JRC Scientific and Technical Reports

Privacy Preserving Data Mining, Evaluation Methodologies

Igor Nai Fovino and Marcelo Masera

EUR 23069 EN - 2008

JRC
EUROPEAN COMMISSION

ipsc
Institute for the Protection and Security of the Citizen

The Institute for the Protection and Security of the Citizen provides research-based, systems-oriented support to EU policies so as to protect the citizen against economic and technological risk. The Institute maintains and develops its expertise and networks in information, communication, space and engineering technologies in support of its mission. The strong cross-fertilisation between its nuclear and non-nuclear activities strengthens the expertise it can bring to the benefit of customers in both domains.

European Commission
Joint Research Centre
Institute for the Protection and Security of the Citizen

Contact information
Address: Via E Fermi I-21020 Ispra (VA) ITALY
E-mail: igor.nai@jrc.it
Tel.: +39 0332786541
Fax: +39 0332789576

http://ipsc.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

JRC JRC42699

EUR 23069 EN
ISSN 1018-5593

Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

Privacy Preserving Data Mining, Evaluation Methodologies

Igor Nai Fovino, Marcelo Masera

16th January 2008

Contents

Introduction

Introduction

Several data mining techniques, incorporating privacy protection mechanisms, have been developed based on different approaches. For instance, various sanitization techniques have been proposed for hiding sensitive items or patterns that are based on removing reserved information or inserting noise into data. Privacy preserving classification methods, instead, prevent a miner from building a classifier able to predict sensitive data. Additionally, privacy preserving clustering techniques have been recently proposed, which distort sensitive numerical attributes, while preserving general features for clustering analysis.

Given the number of different privacy preserving data mining (PPDM) techniques that have been developed over the last years, there is an emerging need of moving toward standardization in this new research area, as discussed in [67]. We believe that one step toward this essential process is the definition of a framework identifying the various parameters which characterize a PPDM algorithm, thus making it possible to assess and compare such techniques according to a fixed set of evaluation criteria. Because all the various techniques differ among each other with respect to a number of criteria, like performance, data quality, privacy level, it is important to provide a systematic and comprehensive framework for their evaluation. In many cases, no technique is better than the other ones with respect to all criteria. Thus, one has to select the privacy preserving technique based on the criterion to be optimized.

An unified framework allowing to satisfy these goals is essential in order to select the privacy preserving technique which is more adequate based on the data and the requirements of the application domain. Moreover, a major feature of PPDM techniques is that they usually entail modifications to the data in order to sanitize them from sensitive information (both private data items and complex data correlations) or to anonymize them with some uncertainty level. Therefore, in evaluating a PPDM algorithm it is important to assess the quality of the transformed data. To do this, we need methodologies for the assessment of the quality of data, intended as the state of the individual items in the database resulting from the application of a privacy preserving technique, as well as the quality of the information that is extracted from the modified data by using a given data mining method. As we already explained, the former notion of data quality is strictly related to the use the data are intended for. Moreover, some of those algorithms can be computationally very expensive and

thus cannot be used when very large sets of data need to be frequently released. Therefore, in addition to data quality, performance also needs to be carefully assessed. Other aspects, like scalability, need also to be taken into account since the data collected, stored and managed for the mining process grow enormously. We thus clearly need a comprehensive evaluation framework characterized by several metrics relevant for assessing the various aspects of PPDM algorithms.

In this report, we present results of our researches on this topic [10,11], and more in detail, a framework which allows one to compare the various privacy preserving techniques on a common platform. The framework consists of a number of evaluation criteria and a set of tools for data pre-processing and PPDM algorithm evaluation. The framework has been extensively used for assessing a suite of PPDM algorithms developed as part of the CODMINE project [97]. Moreover, due to the fact that the problem of disclosing private information when partially or totally releasing data storage is also addressed in the area of statistical databases[1] we also analyze and compare some of these methods along with the metrics used for evaluating them.

[1]We remember here that the statistical disclosure control (SDC) aims at protecting individual data, referred to as *microdata* according to the SDC terminology, when releasing some relevant statistics by means of statistics-based techniques, some of which are also adopted in the area of data mining

Chapter 1

State of the art

We already explained in the introduction that no unified framework exists supporting the evaluation of PPDM algorithms. Moreover a set of universally accepted general parameters on the basis of which one can evaluate some particular aspects of PPDM algorithms not exist. In this chapter, we will briefly focus our attention on the parameters used by the different authors of PPDM algorithms in order to prove the properties of their algorithms. Moreover, on the basis of the presented parameters we will make some considerations on the goals a PPDM must (or should) satisfy. This short presentation will act as the basis over which we will build the remaining part of this report.

1.1 Evaluation, State of The Art

Oliveira and Zaiane [68] in the evaluation of their heuristic-based framework for preserving privacy in mining frequent itemsets introduce some measures quantifying the effectiveness and the efficiency of their algorithms. The first parameter is evaluated in terms of

- *Hiding Failure*, that is, the percentage of restrictive patterns that are discovered from the sanitized database.
- *Misses Cost*, that is, the percentage of non-restrictive patterns that are hidden after the sanitization process.
- *Artifactual Pattern*, measured in terms of the percentage of discovered patterns that are artifacts.

