A Survey of Quantification of Privacy Preserving Data Mining Algorithms

Elisa Bertino, Dan Lin, and Wei Jiang

Abstract The aim of privacy preserving data mining (PPDM) algorithms is to extract relevant knowledge from large amounts of data while protecting at the same time sensitive information. An important aspect in the design of such algorithms is the identification of suitable evaluation criteria and the development of related benchmarks. Recent research in the area has devoted much effort to determine a trade-off between the right to privacy and the need of knowledge discovery. It is often the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria. Therefore, it is crucial to provide a comprehensive view on a set of metrics related to existing privacy preserving algorithms so that we can gain insights on how to design more effective measurement and PPDM algorithms. In this chapter, we review and summarize existing criteria and metrics in evaluating privacy preserving techniques.

1 Introduction

Privacy is one of the most important properties that an information system must satisfy. For this reason, several efforts have been devoted to incorporating privacy preserving techniques with data mining algorithms in order to prevent the disclosure of sensitive information during the knowledge discovery. The existing privacy preserving data mining techniques can be classified according to the following five different dimensions [32]: (i) data distribution (centralized or distributed); (ii) the modification applied to the data (encryption, perturbation, generalization, and so on) in or-

Elisa Bertino
Purdue University, 305 N. University St., West Lafayette, IN, USA, e-mail: bertino@cs.purdue.edu

Dan Lin
Purdue University, 305 N. University St., West Lafayette, IN, USA, e-mail: lindan@cs.purdue.edu

Wei Jiang
Purdue University, 305 N. University St., West Lafayette, IN, USA, e-mail: wjiang@cs.purdue.edu

der to sanitize them; (iii) the data mining algorithm which the privacy preservation technique is designed for; (iv) the data type (single data items or complex data correlations) that needs to be protected from disclosure; (v) the approach adopted for preserving privacy (heuristic or cryptography-based approaches). While heuristic-based techniques are mainly conceived for centralized datasets, cryptography-based algorithms are designed for protecting privacy in a distributed scenario by using encryption techniques. Heuristic-based algorithms recently proposed aim at hiding sensitive raw data by applying perturbation techniques based on probability distributions. Moreover, several heuristic-based approaches for hiding both raw and aggregated data through a hiding technique (k-anonymization, adding noises, data swapping, generalization and sampling) have been developed, first, in the context of association rule mining and classification and, more recently, for clustering techniques.

Given the number of different privacy preserving data mining (PPDM) techniques that have been developed in these years, there is an emerging need of moving toward standardization in this new research area, as discussed by Oliveira and Zaiane [23]. One step toward this essential process is to provide a quantification approach for PPDM algorithms to make it possible to evaluate and compare such algorithms. However, due to the variety of characteristics of PPDM algorithms, it is often the case that no privacy preserving algorithm exists that outperforms all the others on all possible criteria. Rather, an algorithm may perform better than another one on specific criteria like privacy level, data quality. Therefore, it is important to provide users with a comprehensive set of privacy preserving related metrics which will enable them to select the most appropriate privacy preserving technique for the data at hand, with respect to some specific parameters they are interested in optimizing [6].

For a better understanding of PPDM related metrics, we next identify a proper set of criteria and the related benchmarks for evaluating PPDM algorithms. We then adopt these criteria to categorize the metrics. First, we need to be clear with respect to the concept of “privacy” and the general goals of a PPDM algorithm. In our society the *privacy* term is overloaded, and can, in general, assume a wide range of different meanings. For example, in the context of the HIPAA[1] Privacy Rule, *privacy* means the individual’s ability to control who has the access to personal health care information. From the organizations point of view, *privacy* involves the definition of policies stating which information is collected, how it is used, and how customers are informed and involved in this process. Moreover, there are many other definitions of privacy that are generally related with the particular environment in which the privacy has to be guaranteed. What we need is a more generic definition, that can be instantiated to different environments and situations. From a philosophical point of view, Schoeman [26] and Walters [33] identify three possible definitions of privacy:

- Privacy as the right of a person to determine which personal information about himself/herself may be communicated to others.

[1] Health Insurance Portability and Accountability Act

- Privacy as the control over access to information about oneself.
- Privacy as limited access to a person and to all the features related to the person.

In three definitions, what is interesting from our point of view is the concept of "Controlled Information Release". From this idea, we argue that a definition of privacy that is more related with our target could be the following: "*The right of an individual to be secure from unauthorized disclosure of information about oneself that is contained in an electronic repository*". Performing a final tuning of the definition, we consider privacy as "*The right of an entity to be secure from unauthorized disclosure of sensible information that are contained in an electronic repository or that can be derived as aggregate and complex information from data stored in an electronic repository*". The last generalization is due to the fact that the concept of individual privacy does not even exist. As in [23] we consider two main scenarios.

The first is the case of a Medical Database where there is the need to provide information about diseases while preserving the patient identity. Another scenario is the classical "Market Basket" database, where the transactions related to different client purchases are stored and from which it is possible to extract some information in form of association rules like "If a client buys a product X, he/she will purchase also Z with y% probability". The first is an example where individual privacy has to be ensured by protecting from unauthorized disclosure sensitive information in form of specific data items related to specific individuals. The second one, instead, emphasizes how not only the raw data contained into a database must be protected, but also, in some cases, the high level information that can be derived from non sensible raw data need to protected. Such a scenario justifies the final generalization of our privacy definition. In the light of these considerations, it is, now, easy to define which are the main goals a PPDM algorithm should enforce:

1. A PPDM algorithm should have to prevent the discovery of sensible information.
2. It should be resistant to the various data mining techniques.
3. It should not compromise the access and the use of non sensitive data.
4. It should not have an exponential computational complexity.

