner or ombudsman who is authorized to handle complaints.

Canada also promulgates an important act concerning more specifically the private sector. The 2000 Personal Information Protection and Electronic Documents Act, or PIPED Act, regulates, in provinces without a similar legislation, how private sector organizations collect, use and disclose personal information in the course of business activities. This Act has been implemented in three stages: in 2001 the federally regulated private sector, for example banks and international air carriers, is covered; in 2002 it covered personal health information collected, used or disclosed by federally regulated organizations;

from 2004 it covers information collected in the course of any commercial activity in any province or territory in Canada, including provincially regulated organizations. At the beginning of 2004, the only province exempted from the federal PIPED Act was Quebec. Quebec businesses are not covered by the PIPED Act, but must comply with the Quebec private sector privacy law.

The PIPED Act establishes ten principles that organizations must follow when collecting, using and disclosing personal information in the course of commercial activity. Among them: identifying purpose, consent, limiting collection, limiting use, disclosure and retention. Substantially it supports and promotes e-commerce by "protecting personal information that is collected, used or disclosed in certain circumstances, by providing for the use of electronic means to communicate or record information or transactions." It states that an organization "may collect, use or disclose personal information only for purposes that a reasonable person would consider are appropriate in the circumstance," but does not apply to data "rendered anonymous" and does not mention any example of "reasonable" method of identification.

In 2004, the Canadian Institutes of Health Research (CIHR) proposed a clarification of PIPEDA that offers an interpretation of "reasonableness" as a reasonably foreseeable method of identification or linking of data with a specific individual. However, it also refers to anonymized data as information permanently stripped of all identifiers, such that the data has no reasonable potential for any organization to make an identification. Finally, it states that reasonable foreseeability should be assessed with regard to the circumstances prevailing at the time of the proposed collection, use, or disclosure.

1.4.2 Privacy Laws in the United States

In the United States, the right to privacy is the right to be let alone, in the absence of some "reasonable" public interest in a person's activities, like those of celebrities or participants in newsworthy events. Invasion of the right to privacy can be the basis for a lawsuit for damages against the person or entity violating the right. The right to privacy is not mentioned in the Constitution, but the Supreme Court has interpreted several of the amendments as creating this right. One of the amendments is the Fourth Amendment, which stops the police and other government agents from searching us or our property without "probable cause" to believe that we have committed a crime. Other amendments protect our freedom to make certain decisions about our bodies and our private lives without interference from the government. Rights derived from the Fourth Amendment are limited by the legal requirement of a "reasonable expectation of privacy." The due process clause of the 14th amendment generally only protects privacy of family, marriage, motherhood, procreation, and child rearing. The Constitution, however, only protects against state actors. Invasions of privacy by individuals can only be remedied under previous court decisions.

Invasion of privacy is a commonly used cause of action in legal pleadings.

In the United States, the development of the doctrine regarding this tort was largely spurred by an 1890 Harvard Law Review article written by Samuel D. Warren and Louis D. Brandeis on The Right of Privacy. Modern tort law includes four categories of invasion of privacy, i.e., intrusion of solitude, public disclosure of private facts, false light, and appropriation. In particular, public disclosure of private facts arises where one person reveals information which is not of public concern, and the release of which would offend a reasonable person. Disclosure of private facts includes publishing or widespread dissemination of little-known, private facts that are non-newsworthy, not part of public records, public proceedings, not of public interest, and would be offensive to a reasonable person if made public. Although partial regulations exist, there is no all-encompassing law regulating the acquisition, storage, or use of personal data in the US. In general terms, in the US, whoever can be troubled to key in the data is deemed to own the right to store and use it, even if the data were collected without permission. Moreover, in the United States today there are separate privacy laws for medical information, financial information, library records, video rental records, GPS tracking, and numerous other classes of data.

In the US, the use of GPS trackers by police requires a search warrant in some circumstances, but use by a private citizen does not, as the Fourth Amendment does not limit the actions of private citizens. Other laws, like the common law invasion of privacy tort as well as state criminal wiretapping statutes (for example, the wiretapping statute of the Commonwealth of Massachusetts, which is extremely restrictive) potentially cover the use of GPS tracking devices by private citizens without consent of the individual being so tracked.

1.4.3 Privacy Laws in the European Union

Unlike in the United States, the right to data privacy is heavily regulated and rigidly enforced in Europe. Article 8 of the European Convention on Human Rights (ECHR) provides a right to respect for one's "private and family life, his home and his correspondence," subject to certain restrictions. The European Court of Human Rights has given this article a very broad interpretation in its jurisprudence. According to the Court's case law, the collection of information by officials of the state about an individual without his consent always falls within the scope of Article 8. Thus, gathering information for the official census, recording fingerprints and photographs in a police register, collecting medical data or details of personal expenditures and implementing a system of personal identification has been judged to raise data privacy issues.