Moreover, the specification of a *disclosure threshold* ϕ, representing the percentage of sensitive transactions that are not sanitized, allows one to find a balance between the hiding failure and the number of misses. The efficiency of the algorithms is measured in terms of CPU time, by first keeping constant both the size of the database and the set of restrictive patterns, and then by increasing the size of the input data in order to assess the algorithm scalability. Moreover,

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.

In [92], instead, Sweeney proposes a heuristic-based approach for protecting raw data through generalization and suppression techniques. The method she proposes provides *K-Anonymity*. As already explained in Chapter 2, in some way, the cell distortion that normally affects a database sanitized by K-anonymity, can be identified as a measure of DQ impact of the sanitization on the target database. Sweeney measures the cell distortion as the ratio of the domain of the attribute to the height of the attribute generalization which is a hierarchy. In the same article the concept of *precision* is also introduced. Given a table T, the *precision* represents the information loss incurred by the conversion process from a table T to a K-Anonymous Table T^k. More in detail the *precision* of a table T^k is measured as follows:

Given a database DB with N_A attributes and N transactions, if we identify as generalization scheme a domain generalization hierarchy GT with a depth h, it is possible to measure the quality of a sanitized database SDB as:

$$Quality(SDB) = 1 - \frac{\sum_{i=1}^{i=N_A} \sum_{j=1}^{i=N} \frac{h}{|GT_{Ai}|}}{|DB| * |N_A|} \tag{1.1}$$

where $\frac{h}{|GT_{Ai}|}$ represent the detail loss for each cell sanitized.

Agrawal and Srikant in [4] introduce a quantitative measure to evaluate the amount of privacy offered by a method and evaluate the proposed method against this measure. More specifically, if one can estimate with c% confidence that a value x lies in an interval, then the width of such interval defines the amount of privacy with a c% confidence level. They also assess the accuracy of the proposed algorithms for Uniform and Gaussian perturbation and for fixed privacy level. In [2] Agrawal and Aggarwal propose some metrics in order to evaluate privacy and information loss. Unlike the approach in [4], the privacy metric proposed by Agrawal and Aggarwal takes into account the fact that both the perturbed individual records and the reconstructed distribution are available to the user as well as the perturbing distribution, as it is specified in [36]. This metric is based on the concept of mutual information between the original and perturbed records. The average conditional privacy of an attribute A, given some other information, modeled with a random variable B, is defined 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. The information loss, instead, measures the lack of precision in estimating the original distribution from the perturbed data. It is defined as half the expected value of the L_1-norm between the original distribution and the reconstructed one. The proposed metric for evaluating information loss is related to the amount of mismatch between the original distribution and its estimate in terms of area. Both the proposed metrics are universal in the sense

that they can be applied to any reconstruction algorithm, independently from the particular data mining task applied.

Evfimievski et al. [37], in order to evaluate the privacy breaches, 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. Finally, the itemset size giving the worst value for each of these two values is selected.

Rivzi and Haritsa [81] propose a privacy measure dealing with the probability with which the user's distorted entries can be reconstructed. 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.

Kantarcioglu and Clifton in [53] evaluate the methods they propose in terms of communication and computation costs:

- Communication Cost: is expressed in terms of the number of messages exchanged among the sites, that are required by the protocol for securely counting the frequency of each rule.
- Computation cost: is expressed in terms of the number of encryption and decryption operations required by the specific algorithm.

Oliveira and Zaiane in their work on Clustering PPDM [69] define a performance measure that quantifies the fraction of data points which are preserved in the corresponding clusters mined from the distorted database. More in detail, a specific parameter, called *misclassification error*, is also introduced for measuring the amount of legitimate data points that are not well-classified in the distorted database. Finally, the privacy ensured by such techniques is measured as the variance difference between the actual and the perturbed values.

In the context of statistical disclosure control a large number of methods, called *masking* methods in the SDC jargon, have been developed to preserve individual privacy when releasing aggregated statistics on data, and more specifically to anonymize the released statistics from those data items that can identify one among the individual entities (person, household, business, etc.) whose features are described by the statistics, also taking into account, additionally, related information publicly available [106]. In [27] a description of the most relevant masking methods proposed so far is presented. Among the perturbative methods specifically designed for continuous data, the following masking techniques are described: additive noise, data distortion by probability distribution, resampling, microaggregation, rank swapping, and so on. 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. All these masking methods are assessed according

to the two main parameters: the *information loss* and the *disclosure risk*, that is, the risk that a piece of information be linked to a specific individual. Several metrics are presented in the paper for assessing the *information loss* and the *disclosure risk* given by a SDC method. Additionally, in order to provide a trade-off level between these two metrics, a score is defined that gives the same importance to disclosure risk and information loss.