Correspondingly, we identify the following set of criteria based on which a PPDM algorithm can be evaluated.

- *Privacy level* offered by a privacy preserving technique, which indicates how closely the sensitive information, that has been hidden, can still be estimated.
- *Hiding failure*, that is, the portion of sensitive information that is not hidden by the application of a privacy preservation technique;
- *Data quality* after the application of a privacy preserving technique, considered both as the quality of data themselves and the quality of the data mining results after the hiding strategy is applied;
- *Complexity*, that is, the ability of a privacy preserving algorithm to execute with good performance in terms of all the resources implied by the algorithm.

For the rest of the chapter, we first present details of each criteria through analyzing existing PPDM techniques. Then we discuss how to select proper metric under a specified condition. Finally, we summarize this chapter and outline future research directions.

2 Metrics for Quantifying Privacy Level

Before presenting different metrics related to privacy level, we need to take into account two aspects: (i) sensitive or private information can be contained in the original dataset; and (ii) private information that can be discovered from data mining results. We refer to the first one as data privacy and the latter as result privacy.

2.1 Data Privacy

In general, the quantification used to measure data privacy is the degree of uncertainty, according to which original private data can be inferred. The higher the degree of uncertainty achieved by a PPDM algorithm, the better the data privacy is protected by this PPDM algorithm. For various types of PPDM algorithms, the degree of uncertainty is estimated in different ways. According to the adopted techniques, PPDM algorithms can be classified into two main categories: heuristic-based approaches and cryptography-based approaches. Heuristic-based approaches mainly include four sub-categories: additive noise, multiplicative noise, k-anonymization, and statistical disclosure control based approaches. In what follows, we survey representative works of each category of PPDM algorithms and review the metrics used by them.

2.1.1 Additive-Noise-based Perturbation Techniques

The basic idea of the additive-noise-based perturbation technique is to add random noise to the actual data. In [2], Agrawal and Srikant uses an additive-noise-based technique to perturb data. They then estimate the probability distribution of original numeric data values in order to build a decision tree classifier from perturbed training data. They introduce a quantitative measure to evaluate the amount of privacy offered by a method and evaluate the proposed method against this measure. The privacy is measured by evaluating how closely the original values of a modified attribute can be determined. In particular, if the perturbed value of an attribute can be estimated, with a confidence c, to belong to an interval $[a,b]$, then the privacy is estimated by $(b-a)$ with confidence c. However, this metric does not work well because it does not take into account the distribution of the original data along with the perturbed data. Therefore, a metric that considers all the informative content of data available to the user is needed. Agrawal and Aggarwal [1] address this problem by introducing a new privacy metric based on the concept of information entropy. More specifically, they propose an Expectation Maximization (EM) based algorithm for distribution reconstruction, which converges to the maximum likelihood estimate of the original distribution on the perturbed data. The measurement of privacy given by them considers the fact that both the perturbed individual record and the reconstructed distribution are available to the user as well as the perturbing distribution, as it is specified in [10]. This metric defines the average conditional privacy of an attribute A given other information, modeled with a random variable B, as $2^{h(A|B)}$, where $h(A|B)$ is the conditional differential entropy of A given B representing a measure of uncertainty inherent in the value of A, given the value of B.

Another additive-noise-based perturbation technique is by Rivzi and Haritsa [24]. They propose a distortion method to pre-process the data before executing the mining process. Their privacy measure deals with the probability with which the user's distorted entries can be reconstructed. Their goal is to ensure privacy at the level of individual entries in each customer tuple. In other words, the authors estimate the probability that a given 1 or 0 in the true matrix representing the transactional database can be reconstructed, even if for many applications the 1's and 0's values do not need the same level of privacy.

Evfimievski et al. [11] propose a framework for mining association rules from transactions consisting of categorical items, where the data has been randomized to preserve privacy of individual transactions, while ensuring at the same time that only true associations are mined. They also provide a formal definition of privacy breaches and a class of randomization operators that are much more effective in limiting breaches than uniform randomization. According to Definition 4 from [11], an itemset A results in a privacy breach of level ρ if the probability that an item in A belongs to a non randomized transaction, given that A is included in a randomized transaction, is greater than or equal to ρ. In some scenarios, being confident that an item not present in the original transaction may also be considered a privacy breach. In order to evaluate the privacy breaches, the approach taken by Evfimievski et al. is to count the occurrences of an itemset in a randomized transaction and in its sub-items in the corresponding non randomized transaction. Out of all sub-items of an itemset, the item causing the worst privacy breach is chosen. Then, for each combination of transaction size and itemset size, the worst and the average value of this breach level are computed over all frequent itemsets. The itemset size giving the worst value for each of these two values is selected.

Finally, we introduce a universal measure of data privacy level, proposed by Bertino et al. in [6]. The measure is developed based on [1]. The basic concept used by this measure is information entropy, which is defined by Shannon [27]: let X be a random variable which takes on a finite set of values according to a probability distribution $p(x)$. Then, the entropy of this probability distribution is defined as follows:

$$h(X) = -\sum p(x) \log_2(p(x)) \tag{1}$$

or, in the continuous case:

$$h(X) = -\int f(x) \log_2(f(x)) dx \tag{2}$$

where $f(x)$ denotes the density function of the continuous random variable x. Information entropy is a measure of how much "choice" is involved in the selection of an event or how uncertain we are of its outcome. It can be used for quantifying the amount of information associated with a set of data. The concept of "information associated with data" can be useful in the evaluation of the privacy achieved by a PPDM algorithm. Because the entropy represents the information content of a datum, the entropy after data sanitization should be higher than the entropy before the sanitization. Moreover the entropy can be assumed as the evaluation of the uncertain forecast level of an event which in our context is evaluation of the right value of a datum. Consequently, the level of privacy inherent in an attribute X, given some information modeled by Y, is defined as follows:

$$\Pi(X|Y) = 2^{-\int f_{X,Y}(x,y)\log_2 f_{X|Y=y}(x))dxdy} \tag{3}$$

The privacy level defined in equation 3 is very general. In order to use it in the different PPDM contexts, it needs to be refined in relation with some characteristics like the type of transactions, the type of aggregation and PPDM methods. In [6], an example of instantiating the entropy concept to evaluate the privacy level in the context of "association rules" is presented.