The government is not the only entity which may pose a threat to data privacy. The Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data was concluded within the Council of Europe in 1981. This convention obliges the signatories to enact legislation concerning the automatic processing of personal data, which many duly did.

As all the member states of the European Union are also signatories of the European Convention on Human Rights and the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, the European Commission was concerned that diverging data protection legislation would emerge and impede the free flow of data within the EU zone. Therefore the European Commission decided to harmonize data protection regulation and proposed the Directive (95/46/CE) on the protection of personal data, which was voted in the 1995 by the European Parliament, and which member states had to transpose into law by the end of 1998.

Personal data covers both facts and opinions about the individual. It also includes information regarding the intentions of the data controller towards the individual, although in some limited circumstances exemptions will apply. With processing, the definition is far wider than before. For example, it incorporates the concepts of "obtaining," "holding" and "disclosing." Also, mobility data has been considered within this jurisdictional framework by the Directive 2006/24/CE of the European Commission.

All EU member states adopted legislation pursuant to this directives or adapted their existing laws. Each country also has its own supervisory authority to monitor the level of protection. In particular, a critical principle of the EU Directive, which has also been proposed in national jurisdictions, is the 26th considerandum, which states:

> Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable;...

This principle raises important discussions on how to measure the "reasonableness" of identifying means. Clearly, many parameters are likely to be involved in the definition of *reasonable*, such as, computational resources to be employed in term of time and money, number of possible linked sources to be considered for re-identification, and so on. A considerable effort should be then undertaken by both computer scientists and law experts in order to provide a clear and usable framework to guide privacy policy definitions, and to support the decisions of judges and lawyers.

1.5 Anonymity in Complex Data

We have seen how many research efforts have focused on privacy-preserving data mining and data publishing. Most research, however, addresses the

anonymity problems in the context of general tabular data, while relatively little work has addressed more complex forms of data in specific domains, although this kind of data is growing rapidly: examples include social networking data, spatio-temporal data, query log data, and more. The analysis of these data is very interesting as they are semantically rich: such richness makes such data also very difficult to anonymize, because the extra semantics may offer unexpected means to the attacker to link data to background knowledge. Traditional techniques used for tabular datasets cannot be directly applied, so typically the standard approaches must be adjusted appropriately. A survey of techniques for anonymity of query log data is presented in [19]. In this work the author seeks to assess some anonymity techniques against three criteria: a) how well the technique protects privacy, b) how well the technique preserves the utility of the query logs, and c) how well the technique might be implemented as a user control. In [72] Zhou et al. propose a brief systematic review of the existing anonymity techniques for privacy preserving publishing of social network data. Another interesting work is presented in [47], where Malin introduces a computational method for the anonymization of a collection of person-specific DNA database sequences. The analysis of person-specific DNA sequences is important but poses serious challenges to the protection of the identities to which such sequences correspond.

In this section we focus our discussion on spatio-temporal data showing that in recent years some reasonable results have been obtained by solutions that consider the particular nature of these data. The increasing availability of spatio-temporal data is due to the diffusion of mobile devices (e.g., mobile phones, RFID devices and GPS devices) and of new applications, where the discovery of consumable, concise, and applicable knowledge is the key step. Clearly, in these applications privacy is a concern, since a pattern can reveal the behavior of a group of few individuals, compromising their privacy. Spatio-temporal datasets present a new challenge for the privacy-preserving data mining community because of their spatial and temporal characteristics.

Standard approaches developed for tabular data do not work for spatio-temporal datasets. For example, randomization techniques, discussed above, which modify a dataset to guarantee respondents' privacy while preserving data utility for analyses, are not applicable on spatio-temporal data, due to their particular nature. Therefore, alternative solutions have been suggested: some of them belong to the category of *confusion-based algorithms* others belong to the category of approaches of *k-anonymity for location position collection*. All these techniques try to guarantee location privacy for trajectories.