1.2 Considerations

In order to define which set of parameters is the most suitable to evaluate PPDM algorithms, it is previously necessary to define which are the main goals a PPDM algorithm should satisfy and then, starting from these considerations, reflect on the dimensions to be taken into account in the evaluation phase. On the basis of the content of the previous section, it is evident that a PPDM algorithm must satisfy the following requirements:

1. It should prevent the discovery of sensible information.
2. The sanitized database should be resistant to the various data mining techniques.
3. It should not compromise the access and use of non sensitive data.
4. It should be usable on large amounts of data.
5. It should not have an exponential computational complexity.
6. It should not consume an high amount of resources

Current PPDM algorithms do not satisfy all these goals at the same time; for instance, only few of them satisfy the point (2). The above list of goals helps us to understand how to evaluate these algorithms in a general way. The framework we have identified is based on the following evaluation dimensions:

- **Efficiency**, that is, the ability of a privacy preserving algorithm to execute with good performance in terms of all the resources implied by the algorithm. It can be used to measure the goal number six.
- **Scalability**, which evaluates the efficiency trend of a PPDM algorithm for increasing sizes of the data from which relevant information is mined while ensuring privacy. It can be used to measure the goal number four.
- **Data Quality** which evaluate the impact of the sanitization on the database DQ. It is related with the goal number three.
- **Hiding Failure**, that is, the portion of sensitive information that is not hidden by the application of a privacy preservation technique. It is related to the goal number one.

- **Privacy level** offered by a privacy preserving technique, which estimates the degree of uncertainty, according to which sensitive information, that has been hidden, can still be predicted. It can be used to give an alternative measure to the goal one and partially to the goal two.

An important question is which one among the presented "dimensions" is the most relevant for a given privacy preserving technique. Dwork and Nissim [32] 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, flexibility 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 quality of data 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. A trade-off metric can help us to state a unique value measuring the effectiveness of a PPDM algorithm. In Domingo-Ferrer and Torra [27] 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 Duncan, Keller-McNulty and Stokes [31] 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 the 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. In the current work, however, we mainly focus on the different evaluation criteria characterizing a PPDM algorithm.

In the following, we discuss in deep each evaluation criteria we have identified. More in detail, we divided these criteria into three groups related to different aspect to be assessed: the *Operational Parameters* are mainly related to computational measures, the *Privacy Parameters* strongly related to the not evident definition of Privacy and the *Data Quality Parameters* related to the already presented concepts of DQ.

Chapter 2

Operational parameters

The *Operational Parameters* are mainly related to the computational properties of the algorithms. In this class of parameters we consider the efficiency, the scalability, the hiding failure and the complexity. In what follows we give a detailed definition of these parameters.

2.1 Efficiency

The assessment of the resources used by a privacy preserving data mining algorithm is given by its *Efficiency*, which represents the ability of the algorithm to execute with good performance in terms of all used resources. Performance is assessed, as usually, in terms of time and space, and, in case of distributed algorithms, in terms of the communication costs incurred during information exchange.

Time requirements can be evaluated in terms of CPU time, or computational cost, or even the average of the number of operations required by the PPDM technique. Clearly, it would be desirable that the algorithms have a polynomial complexity rather than an exponential one. Anyway, it can be useful to compare the performance of the privacy preserving method with the performance of the data mining algorithm for which the privacy preserving method has been developed. Our initial expectation is that the execution times of the hiding strategies be proportional to the execution times of the mining algorithms that extract the sensitive information. Such an expectation was contradicted by tests performed in some cases. That can be easily explained. In fact in the cases in which the PPDM algorithm acts simply modifying data without taking into account some complex criteria like the data quality, the efficiency of the PPDM algorithm and the DM algorithm for which it was designed to fight against are similar. Obviously that will not be the same in the case in which, in order to make the better modifications, the PPDM algorithm adopts more complex strategies.

Space requirements are assessed according to the amount of memory that must be allocated in order to implement the given algorithm.

Finally, communication requirements are evaluated for those data mining algorithms, which require information exchanges during the secure mining process, as the cryptography-based techniques. It is measured in terms of the number of communications among all the sites involved in the distributed data mining task.

2.2 Scalability

As in the case of DQ different definitions of scalability have been proposed. For this reason it is not so simple to define "What we want to measure" in our particular case. However, one of the fields in which this concept is well explored is the field of multiprocessor systems. Starting from experiences in this context we will give our definition of scalability for PPDM algorithms. In [33] Eager claims that intuitively scalability implies a favorable comparison between a larger version of some parallel system with either a sequential version of that same system or a theoretical parallel machine. He relates scalability to the concept of *speedup*. More in detail, let $time(n, x)$ be the time required by an $n - processor$ system to execute a program to solve a of problem of size x, the speedup on a problem of size x with n processors is the execution time on one processor divided by the time on n processors:

$$speedup(n, x) = \frac{time(1, x)}{time(n, x)} \tag{2.1}$$

Moreover, Eager relates Efficiency with speedup as follows:

$$efficiency(n, x) = \frac{speedup(n, x)}{n} = \frac{time(1, x)}{n * time(n, x)} \tag{2.2}$$

Intuitively, a system with a linear speedup ($speedup(n, x) = n$) can be assumed to be scalable. We can thus propose a first definition:

Definition 1 *A system is scalable if efficiency (n, x) = 1 for all algorithms, number of processors n and problem sizes x.*

This is however a not useful definition. As Amdhal [6] notes, *many parallel algorithms have a sequential (or at least not completely parallel) component, yielding poor efficiency for a sufficiently large number of processors.* Moreover there exists the size problem (it is constant or not?). Some approaches to scalability are based on the concept of *theoretical parallel machines* [40] and on the comparison between the efficiency of the real machine with the theoretical one.