However, it is worth noting that the value of the privacy level depends not only on the PPDM algorithm used, but also on the knowledge that an attacker has about the data before the use of data mining techniques and the relevance of this knowledge in the data reconstruction operation. This problem is underlined, for example, in [29, 30]. In [6], this aspect is not considered, but it is possible to introduce assumptions on attacker knowledge by properly modeling Y.

2.1.2 Multiplicative-Noise-based Perturbation Techniques

According to [16], additive random noise can be filtered out using certain signal processing techniques with very high accuracy. To avoid this problem, random projection-based multiplicative perturbation techniques has been proposed in [19]. Instead of adding some random values to the actual data, random matrices are used to project the set of original data points to a randomly chosen lower-dimensional space. However, the transformed data still preserves much statistical aggregates regarding the original dataset so that certain data mining tasks (e.g., computing inner product matrix, linear classification, K-means clustering and computing Euclidean distance) can be performed on the transformed data in a distributed environment (data are either vertically partitioned or horizontally partitioned) with small errors.

In addition, this approach provides a high degree of privacy regarding the original data. As analyzed in the paper, even if the random matrix (i.e., the multiplicative noise) is disclosed, it is impossible to find the exact values of the original dataset, but finding approximation of the original data is possible. The variance of the approximated data is used as privacy measure.

Oliveira and Zaiane [22] also adopt a multiplicative-noise-based perturbation technique to perform a clustering analysis while ensuring at the same time privacy preservation. They have introduced a family of geometric data transformation methods where they apply a noise vector to distort confidential numerical attributes. The privacy ensured by such techniques is measured as the variance difference between the actual and the perturbed values. This measure is given by $Var(X-Y)$, where X represents a single original attribute and Y the distorted attribute. This measure can be made scale invariant with respect to the variance of X by expressing security as $Sec = Var(X-Y)/Var(X)$.

2.1.3 *k*-Anonymization Techniques

The concept of k-anonymization is introduced by Samarati and Sweeney in [25, 28]. A database is k-anonymous with respect to quasi-identifier attributes (a set of attributes that can be used with certain external information to identify a specific individual) if there exist at least k transactions in the database having the same values according to the quasi-identifier attributes. In practice, in order to protect sensitive dataset T, before releasing T to the public, T is converted into a new dataset T^* that guarantees the k-anonymity property for a sensible attribute by performing some value generalizations on quasi-identifier attributes. Therefore, the degree of uncertainty of the sensitive attribute is at least $1/k$.

2.1.4 Statistical-Disclosure-Control-based Techniques

In the context of statistical disclosure control, a large number of methods have been developed to preserve individual privacy when releasing aggregated statistics on data. To anonymize the released statistics from those data items such as person, household and business, which can be used to identify an individual, not only features described by the statistics but also related information publicly available need to be considered [35]. In [7] a description of the most relevant perturbation methods proposed so far is presented. Among these methods specifically designed for continuous data, the following masking techniques are described: additive noise, data distortion by probability distribution, resampling, microaggregation, rank swapping, etc. For categorical data both perturbative and non-perturbative methods are presented. The top-coding and bottom-coding techniques are both applied to ordinal categorical variables; they recode, respectively, the first/last p values of a variable into a new category. The global-recoding technique, instead, recodes the p lowest frequency categories into a single one.

The privacy level of such method is assessed by using the *disclosure risk*, that is, the risk that a piece of information be linked to a specific individual. There are several approaches to measure the disclosure risk. One approach is based on the computation of the distance-based record linkage. An intruder is assumed to try to link the masked dataset with the external dataset using the key variables. The distance between records in the original and the masked datasets is computed. A record in the masked dataset is labelled as “linked” or “linked to 2nd nearest” if the nearest or 2nd nearest record in the original dataset turns out to be the corresponding original record. Then the disclosure risk is computed as the percentage of “linked” and “linked to 2nd nearest”. The second approach is based on the computation of the probabilistic record linkage. The linear sum assignment model is used to ‘pair’ records in the original file and the masked file. The percentage of correctly paired records is a measure of disclosure risk. Another approach computes rank intervals for the records in the masked dataset. The proportion of original values that fall into the interval centered around their corresponding masked value is a measure of disclosure risk.

2.1.5 Cryptography-based Techniques

The cryptography-based technique usually guarantees very high level of data privacy. In [14], Kantarcioglu and Clifton address the problem of secure mining of association rules over horizontally partitioned data, using cryptographic techniques to minimize the information shared. Their solution is based on the assumption that each party first encrypts its own itemsets using commutative encryption, then the already encrypted itemsets of every other party. Later on, an initiating party transmits its frequency count, plus a random value, to its neighbor, which adds its frequency count and passes it on to other parties. Finally, a secure comparison takes place between the final and initiating parties to determine if the final result is greater than the threshold plus the random value.

Another cryptography-based approach is described in [31]. Such approach addresses the problem of association rule mining in vertically partitioned data. In other words, its aim is to determine the item frequency when transactions are split across different sites, without revealing the contents of individual transactions. The security of the protocol for computing the scalar product is analyzed.

Though cryptography-based techniques can well protect data privacy, they may not be considered good with respect to other metrics like efficiency that will be discussed in later sections.