The approaches in [35, 39, 40, 27] belong to the first category and provide confusion/obfuscation algorithms to prevent an attacker from tracking a complete user trajectory. The main idea is to modify true trajectories or generate fake trajectories in order to confuse the attacker. In [13, 14, 34, 16] authors presented techniques belonging to the second category. The main aim of these techniques is to preserve the anonymity of a user obscuring his route. They use the notion of k-anonymity adapted for the spatio-temporal context.

k-anonymity is the most popular method for the anonymization of spatio-temporal data. It is often used both in the works on privacy issues in location-based services (LBSs) [15, 48] and in the works of anonymity of trajectories [1, 51, 69]. In the work presented in [1], the authors study the problem of privacy-preserving publishing of moving object databases. They propose the notion of (k, δ)-anonymity for moving object databases, where δ represents the possible location imprecision. In particular, this is a novel concept of k-anonymity based on co-localization that exploits the inherent uncertainty of the moving objects whereabouts. In this work authors also propose an approach, called *Never Walk Alone*, for obtaining a (k, δ)-anonymous moving object database. The method is based on trajectory clustering and spatial translation. In [51] Nergiz et al. address privacy issues regarding the identification of individuals in static trajectory datasets. They provide privacy protection by: (1) first enforcing k-anonymity, meaning every released information refers to at least k users/trajectories, (2) then reconstructing randomly a representation of the original dataset from the anonymization. Another approach based on the concept of k-anonymity is proposed in [56], where Pensa et al. present a framework for k-anonymization of sequences of regions/locations. The authors also propose an approach that is an instance of the proposed framework and that allows to publish protected datasets while preserving the data utility for sequential pattern mining tasks. This approach, called *BF-P2kA*, consists of three steps. During the first step, the sequences in the input dataset D are used to build a prefix tree, representing the dataset. The second step, given a minimum support threshold k, anonymizes the prefix tree. This means that sequences, whose support is less than k, are pruned from the prefix tree. Then, part of these infrequent sequences is re-appended in the prefix tree, by using the notion of longest common sub-sequence. The third and last step is to post-process the anonymized prefix tree, as obtained in the previous step, to generate the anonymized dataset of sequences D'. Yarovoy et al. in [69] study problem of k-anonymization of moving object databases for the purpose of their publication. They observe the fact that different objects in this context may have different quasi-identifiers and so, anonymization groups associated with different objects may not be disjoint. Therefore, a novel notion of k-anonymity based on spatial generalization is provided. In this work, authors propose two approaches that generate anonymity groups satisfying the novel notion of k-anonymity. These approaches are called *Extreme Union* and *Symmetric Anonymization*.

Finally, we also mention the very recent work [64], where Terrovitis and Mamoulis suggest a suppression-based algorithm that, given the head of a trajectory, reduces the probability of disclosing its tail. This work is based on the assumption that different attackers know different and disjoint portions of the trajectories and the data publisher knows the attacker's knowledge. So, the proposed solution is to suppress all the dangerous observations in the database.

1.6 Conclusion

In this chapter, we presented an overview of the main techniques for ensuring anonymity in data publishing and mining. In particular, we described the approaches proposed both by the data mining community and by the statistical disclosure control community. Often, the work done by these communities is very similar. Moreover, we presented an overview on how anonymity has been considered in the privacy jurisdiction of different countries. Finally, we concluded with a discussion about the anonymity issues in complex data, focusing our attention on spatio-temporal data. Usually, the anonymity techniques proposed for tabular data are not suitable for complex data, such as social networking data, spatio-temporal data, query log data and web log data; therefore new approaches have to be developed. We presented very recent works tackling the problem of anonymity in the particular context of spatio-temporal data.

What have we learned by this critical perspective? The main lesson is that if we strive to a general concept of anonymity in personal data, we are doomed to obtain rather weak definitions, both in the legal and in the data analysis fields. Talking in general of tabular data describing personal information leads to slippery concepts such "reasonableness," "disproportionate effort" referred to re-identification; from the analytical side, a clear trade-off for balancing analytical utility and risk of re-identification is missing. Researchers in data mining and statistic disclosure control seem to have learned this lesson, and the most recent results are focusing on specific, yet interesting, forms of data: the assumption of a particular data semantics helps in defining convincing background knowledge assumptions for adversarial attacks, and, in turn, precise countermeasures and formal protection models. This is a crucial step in identifying formal measures of privacy risk, that may in turn affect regulations, and provide quantifiable counterparts of the adjective "reasonable." We tried to provide one instance of this "purpose-oriented anonymity," with reference to mobility data: other examples can be found in the other chapters of this book.

Anonymity in data publishing and mining is a young, exciting research arena, with plenty of open issues, some of which, if solved, can have a great impact on society, and change the way information technology is perceived today.

Acknowledgments

Ruggero G. Pensa is co-funded by Regione Piemonte. Dino Pedreschi acknowledges support by Google, under the Google Research Award program.