Starting from these considerations we give a definition tailored for our particular case. In order to do this, some consideration must be made. We are not interested to evaluate the case in which an algorithm is executed on a single processor or on a multiprocessor; we want a measure completely independent from "hardware constraint". Moreover in the database context the prominent

position is occupied by data and its size. We want to link the scalability with the "database size". For this reason, a PPDM algorithm has to be designed and implemented for being scalable with larger and larger datasets. The less rapid is the decrease in the efficiency of a PPDM algorithm for increasing data dimensions, the better is its scalability. By recalling equations 2.1 and 2.2 we can define the scalability as follows:

Definition 2 *Let A an algorithm for PPDM, let D a database to be sanitized, we define the scalability of the algorithm A as the efficiency trend for increasing values in data sizes of D*

Therefore, such parameter concerns the increase of both performance and storage requirements together with the costs of the communications required by a distributed technique when data sizes increase.

Formally, we define the scalability as the speedup of a mono-processor computation in function of the size increase of the database.

$$Scalability = (speedup(1, size(t))) \tag{2.3}$$

Obviously, given this definition, it is easy to extend it in the case in which we want to evaluate the PPDM algorithm in a multi-processor context. However we are not actually interested to this type of application.

2.3 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 privacy preserving data mining 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. It is important to underline that, as we will explain well after, the hiding failure is not related to the Privacy level Measure. In fact, the hiding failure measures the percentage of failure in the hiding process, or, roughly speaking, the number of sensitive information (clusters, rules etc.) not hidden. The privacy parameter, instead, starting from the assumption that all the sensitive rules are hidden, measures how strongly the information is hidden. This concept will be extensively explained in the following chapters.

2.4 Complexity

If an algorithm halts, we define its running time to be the sum of the costs of each instruction carried out. Within the RAM model of computation, arithmetic operations involve a single instruction and could be assumed to have a unit cost.

By computing then the computational cost of an algorithm we have in some way an alternative measure of the efficiency and scalability of an algorithm. It represents the theoretical measure of the algorithm behavior. It is however important, when considering the computational complexity, to take into account the dimension and the type of the input. In fact, when comparing different algorithms, if these properties are very different, it does not probably make sense to consider as discriminant the complexity of the algorithms.

Chapter 3

Privacy Level

In our society the term "*Privacy*" 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 [102] and Walters [103] identify three possible definitions of privacy:

1. Privacy as the right of a person to determine which personal information about himself/herself may be communicated to others.
2. Privacy as the control over access to information about oneself.
3. Privacy as limited access to a person and to all the features related to the person.

These three definitions are very similar apart from some philosophical differences that are not in the scope of our work. What is interesting from our point of view is the concept of "Controlled Information Release" emerging from the previous definitions. 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:
Definition 3 "*The right of an entity to be secure from unauthorized disclosure*

[1] Health Insurance Portability and Accountability Act

of sensitive 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 [67] 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 in which 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 sensitive raw data need to protected. Such a scenario justifies the final generalization of our privacy definition.

As we already noted before in this work, PPDM algorithms act in very different ways in order to hide the sensitive information. The question is then the following: "The different sanitizations are all equally robust?" or, more clearly "Which is the effort a malicious agent must spend in order to discover the hidden information?". Obviously not all the PPDM algorithms are able to guarantee the same robustness. For this reason we believe that when evaluating a set of PPDM algorithm it is necessary to take into account such considerations. The *Privacy Level* parameter allows us to measure how strong is the sanitization performed.

3.1 The Privacy Evaluation in PPDM

As we have explained before, we are interested in assessing the privacy introduced in a database by a sanitization operation. This is a not simple task. Privacy is an abstract concept and then it is not possible to measure it directly. However, as with every type of abstract concept, it is possible to measure privacy in an indirect way, by finding and measuring some phenomena that can be in some way identified as a direct effect of a privacy variation. The question we want to solve in this section is then the following "Are there some observable phenomena linked with the privacy variation?". An answer to this question is not straightforward. Moreover, the phenomena we search should have some desirable characteristics:

- They should be observable in any type of database (we want to provide an general measure).
- They should be applicable to any type of PPDM sanitization.

- They should be not directly linked to a sensitive information (otherwise it could be possible to use the same measure to make a privacy breach).

By analyzing the privacy definition, it is evident that it is strongly related to the information contained in the sanitized database. Moreover a privacy breach happens when a malicious user is able to link some information to a specific object.