2.2 Result Privacy

So far, we have seen privacy metrics related to the data mining process. Many data mining tasks produce aggregate results, such as Bayesian classifiers. Although it is possible to protect sensitive data when a classifier is constructed, can this classifier be used to infer sensitive data values? In other words, do data mining results violate privacy? This issue has been analyzed and a framework is proposed in [15] to test if a classifier C creates an inference channel that could be adopted to infer sensitive data values.

The framework considers three types of data: public data (P), accessible to every one including the adversary; private/sensitive data (S), must be protected and unknown to the adversary; unknown data (U), not known to the adversary, but the release of this data might cause privacy violation. The framework assumes that S depends only on P and U, and the adversary has at most t data samples of the form (p_i, s_i). The approach to determine whether an inference channel exists is comprised of two steps. First, a classifier C_1 is built on the t data samples. To evaluate the impact of C, another classifier C_2 is built based on the same t data samples plus the classifier C. If the accuracy of C_2 is significantly better than C_1, we can say that C provides an inference channel for S.

Classifier accuracy is measured based on Bayesian classification error. Suppose we have a dataset $\{x_1, \ldots, x_n\}$, and we want to classify x_i into m classes labelled as $\{1, \ldots, m\}$. Given a classifier C:

$$C : x_i \to C(x_i) \in \{1, \ldots, m\}, \quad i = 1, \ldots, n$$

The classifier accuracy for C is defined as:

$$\sum_{j=1}^{m} Pr(C(x_i) \neq j | z = j) Pr(z = j)$$

where z is the actual class label of x_i. Since cryptography-based PPDM techniques usually produce the same results as those mined from the original dataset, analyzing privacy implications from the mining results is particular important to this class of techniques.

3 Metrics for Quantifying Hiding Failure

The percentage of sensitive information that is still discovered, after the data has been sanitized, gives an estimate of the *hiding failure* parameter. Most of the developed privacy preserving algorithms are designed with the goal of obtaining zero hiding failure. Thus, they hide all the patterns considered sensitive. However, it is well known that the more sensitive information we hide, the more non-sensitive information we miss. Thus, some PPDM algorithms have been recently developed which allow one to choose the amount of sensitive data that should be hidden in order to find a balance between privacy and knowledge discovery. For example, in [21], Oliveira and Zaiane define the *hiding failure* (HF) as the percentage of restrictive patterns that are discovered from the sanitized database. It is measured as follows:

$$HF = \frac{\#R_P(D')}{\#R_P(D)} \tag{4}$$

where $\#R_P(D)$ and $\#R_P(D')$ denote the number of restrictive patterns discovered from the original data base D and the sanitized database D' respectively. Ideally, HF

should be 0. In their framework, they give a specification of a *disclosure threshold* ϕ, representing the percentage of sensitive transactions that are not sanitized, which allows one to find a balance between the hiding failure and the number of misses. Note that ϕ does not control the *hiding failure* directly, but indirectly by controlling the proportion of sensitive transactions to be sanitized for each restrictive pattern.

Moreover, as pointed out in [32], it is important not to forget that intruders and data terrorists will try to compromise information by using various data mining algorithms. Therefore, a PPDM algorithm developed against a particular data mining techniques that assures privacy of information, may not attain similar protection against all possible data mining algorithms. In order to provide for a complete evaluation of a PPDM algorithm, we need to measure its hiding failure against data mining techniques which are different from the technique that the PPDM algorithm has been designed for. The evaluation needs the consideration of a class of data mining algorithms which are significant for our test. Alternatively, a formal framework can be developed that upon testing of a PPDM algorithm against pre-selected data sets, we can transitively prove privacy assurance for the whole class of PPDM algorithms.

4 Metrics for Quantifying Data Quality

The main feature of the most PPDM algorithms is that they usually modify the database through insertion of false information or through the blocking of data values in order to hide sensitive information. Such perturbation techniques cause the decrease of the data quality. It is obvious that the more the changes are made to the database, the less the database reflects the domain of interest. Therefore, data quality metrics are very important in the evaluation of PPDM techniques. Since the data is often sold for making profit, or shared with others in the hope of leading to innovation, data quality should have an acceptable level according also to the intended data usage. If data quality is too degraded, the released database is useless for the purpose of knowledge extraction.

In existing works, several data quality metrics have been proposed that are either generic or data-use-specific. However, currently, there is no metric that is widely accepted by the research community. Here we try to identify a set of possible measures that can be used to evaluate different aspects of data quality. In evaluating the data quality after the privacy preserving process, it can be useful to assess both the *quality of the data* resulting from the PPDM process and the *quality of the data mining results*. The quality of the data themselves can be considered as a general measure evaluating the state of the individual items contained in the database after the enforcement of a privacy preserving technique. The quality of the data mining results evaluates the alteration in the information that is extracted from the database after the privacy preservation process, on the basis of the intended data use.

4.1 Quality of the Data Resulting from the PPDM Process

The main problem with data quality is that its evaluation is relative [18], in that it usually depends on the context in which data are used. In particular, there are some aspects related to data quality evaluation that are heavily related not only with the PPDM algorithm, but also with the structure of the database, and with the meaning and relevance of the information stored in the database with respect to a well defined context. In the scientific literature data quality is generally considered a multi-dimensional concept that in certain contexts involves both objective and subjective parameters [3, 34]. Among the various possible parameters, the following ones are usually considered the most relevant:

- *Accuracy*: it measures the proximity of a sanitized value to the original value.
- *Completeness*: it evaluates the degree of missed data in the sanitized database.
- *Consistency*: it is related to the internal constraints, that is, the relationships that must hold among different fields of a data item or among data items in a database.

4.1.1 Accuracy

The accuracy is closely related to the *information loss* resulting from the hiding strategy: the less is the information loss, the better is the data quality. This measure largely depends on the specific class of PPDM algorithms. In what follows, we discuss how different approaches measure the accuracy.