References

[1] O. Abul, F. Bonchi, and M. Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *ICDE*, pages 376–385, 2008.

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

[3] C. C. Aggarwal. On randomization, public information and the curse of dimensionality. In *ICDE*, pages 136–145, 2007.

[4] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. Achieving anonymity via clustering. In *PODS*, pages 153–162. ACM, 2006.

[5] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *ICDT*, volume 3363 of *LNCS*, pages 246–258, 2005.

[6] D. Agrawal and C.C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *PODS*. ACM, 2001.

[7] R. Agrawal. Privacy and data mining. In *ECML/PKDD*, 2004.

[8] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD*, pages 439–450. ACM, 2000.

[9] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. Blocking anonymity threats raised by frequent itemset mining. In *ICDM*, pages 561–564. IEEE Computer Society, 2005.

[10] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. k-anonymous patterns. In *PKDD*, pages 10–21, 2005.

[11] M. Atzori, F. Bonchi, F. Giannotti, and D. Pedreschi. Anonymity preserving pattern discovery. *VLDB Journal*, 17(4):703–727, 2008.

[12] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE 2005*, pages 217–228, 2005.

[13] A. R. Beresford and F. Stajano. Location privacy in pervasive computing. *IEEE Pervasive Computing*, 2(1):46–55, 2003.

[14] A. R. Beresford and F. Stajano. Mix zones: user privacy in location-aware services. In *PerCom Workshops*, pages 127–131. IEEE Computer Society, 2004.

[15] S. Mascetti and C. Bettini. Preserving k-anonymity in spatio-temporal datasets and location-based services. First Italian workshop on PRIvacy and SEcurity (PRISE), Rome, June 2006.

[16] C. Bettini, X. S. Wang, and S. Jajodia. Protecting privacy against location-based personal identification. In *VLDB Workshop SDM 2005*, volume 3674 of *LNCS*, pages 185–199. Springer, 2005.

[17] J. Burridge, L. Franconi, S. Polettini, and J. Stander. A methodological framework for statistical disclosure limitation of business microdata. *Technical Report 1.1-D4*, CASC Project, 2002.

[18] K. Chen and L. Liu. Privacy preserving data classification with rotation perturbation. In *ICDM*, pages 589–592. IEEE Computer Society, 2005.

[19] A. Cooper. A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Trans. Web*, 2(4):1–27, 2008.

[20] L. H. Cox and J. J. Kim. Effects of rounding on the quality and confidentiality of statistical data. In *Privacy in Statistical Databases*, volume 4302 of *LNCS*, pages 48–56, 2006.

[21] R. A. Dandekar, J. Domingo-Ferrer, and F. Sebé. Lhs-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In *Inference Control in Statistical Databases*, pages 153–162. Springer-Verlag, 2002.

[22] A. G. DeWaal and L. C. R. J. Willenborg. Global recodings and local suppressions in microdata sets. In *Proceedings of Statistics Canada Symposium'95*, page 121132, 1995.

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

[24] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE TKDE*, 14(1):189–201, 2002.

[25] J. Domingo-Ferrer and V. Torra. *A Quantitative Comparison of Disclosure Control Methods for Microdata*, pages 111–133. Elsevier, 2001.

[26] J. Domingo-Ferrer and V. Torra. Distance-based and probabilistic record linkage for re-identification of records with categorical variables. *Butlletí de l'ACIA*, 28:243–250, 2002.

[27] M. Duckham and L. Kulik. A formal model of obfuscation and negotiation for location privacy. In *Pervasive*, pages 152–170, 2005.

[28] A. V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *SIGKDD*, pages 217–228. ACM, 2002.

[29] S. E. Fienberg and J. McIntyre. Data swapping: Variations on a theme by dalenius and reiss. In *Privacy in Statistical Databases*, volume 3050 of *LNCS*, pages 14–29. Springer, 2004.

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

[31] A. Florian. An efficient sampling scheme: Updated latin hypercube sampling. *J. Probabilistic Engineering Mechanics*, 7(2):123–130, 1992.

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

[33] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, pages 205–216, 2005.

[34] M. Gruteser and X. Liu. Protecting privacy in continuous location-tracking applications. *IEEE Security & Privacy*, 2(2):28–34, 2004.

[35] B. Hoh and M. Gruteser. Protecting location privacy through path confusion. In *SECURECOMM '05*, pages 194–205. IEEE Computer Society, 2005.

[36] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, pages 279–288. ACM, 2002.