The sanitization acts as a "Confusion Agent" that avoids the Malicious User be able to see clearly the reality. In physics the confusion of a system is strongly related to the Entropy of a system. The English Oxford Dictionary gives as the first definition of entropy the following one: "For a closed system, the quantitative measure of the amount of thermal energy not available to do work". If we substitute "Thermal Energy" with "Information", we obtain that the entropy in some way is the amount of information not available to do work (to be used). Moreover, in a well known book of R. Feynman [39] the author claims:

So we now have to talk about what we mean by disorder and what we mean by order. ... Suppose we divide the space into little volume elements. If we have black and white molecules, how many ways could we distribute them among the volume elements so that white is on one side and black is on the other? On the other hand, how many ways could we distribute them with no restriction on which goes where? Clearly, there are many more ways to arrange them in the latter case. We measure "disorder" by the number of ways that the insides can be arranged, so that ***from the outside it looks the same****. The logarithm of that number of ways is the entropy. The number of ways in the separated case is less, so the entropy is less, or the "disorder" is less.*

The phrase in bold style is the key of our idea. In fact, the aim of the sanitization is to hide information in such a way the external users are unable to discover the modification. Moreover if we assume the original db as an ordered universe, the sanitization introduces some disorder. The biggest is the disorder, the biggest is the number of possibilities in which it is possible to rearrange the universe, and then more difficult is to recover the sensitive information.

To summarize, the intuition is that in some way the entropy of the db is related to the privacy introduced by the sanitization. That is, however, not sufficient to define any type of privacy measure. We need to make another little step. In the 1948 C. Shannon wrote his most famous paper, "A Mathematical Theory of Communication". In such paper he give an interpretation of entropy applicable to the information theory. What it is interesting to underline is that, when talking about information and communication channel, he defined the concept of *Information Content* claiming that the information contained in a data sent along a communication channel is inversely related to the probability of occurrence. In other words, the information associated with an event having a low probability of occurrence is bigger than the one associated with an event having an high probability of occurrence. He links the concept of Information Content to the definition of Information Entropy. Thus, our hypothesis is that it is possible to measure the Privacy introduced by a PPDM algorithm measuring the variation of Information Content associated to the database.

3.2 The Privacy Level Measure

In order to evaluate the privacy protection offered by a PPDM algorithm, we need to define a unique parameter quantifying the privacy level ensured by these algorithms. As previously stated, a metric for evaluating the privacy level offered by a PPDM method is proposed in [4]: 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. This metric does not work well because it does not take into account the distribution of the original data along with the perturbed data. We need, therefore, a metric that considers all the informative contents of data available to the user. Agrawal and Aggarwal [2] address this problem by introducing a new privacy metric based on the concept of information entropy.

Shannon in formulating his most well-known theorem [87] defines the concept of *Information Entropy* as follows: 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:

$$h(X) = -\sum p(x) \log_2(p(x)) \tag{3.1}$$

or, in the continuous case:

$$h(X) = -\int f(x) \log_2(f(x))dx \tag{3.2}$$

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 evaluating the privacy achieved by a PPDM algorithm as we mentioned in the previous section. 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 considered as the evaluation of the uncertain forecast level of an event which in our context is evaluation of the right value of a datum.

As in [2], we measure the level of privacy inherent in an attribute X, given some information modeled by Y, by the following quantity:

$$\Pi(X|Y) = 2^{-\int f_{X,Y}(x,y) \log_2 f_{X|Y=y}(x))dxdy} \tag{3.3}$$

in which the exponent is the conditional entropy of a random variable X (modeling the original data)given a random variable Y (modeling the sanitized (perturbed) data)

However, we have to notice 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 [95, 96]. In order to solve this problem, with respect to the expression 3.3, it is possible to introduce assumptions on the attacker knowledge by properly modeling Y. Due to the fact that we are interested actually in measuring the privacy level of the PPDM algorithm without making assumption on the environment in which the algorithm will work, we actually do not consider this extension.

The measure of the entropy level, and thus of the privacy level, is very general and in order to use it in the different PPDM contexts, it must be refined with respect to some characteristics like the type of transactions, the type of aggregation and PPDM methods. Here, for example, we show how the entropy concept can be instantiated in order to evaluate the privacy level in the context of "association rules". Our approach is based on the work of Smyth and Goodman [91] that use the concept of Information Entropy in order to measure the amount of information contained in the association rules extracted from a database, with the aim of ranking and thus characterizing the most important rules in terms of information they contain. They think of a rule $y \Rightarrow x$ as a condition *if Y=y then X=x* with a certain probability p. Intuitively the two random variables can be viewed as being connected by a discrete memoryless channel. The channel transition probabilities are the conditional probabilities between the two variables. A rule corresponds to a particular input event $\mathbf{Y}$=y, rather than the average over all input events as is defined for communication channels, and p, the rule probability, is the transition probability $p(X = x|Y = y)$. Starting from these assumptions, it is possible then define the *Instantaneous Information* (i.e. the information we have about X knowing that $\mathbf{Y}$=y occurs) as a function $f(\mathbf{X{:}Y} = y)$. As defined by Shannon in [87] a requirement for f is that

$$E_y[f(\mathbf{X{:}Y} = y)] = I(\mathbf{X{:}Y}) \tag{3.4}$$

where E_y is the expectation with respect to the random variable $\mathbf{Y}$. In [12], Blachman showed that f($\mathbf{X{:}Y}$=y) is not unique and it has two possible solutions, the *i-measure* and the *j-measure*.