As for heuristic-based techniques, we distinguish the following cases based on the modification technique that is performed for the hiding process. If the algorithm adopts a perturbation or a blocking technique to hide both raw and aggregated data, the information loss can be measured in terms of the dissimilarity between the original dataset D and the sanitized one D'. In [21], Oliveira and Zaiane propose three different methods to measure the *dissimilarity* between the original and sanitized databases. The first method is based on the difference between the frequency histograms of the original and the sanitized databases. The second method is based on computing the difference between the sizes of the sanitized database and the original one. The third method is based on a comparison between the contents of two databases. A more detailed analysis on the definition of dissimilarity is presented by Bertino et al. in [6]. They suggest to use the following formula in the case of transactional dataset perturbation:

$$Diss(D,D') = \frac{\sum_{i=1}^{n} |f_D(i) - f_{D'}(i)|}{\sum_{i=1}^{n} f_D(i)} \tag{5}$$

where i is a data item in the original database D and $f_D(i)$ is its frequency within the database, whereas *i*’ is the given data item after the application of a privacy preservation technique and $f_{D'}(i)$ is its new frequency within the transformed database D'. As we can see, the information loss is defined as the ratio between the sum of the absolute errors made in computing the frequencies of the items from a sanitized database and the sum of all the frequencies of items in the original database. The formula 5 can also be used for the PPDM algorithms which adopt a blocking technique for inserting into the dataset uncertainty about some sensitive data items or their correlations. The frequency of the item i belonging to the sanitized dataset D' is then given by the mean value between the minimum frequency of the data item i, computed by considering all the blocking values associated with it equal to zero, and the maximum frequency, obtained by considering all the question marks equal to one.

In case of data swapping, the information loss caused by an heuristic-based algorithm can be evaluated by a parameter measuring the *data confusion* introduced by the value swappings. If there is no correlation among the different database records, the *data confusion* can be estimated by the percentage of value replacements executed in order to hide specific information.

For the multiplicative-noise-based approaches [19], the quality of the perturbed data depends on the size of the random projection matrix. In general, the error bound of the inner product matrix produce by this perturbation technique is 0 on average and the variance is bounded by the inverse of the dimensionality of the reduced space. In other words, when the dimensionality of the random projection matrix is close to that of the original data, the result of computing the inner product matrix based on the transformed or projected data is also close to the actual value. Since inner product is closely related to many distance-based metrics (e.g., Euclidean distance, cosine angle of two vectors, correlation coefficient of two vectors, etc), the analysis on error bound has direct impact on the mining results if these data mining tasks adopt certain distance-based metrics.

If the data modification consists of aggregating some data values, the information loss is given by the loss of detail in the data. Intuitively, in this case, in order to perform the hiding operation, the PPDM algorithms use some type of “Generalization or Aggregation Scheme” that can be ideally modeled as a tree scheme. Each cell modification applied during the sanitization phase using the Generalization tree introduces a data perturbation that reduces the general accuracy of the database. As in the case of the k-anonymity algorithm presented in [28], we can use the following formula. Given a database T with N_A fields and N transactions, if we identify as generalization scheme a domain generalization hierarchy GT with a depth h, it is possible to measure the *information loss* (IL) of a sanitized database T^* as:

$$IL(T^*) = \frac{\sum_{i=1}^{i=N_A} \sum_{j=1}^{i=N} \frac{h}{|GT_{Ai}|}}{|T| * |N_A|} \tag{6}$$

where $\frac{h}{|GT_{Ai}|}$ represent the detail loss for each cell sanitized. For hiding techniques based on sampling approach, the quality is obviously related to the size of the considered sample and, more generally, on its features.

There are some other precision metrics specifically designed for k-anonymization approaches. One of the earliest data quality metrics is based on the height of generalization hierarchies [25]. The height is the number of times the original data value has been generalized. This metric assumes that a generalization on the data represents an information loss on the original data value. Therefore, data should be generalized as fewer steps as possible to preserve maximum utility. However, this metric does not take into account that not every generalization steps are equal in the sense of information loss.

Later, Iyengar [13] proposes a general *loss metric* (*LM*). Suppose T is a data table with n attributes. The *LM* metric is thought as the average information loss of all data cells of a given dataset, defined as follows:

$$LM(T^*) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{f(T^*[i][j])-1}{g(A_i)-1}}{|T| \cdot n} \tag{7}$$

In equation 7, T^* is the anonymized table of T, f is a function that given a data cell value $T^*[i][j]$, returns the number of distinct values that can be generalized to $T^*[i][j]$, and g is a function that given an attribute A_i, returns the number of distinct values of A_i.

The next metric, *classification metric* (*CM*), is introduced by Iyengar [13] to optimize a k-anonymous dataset for training a classifier. It is defined as the sum of the individual penalties for each row in the table normalized by the total number of rows N.

$$CM(T^*) = \frac{\sum_{all\ rows} penalty(row\ r)}{N} \tag{8}$$

The penalty value of row r is 1, i.e., row r is penalized, if it is suppressed or if its class label is not the majority class label of its group. Otherwise, the penalty value of row r is 0. This metric is particularly useful when we want to build a classifier over anonymous data.

Another interesting metric is the *discernibility metric*(*DM*) proposed by Bayado and Agrawal [4]. This discernibility metric assigns a penalty to each tuple based on how many tuples in the transformed dataset are indistinguishable from it. Let t be a tuple from the original table T, and let $G_{T^*}(t)$ be the set of tuples in an anonymized table T^* indistinguishable from t or the set of tuples in $T*$ equivalent to the anonymized value of t. Then *DM* is defined as follows:

$$DM(T^*) = \sum_{t \in T} |G_{T^*}(t)| \tag{9}$$

Note that if a tuple t has been suppressed, the size of $G_{T^*}(t)$ is the same as the size of T^*. In many situation, suppressions are considered to be most expensive in the sense of information loss. Thus, to maximize data utility, tuple suppression should be avoided whenever possible.