[37] W. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. *Contemporary Math.*, 26:189–206, 1984.

[38] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques. In *ICDM '03*, page 99. IEEE Computer Society, 2003.

[39] H. Kido, Y. Yanagisawa, and T. Satoh. An anonymous communication technique using dummies for location-based services. In *International Conference on Pervasive Services*, pages 88–97. IEEE Computer Society, 2005.

[40] H. Kido, Y.Yanagisawa, and T. Satoh. Protection of location privacy using dummies for location-based services. In *ICDE Workshops*, page 1248, 2005.

[41] J. J. Kim. A method for limiting disclosure in microdata based on random noise and transformation. In *Survey Research Method Section, American Statistical Association*, pages 370–374, 1986.

[42] P. Kooiman, L. Willenborg, and J. Gouweleeuw. Pram: A method for disclosure limitation of microdata. Research paper no. 9705, 1997.

[43] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD*, pages 49–60. ACM, 2005.

[44] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, pages 106–115. IEEE, 2007.

[45] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE TKDE*, 18(1):92–106, 2006.

[46] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE*, page 24. IEEE Computer Society, 2006.

[47] B. Malin. Protecting DNA sequence anonymity with generalization lattices. *Methods of Information in Medicine*, 44(5):687–692, 2005.

[48] S. Mascetti, C. Bettini, X. S. Wang, and S. Jajodia. k-anonymity in databases with timestamped data. In *TIME*, pages 177–186, 2006.

[49] J. M. Mateo-Sanz, A. Martinez-Balleste, and J. Domingo-Ferrer. Fast generation of accurate synthetic microdata. In *Privacy in Statistical Databases, vol.3050 of LNCS*, pages 298–306. Springer, 2004.

[50] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS '04*, pages 223–228. ACM, 2004.

[51] M. E. Nergiz, M. Atzori, and Y. Saygin. Perturbation-driven anonymization of trajectories. Technical Report 2007-TR-017, ISTI-CNR, Pisa, Italy, 2007. 10 pages.

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

[53] S. R. M. Oliveira and O. R. Zaiane. Privacy preserving clustering by data transformation. In *SBBD*, pages 304–318, 2003.

[54] S. R. M. Oliveira and O. R. Zaiane. Data perturbation by rotation for privacypreserving clustering. *Technical Report* TR04-17, Department of Computing Science, University of Alberta, Edmonton, Canada, 2004.

[55] Federal Committee on Statistical Methodology. Statistical policy working paper 22, may 1994. Report on Statistical Disclosure Limitation Methodology.

[56] R. G. Pensa, A. Monreale, F. Pinelli, and D. Pedreschi. Pattern-preserving k-anonymization of sequences and its application to mobility data mining. In *PiLBA*, 2008.

[57] S. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In *VLDB*, pages 682–693, 2002.

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

[59] P. Samarati. Protecting respondents' identities in microdata release. *IEEE TKDE*, 13(6):1010–1027, 2001.

[60] A. C. Singh, F. Yu, and G. H. Dunteman. Massc: A new data mask for limiting statistical information loss and disclosure. In *The Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality*, pages 373–394, 2003.

[61] G. R. Sullivan. *The use of added error to avoid disclosure in microdata releases*. PhD thesis, Ames, IA, USA, 1989. Supervisor-Fuller, Wayne A.

[62] L. Sweeney. Uniqueness of simple demographics in the U.S. population. Technical report, Laboratory for International Data Privacy, Carnegie Mellon University, Pittsburgh, PA, 2000.

[63] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems*, 2002.

[64] M. Terrovitis and N. Mamoulis. Privacy preservation in the publication of trajectories. In *MDM*, pages 65–72, 2008.

[65] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM*, pages 249–256. IEEE Computer Society, 2004.

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

[67] W. E. Winkler. Using simulated annealing for k-anonymity. Technical Report 7, US Census Bureau.

[68] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, pages 139–150. ACM, 2006.

[69] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang. Anonymizing moving objects: how to hide a mob in a crowd? In *EDBT*, pages 72–83, 2009.

[70] J. Z. Zhan, S. Matwin, and L. Chang. Privacy-preserving collaborative association rule mining. In *DBSec*, pages 153–165, 2005.

[71] P. Zhang, Y. Tong, S. Tang, and D. Yang. Privacy preserving Naive Bayes classification. In *ADMA*, volume 3584 of *LNCS*, pages 744–752. Springer, 2005.

[72] B. Zhou, J. Pei, and W. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *SIGKDD Explor. Newsl.*, 10(2):12–22, 2008.