$$i(\mathbf{X{:}Y} = y) = \sum_x p(x) log \left(\frac{1}{p(x)}\right) - \sum_x p(x|y) log \left(\frac{1}{p(x|y)}\right) \tag{3.5}$$

$$j(\mathbf{X{:}Y} = y) = \sum_x p(x|y) log \left(\frac{p(x|y)}{p(x)}\right) \tag{3.6}$$

Between the two measures, the only one, as proved by Blachman, that is not negative is the j measure. Adapting this measure to the case of a rule (a rule give information about the event X=x and its complement $\bar{x}$), it is possible to obtain the following function:

$$j_r(x, Y = y) = p(x|y) \log \frac{p(x|y)}{p(x)} + (1 - p(x|y)) \log \frac{1 - p(x|y)}{1 - p(x)}. \tag{3.7}$$

representing the cross-entropy of a rule.

Finally it is possible to define a J-measure representing the entropy of a rule as:

$$J(x, Y = y) = p(y)j_r(x, Y = y) \tag{3.8}$$

where the term *p(y)* is the probability of the rule antecedent.

If we consider the association rules model and a specific rule $y \Rightarrow x$, the value $p(y)$, that is, the probability of antecedent, is equal to frequency of y and the value $p(x|y)$ is the probability that the variable X assumes the value x, given that y is the value of variable Y. It represents the strength of the rule *if* Y=y *then* X=x and it is referred to as *confidence* of the rule. Now, we define a parameter entropy privacy (EP) as:

$$EP = J(x, Y = y) - J_1(x, Y = y) \tag{3.9}$$

where J_1 is the J-measure after the operation of data hiding. Some preliminary tests executed in the context of this work, show that the simple J-measure does not provide an intuitive evaluation parameter. In fact, as the Information Theory suggests, we would expect as result that when the confidence decreases, the level of entropy increases (see Figure 3.1). Indeed, in some particular cases, the trend obtained is the one shown in Figure 3.2. This is due to the fact that the J-measure represents the average conditional mutual information, or, in other words, the difference between the "a priori" and "a posteriori" probabilistic distributions of the two random variables X and Y. On the base of this observation, we note that if:

- $P(X \wedge Y) < P(X) \times P(Y)$ the two variables X and Y are negatively correlated
- $P(X \wedge Y) > P(X) \times P(Y)$ the two variables X and Y are positively correlated
- $P(X \wedge Y) = P(X) \times P(Y)$ the two variables X and Y are independent

By remembering that:

$$\frac{P(X \wedge Y)}{P(Y)} = P(X|Y) \tag{3.10}$$

we observe than the J-measure does not take into account the type of correlation between the involved random variables. In fact that happen only in the case in which during the sanitization process, the confidence of the rule remains under the value of the support. In this case, when the confidence value decrease, the J-measure value increase. Studying the J-measure function, it is possible to see that it always has a minimum. The derived function is negative when $p(x|y) < p(x)$ and positive when $p(x|y) > p(x)$. For this reason, we finally adopt as measure the derivative of the J-measure (for make easy to understand the steps, s is equal to $p(x)$, b is equal to $p(y)$ and x is equal to $(p(x|y))$):

$$J'(X; Y = y) = \left(b * \left(x * log_2\left(\frac{x}{a}\right) + (1 - x) * log_2\left(\frac{1 - x}{1 - a}\right)\right)\right) = \tag{3.11}$$

Rules	Sup. Ant.	Conf	Sup. Post.	j	J
$12 \rightarrow 21$	0.302466208	0.310074481	0.313018734	**0.00002**	**0.000006**
After sanitization	0.290372303	0.203756635	0.29049087	**0.019570**	**0.005683**
$43 \rightarrow 12$	0.22290728	0.325	0.302466208	**0.0018**	**0.000265**
After sanitization	0.198482333	0.118876941	0.278041262	**0.074538**	**0.014794**
$45 \rightarrow 8$	0.296063552	0.315178214	0.298672042	**0.000644**	**0.000191**
After sanitization	0.261204648	0.090331366	0.263813137	**0.095670**	**0.024989**