For any given metric M, if $M(T) > M(T')$, we say T has a higher information loss, or is less precise, than b. In other words, data quality of T is worse than that of T'. Is this true for all metrics? What is a good metric? It is not easy to answer these kinds of questions. As shown in [20], *CM* works better than *LM* in classification application. In addition, *LM* is better for association rule mining. It is apparent that to judge how good a particular metric is, we need to associate our judgement with specific applications (e.g., classification, mining association rules).

The *CM* metric and the information gain privacy loss ratio [12] are more interesting measure of utility because it considers the possible application for the data. Nevertheless, it is unclear what to do if we want to build classifiers on various attributes. In addition, these two metrics only work well if the data are intended to be used for building classifiers. Is there a utility metric that works well for various applications? Having this in mind, Kifer [17] proposes a utility measure related to Kullback-Leibler divergence. In theory, using this measure, *better* anonymous datasets (for different applications) can be produced. Researchers have measured the utility of the resulting anonymous datasets. Preliminary results show that this metric works well in practical applications.

For the statistical-based perturbation techniques which aim to hide the values of a confidential attribute, the information loss is basically the lack of precision in estimating the original distribution function of the given attribute. As defined in [1], the information loss incurred during the reconstruction of estimating the density function $f_X(x)$ of the attribute X, is measured by computing the following value:

$$I(f_X, \widehat{f}_X) = \frac{1}{2} E \left[\int_{\Omega_X} \left| f_X(x) - \widehat{f}_X(x) \right| dx \right] \tag{10}$$

that is, half of the expected value of L_1 norm between $f_X(x)$ and $\widehat{f}_X(x)$, which are the density distributions respectively before and after the application of the privacy preserving technique.

When considering the cryptography-based techniques which are typically employed in distributed environments, we can observe that they do not use any kind of perturbation techniques for the purpose of privacy preserving. Instead, they use the cryptographic techniques to assure data privacy at each site by limiting the information shared by all the sites. Therefore, the quality of data stored at each site is not compromised at all.

4.1.2 Completeness and Consistency

While the accuracy is a relatively general parameter in that it can be measured without strong assumptions on the dataset analyzed, the completeness is not so general. For example, in some PPDM strategies, e.g. blocking, the completeness evaluation is not significant. On the other hand, the consistency requires to determine all the relationships that are relevant for a given dataset.

In [5], Bertino et al. propose a set of evaluation parameters including the completeness and consistency evaluation. Unlike other techniques, their approach takes into account two more important aspects: relevance of data and structure of database. They provide a formal description that can be used to magnify the aggregate information of interest for a target database and the relevance of data quality properties of each aggregate information and for each attribute involved in the aggregate information. Specifically, the completeness lack (denoted as CML) is measured as follows:

$$CML = \sum_{i=0}^{n} (DMG.N_i.CV \cdot DMG.N_i.CW) \tag{11}$$

In equation 11, DMG is an oriented graph where each node N_i is an attribute class. CV is the completeness value and CW is the consistency value. The consistency lack (denoted as CSL) is given by the number of constraint violations occurred in all the sanitized transaction multiplied by the weight associated with every constraints.

$$CSL = \sum_{i=0}^{n} (DMG.SC_i.csv \cdot DMG.SC_i.cw) + \sum_{j=0}^{m} (DMG.CC_j.csv \cdot DMG.CC_j.cw) \tag{12}$$

In equation 12, *csv* indicates the number of violations, *cw* is the weight of the constraint, SC_i describes a simple constraint class, and CC_j describes a complex constraint class.

4.2 Quality of the Data Mining Results

In some situations, it can be useful and also more relevant to evaluate the quality of the data mining results after the sanitization process. This kind of metric is strictly related to the use the data are intended for. Data can be analyzed in order to mine information in terms of associations among single data items or to classify existing data with the goal of finding an accurate classification of new data items, and so on. Based on the intended data use, the information loss is measured with a specific metric, depending each time on the particular type of knowledge model one aims to extract.

If the intended data usage is data clustering, the information loss can be measured by the percentage of legitimate data points that are not well-classified after the sanitization process. As in [22], a misclassification error M_E is defined to measure the information loss.

$$M_E = \frac{1}{N} \sum_{i=1}^{k} (|Cluster_i(D)| - |Cluster_i(D')|) \tag{13}$$

where N represents the number of points in the original dataset, k is the number of clusters under analysis, and $|Cluster_i(D)|$ and $|Cluster_i(D')|$ represent the number of legitimate data points of the ith cluster in the original dataset D and the sanitized dataset D' respectively. Since a privacy preserving technique usually modify data for the sanitization purpose, the parameters involved in the clustering analysis is almost inevitably affected. In order to achieve high clustering quality, it is very important to keep the clustering results as consistent as possible before and after the application of a data hiding technique.

When quantifying information loss in the context of the other data usages, it is useful to distinguish between: *lost information* representing the percentage of non-sensitive patterns (i.e., association, classification rules) which are hidden as side-effect of the hiding process; and the *artifactual information* representing the percentage of artifactual patterns created by the adopted privacy preserving technique. For example, in [21], Oliveira and Zaiane define two metrics *misses cost* and *artifactual pattern* which are corresponding to *lost information* and *artifactual information* respectively. In particular, misses cost measures the percentage of non-restrictive patterns that are hidden after the sanitization process. This happens when some non-restrictive patterns lose support in the database due to the sanitization process. The misses cost (MC) is computed as follows:

$$MC = \frac{\# \sim R_P(D) - \# \sim R_P(D')}{\# \sim R_P(D)} \tag{14}$$

where $\# \sim R_P(D)$ and $\# \sim R_P(D')$ denote the number of non-restrictive patterns discovered from the original database D and the sanitized database D' respectively. In the best case, MC should be 0%. Notice that there is a compromise between the misses cost and the hiding failure in their approach. The more restrictive patterns they hide, the more legitimate patterns they miss. The other metric, artifactual pattern (AP), is measured in terms of the percentage of the discovered patterns that are artifacts. The formula is:

$$AP = \frac{|P'| - |P \bigcap P'|}{P'} \tag{15}$$

where $|X|$ denotes the cardinality of X. According to their experiments, their approach does not have any artifactual patterns, i.e., AP is always 0.

In case of association rules, the lost information can be modeled as the set of non-sensitive rules that are accidentally hidden, referred to as `lost rules`, by the privacy preservation technique, the artifactual information, instead, represents the set of new rules, also known as `ghost rules`, that can be extracted from the database after the application of a sanitization technique.

Similarly, if the aim of the mining task is data classification, e.g. by means of decision trees inductions, both the lost and artifactual information can be quantified by means of the corresponding lost and ghost association rules derived by the classification tree. These measures allow one to evaluate the high level information that are extracted from a database in form of the widely-used inference rules before and after the application of a PPDM algorithm.

It is worth noting that for most cryptography-based PPDM algorithms, the data mining results are the same as that produced from unsanitized data.

5 Complexity Metrics

The *complexity* metric measures the efficiency and scalability of a PPDM algorithm. Efficiency indicates whether the algorithm can be executed with good performance, which is generally assessed in terms of space and time. Space requirements are assessed according to the amount of memory that must be allocated in order to implement the given algorithm.

For the evaluation of time requirements, there are several approaches. The first approach is to evaluate the CPU time. For example, in [21], they first keep constant both the size of the database and the set of restrictive patterns, and then increase the size of the input data to measure the CPU time taken by their algorithm. An alternative approach would be to evaluate the time requirements in terms of the computational cost. In this case, it is obvious that an algorithm having a polynomial complexity is more efficient than another one with exponential complexity. Sometimes, the time requirements can even be evaluated by counting the average number of operations executed by a PPDM algorithm. As in [14], the performance is measured in terms of the number of encryption and decryption operations required by the specific algorithm. The last two measures, i.e. the computational cost and the average number of operations, do not provide an absolute measure, but they can be considered in order to perform a fast comparison among different algorithms.

In case of distributed algorithms, especially the cryptography-based algorithms (e.g. [14, 31]), the time requirements can be evaluated in terms of communication cost during the exchange of information among secure processing. Specifically, in [14], the communication cost is expressed as the number of messages exchanged among the sites, that are required by the protocol for securely counting the frequency of each rule.

Scalability is another important aspect to assess the performance of a PPDM algorithm. In particular, scalability describes the efficiency trends when data sizes increase. Such parameter concerns the increase of both performance and storage requirements as well as the costs of the communications required by a distributed technique with the increase of data sizes.

Due to the continuous advances in hardware technology, large amounts of data can now be easily stored. Databases along with data warehouses today store and manage amounts of data which are increasingly large. For this reason, a PPDM algorithm has to be designed and implemented with the capability of handling huge datasets that may still keep growing. The less fast is the decrease in the efficiency of a PPDM algorithm for increasing data dimensions, the better is its scalability. Therefore, the scalability measure is very important in determining practical PPDM techniques.

6 How to Select a Proper Metric

In previous section, we have discussed various types of metrics. An important question here is “which one among the presented metrics is the most relevant for a given privacy preserving technique?”.

Dwork and Nissim [9] make some interesting observations about this question. In particular, according to them in the case of statistical databases *privacy* is paramount, whereas in the case of distributed databases for which the privacy is ensured by using a secure multiparty computation technique *functionality* is of primary importance. Since a real database usually contains a large number of records, the performance guaranteed by a PPDM algorithm, in terms of time and communication requirements, is a not negligible factor, as well as its trend when increasing database size. The *data quality* guaranteed by a PPDM algorithm is, on the other hand, very important when ensuring privacy protection without damaging the data usability from the authorized users.

From the above observations, we can see that a trade-off metric may help us to state a unique value measuring the effectiveness of a PPDM algorithm. In [7], the score of a masking method provides a measure of the trade-off between disclosure risk and information loss. It is defined as an average between the ranks of disclosure risk and information loss measures, giving the same importance to both metrics. In [8], a *R-U* confidentiality map is described that traces the impact on disclosure risk *R* and data utility *U* of changes in the parameters of a disclosure limitation method which adopts an additive noise technique. We believe that an index assigning the same importance to both the data quality and the degree of privacy ensured by a PPDM algorithm is quite restrictive, because in some contexts one of these parameters can be more relevant than the other. Moreover, in our opinion the other parameters, even less relevant ones, should be also taken into account. The efficiency and scalability measures, for instance, could be discriminating factors in choosing among a set of PPDM algorithms that ensure similar degrees of privacy and data utility. A *weighted mean* could be, thus, a good measure for evaluating by means of a unique value the quality of a PPDM algorithm.

7 Conclusion and Research Directions

In this chapter, we have surveyed different approaches used in evaluating the effectiveness of privacy preserving data mining algorithms. A set of criteria is identified, which are *privacy level*, *hiding failure*, *data quality* and *complexity*. As none of the existing PPDM algorithms can outperform all the others with respect to all the criteria, we discussed the importance of certain metrics for each specific type of PPDM algorithms, and also pointed out the goal of a good metric.