Table 3.1: Example of rule entropy before and after the sanitization phase

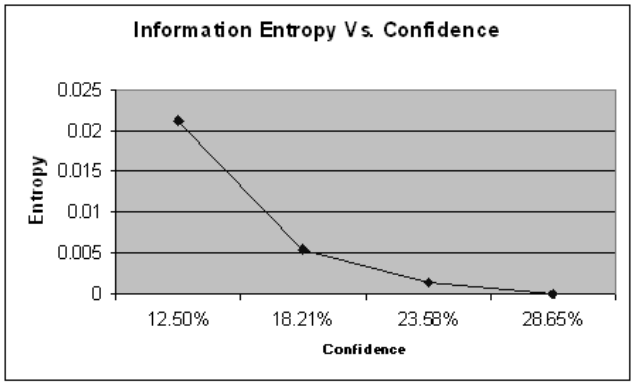

Figure 3.1: Evolution of information entropy with respect to confidence

$$= b*\left(1 * log_2\left(\frac{x}{a}\right) + x * \frac{a}{x} * \frac{1}{a} - 1 * log_2\left(\frac{1-x}{1-a}\right) + (1-x) * \frac{1-a}{1-x} * \frac{-1}{1-a}\right) = \tag{3.12}$$

$$= b * \left(log_2\left(\frac{x}{a}\right) + 1 - log_2\left(\frac{1-x}{1-a}\right) - 1\right) = b * \left(log_2\left(\frac{x}{a}\right) - log_2\left(\frac{1-x}{1-a}\right)\right) \tag{3.13}$$

Making the propers substitutions we obtain:

$$J'(X; Y = y) = p(y) * (log_2(\frac{p(x|y)}{p(x)}) - log_2(\frac{1-p(x|y)}{1-p(x)})) \tag{3.14}$$

Fig 3.3 shows the graph obtained when using J'.
Finally, we measure the amount of privacy introduced by the following expression:

$$Level\_of\_privacy = (J'_1 - J') \tag{3.15}$$

where J'_1 is the calculated after the sanitization and J' is measured before sanitization.
We observe that a decrease in the confidence value, and therefore an increase

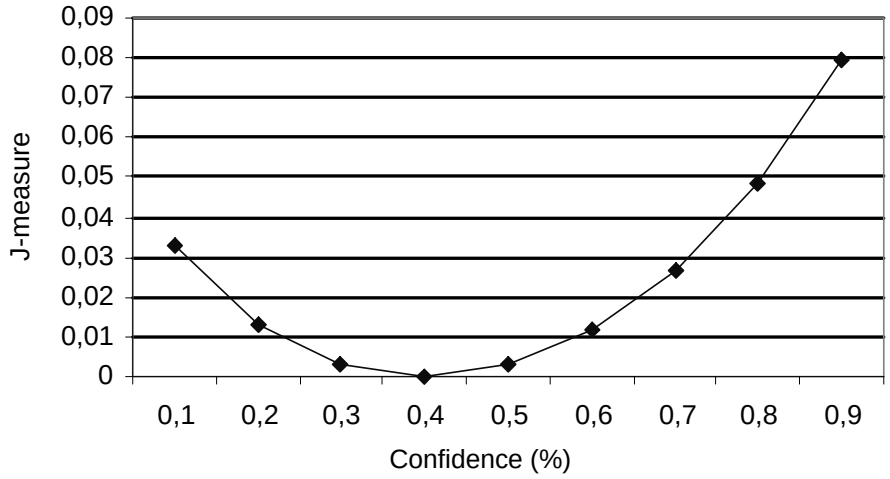

Figure 3.2: Evolution of information entropy with respect to confidence in some particular cases

in the level of privacy, results in an increase of entropy, as shown in Fig. 3.3. This trend is in accordance with the Information Theory, because if a rule or in general an information is well hidden, the informative value of its discovery is higher than an information that is simple to discover.

The measure we provided show how to use the concept of Information Content to measure the Privacy level introduced by a sanitization in the context of Association Rule Mining. Making a little step, it is possible to extend this measure to the case of Classification PPDM algorithms. In fact, in the classification context, we are interested to find a classification model able to describe some elements contained in a database. To each classification model, if we assume to be able to build the classification pattern, it is possible to associate a set of rules like the following:

if A1 A2...An then Ci

where each rule represent a class of the model. Applying then the expression 3.15 to these rules, we are able to compute the privacy lever introduced for every class.

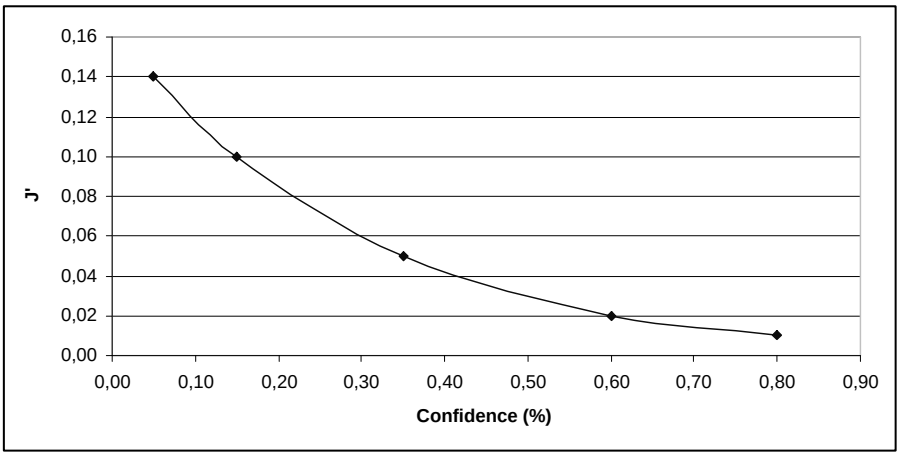

Figure 3.3: Evolution of J' with respect to Confidence

Chapter 4

Data Quality

The quality of the data is one of the most important properties of a database. Higher is the DQ, the better is the real world representation given by the models contained in the database and the higher is the usefulness of the information of the data contained in the database.