There are several future research directions along the way of quantifying a PPDM algorithm and its underneath application or data mining task. One is to develop a comprehensive framework according to which various PPDM algorithms can be evaluated and compared. It is also important to design good metrics that can better reflect the properties of a PPDM algorithm, and to develop benchmark databases for testing all types of PPDM algorithms.

References

1. Agrawal, D., Aggarwal, C.C.: On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principle of Database System, pp. 247–255. ACM (2001)
2. Agrawal, R., Srikant, R.: Privacy preserving data mining. In: Proceeedings of the ACM SIGMOD Conference of Management of Data, pp. 439–450. ACM (2000)
3. Ballou, D., Pazer, H.: Modelling data and process quality in multi input, multi output information systems. Management science **31**(2), 150–162 (1985)
4. Bayardo, R., Agrawal, R.: Data privacy through optimal k-anonymization. In: Proc. of the 21st Int'l Conf. on Data Engineering (2005)
5. Bertino, E., Fovino, I.N.: Information driven evaluation of data hiding algorithms. In: 7th Internationa Conference on Data Warehousing and Knowledge Discovery, pp. 418–427 (2005)
6. Bertino, E., Fovino, I.N., Provenza, L.P.: A framework for evaluating privacy preserving data mining algorithms. Data Mining and Knowledge Discovery **11**(2), 121–154 (2005)
7. Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata. In: L. Zayatz, P. Doyle, J. Theeuwes, J. Lane (eds.) Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 113–134. North-Holland (2002)
8. Duncan, G.T., Keller-McNulty, S.A., Stokes, S.L.: Disclosure risks vs. data utility: The R-U confidentiality map. Tech. Rep. 121, National Institute of Statistical Sciences (2001)
9. Dwork, C., Nissim, K.: Privacy preserving data mining in vertically partitioned database. In: CRYPTO 2004, vol. 3152, pp. 528–544 (2004)
10. Evfimievski, A.: Randomization in privacy preserving data mining. SIGKDD Explor. Newsl. **4**(2), 43–48 (2002)
11. Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Privacy preserving mining of association rules. In: 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217–228. ACM-Press (2002)
12. Fung, B.C.M., Wang, K., Yu, P.S.: Top-down specialization for information and privacy preservation. In: Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE 2005). Tokyo, Japan (2005)
13. Iyengar, V.: Transforming data to satisfy privacy constraints. In: Proc., the Eigth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 279–288 (2002)
14. Kantarcioglu, M., Clifton, C.: Privacy preserving distributed mining of association rules on horizontally partitioned data. In: ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 24–31 (2002)
15. Kantarcıoğlu, M., Jin, J., Clifton, C.: When do data mining results violate privacy? In: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 599–604. Seattle, WA (2004). URL http://doi.acm.org/10.1145/1014052.1014126
16. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy preserving properties of random data perturbation techniques. In: Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03). Melbourne, Florida (2003)
17. Kifer, D., Gehrke, J.: Injecting utility into anonymized datasets. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, pp. 217–228. ACM Press, Chicago, IL, USA (2006)
18. Kumar Tayi, G., Ballou, D.P.: Examining data quality. Communications of the ACM **41**(2), 54–57 (1998)
19. Liu, K., Kargupta, H., Ryan, J.: Random projection-based multiplicative data perturbation for privacy preserving distributed data mining **18**(1), 92–106 (2006)
20. Nergiz, M.E., Clifton, C.: Thoughts on k-anonymization. In: The Second International Workshop on Privacy Data Management held in conjunction with The 22nd International Conference on Data Engineering. Atlanta, Georgia (2006)

21. Oliveira, S.R.M., Zaiane, O.R.: Privacy preserving frequent itemset mining. In: IEEE icdm Workshop on Privacy, Security and Data Mining, vol. 14, pp. 43–54 (2002)
22. Oliveira, S.R.M., Zaiane, O.R.: Privacy preserving clustering by data transformation. In: 18th Brazilian Symposium on Databases (SBBD 2003), pp. 304–318 (2003)
23. Oliveira, S.R.M., Zaiane, O.R.: Toward standardization in privacy preserving data mining. In: ACM SIGKDD 3rd Workshop on Data Mining Standards, pp. 7–17 (2004)
24. Rizvi, S., Haritsa, R.: Maintaining data privacy in association rule mining. In: 28th International Conference on Very Large Databases, pp. 682–693 (2002)
25. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering (TKDE) **13**(6), 1010–1027 (2001). DOI HTTP://doi.ieeecomputersociety.org/10.1109/69.971193
26. Schoeman, F.D.: Philosophical Dimensions of Privacy: An Anthology. Cambridge University Press. (1984)
27. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**, 379–423, 623–656 (1948)
28. Sweeney, L.: Achieving *k*-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems **10**(5), 571–588 (2002)
29. Trottini, M.: A decision-theoretic approach to data disclosure problems. Research in Official Statistics **4**, 7–22 (2001)
30. Trottini, M.: Decision models for data disclosure limitation. Ph.D. thesis, Carnegie Mellon University (2003). Available at `http://www.niss.org/dgii/TR/ThesisTrottini -final.pdf`
31. Vaidya, J., Clifton, C.: Privacy preserving association rule mining in vertically partitioned data. In: 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639–644. ACM Press (2002)
32. Verykios, V.S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., Theodoridis, Y.: State-of-the-art in privacy preserving data mining. SIGMOD Record **33**(1), 50–57 (2004)
33. Walters, G.J.: Human Rights in an Information Age: A Philosophical Analysis, chap. 5. University of Toronto Press. (2001)
34. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems **12**(4), 5–34 (1996)
35. Willenborg, L., De Waal, T.: Elements of statistical disclosure control, *Lecture Notes in Statistics*, vol. 155. Springer (2001)