Therefore a set of operations that in some way contribute to downgrade the DQ of a database may results in huge economical and, in some specific cases, social damages.

The sanitization process performed by a PPDM algorithm constitutes potentially one of these specifications. For this reason it is important, when assessing PPDM algorithms, to take into account this relevant aspect. In what follows we present a set of parameters, called *Raw Parameters*, that can be used to make a non accurate, on the fly, DQ evaluation. Moreover, we present a most sophisticated techniques based on the concept of Information Quality Model. Finally we will describe a three-step framework that allows one to select from a set of PPDM algorithm the most suitable for a target database with respect to both Operational and DQ parameters.

In some cases it can be useful to be able to give a non detailed but fast evaluation of the quality of the data mining results after the sanitization process. For example in order to identify from a set of PPDM algorithms, the subset with a general high impact on the DQ.

In order to perform a fast assessment of damages to DQ, it can be useful to analyze what happens at macroscopic level when a generic PPDM algorithm is applied to a database. Some experimental measurements (contained in the appendix), show that the most evident effect of a PPDM algorithm with an high impact on DQ is generally identified as an *Information Loss* of a database. It is possible then to identify the Information Loss as first crude measure of the DQ impact of PPDM algorithm.

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. Data modification often applied by a privacy preserving technique obviously affects the parameters involved in the clustering analysis. There is, thus, the need to control, as much as possible, the results of such analysis 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
- **Artifactual information** representing the percentage of artifactual patterns created by the adopted privacy preserving technique

In case of association rules, the lost information can be modeled as the set of nonsensitive 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.

4.1 Data Quality in the Context of PPDM

Traditionally DQ is a measure of the consistency between the data views presented by an information system and the same data in the real-world [70]. This definition is strongly related with the classical definition of information system as a "model of a finite subset of the real world" [56]. More in detail Levitin and Redman [59] claim that DQ is the instrument by which it is possible to evaluate if data models are well defined and data values accurate. The main problem with DQ is that its evaluation is relative [94], in that it usually depends from the context in which data are used. In the scientific literature DQ is considered a multi-dimensional concept that in some environments involves both objective and subjective parameters [104, 105]. In the context of PPDM, we are interested in assessing whether, given a target database, the sanitization phase will compromise the quality of the mining results that can be obtained from the sanitized database. The parameters we consider relevant in the context of PPPDM are the following: the *Accuracy*, measuring the proximity of a sanitized value a^I to the original value a; the *Completeness*, evaluating the percentage of data from the original database that are missing from the sanitized

database and finally the *Consistency* that is related to the semantic constraints holding on the data and it measures how many of these constraints are still satisfied after the sanitization. We now present the formal definitions of those parameters for use in the remainder of the discussion. Let OD be the original database and SD be the sanitized database resulting from the application of the PPDM algorithm. Without loosing generality and in order to make simpler the following definitions, we assume that OD (and consequently SD) be composed by a single relation. We also adopt the positional notation to denote attributes in relations. Thus, let od_i (sd_i) be the i-th tuple in OD (SD), then od_{ik} (sd_{ik}) denotes the k^{th} attribute of od_i (sd_i). Moreover, let n be the total number of the attributes of interest, we assume that attributes of positions $1, \ldots, m$ ($m \leq n$) are the primary key attributes of the relation.

Definition 1 ***1:*** Let sd_j be a tuple of SD. We say that sd_j is **Accurate** if $\neg\exists od_i \in OD$ such that $((od_{ik} = sd_{jk})\forall k = 1..m \wedge \exists(od_{if} \neq sd_{jf}), (sd_{jf} \neq NULL), f = m+1, .., n)$.
Definition 2 ***2:***A sd_j is **Complete** if $(\exists od_i \in OD$ such that $(od_{ik} = sd_{jk})\forall k = 1..m) \wedge (\neg\exists(sd_{jf} = NULL), f = m+1, .., n)$.

Let C the set of the constraints defined on database OD, in what follows we denote with c_{ij} the j^{th} constraint on attribute i. We assume here constraints on a single attribute, but, it is easily possible to extend the measure to complex constraints.

Definition 3 ***3:*** An instance sd_k is **Consistent** if $\neg\exists c_{ij} \in C$ such that $c_{ij}(sd_{ki}) = false, i = 1..n$

Chapter 5

Information Driven Data Quality Schema

Current approaches to PPDM algorithms do not take into account two important aspects:

- **Relevance of data**: not all the information stored in the database has the same level of relevance and not all the information can be dealt at the same way.
- **Structure of the database**: information stored in a database is strongly influenced by the relationships between the different data items. These relationships are not always explicit.

We believe that in a context in which a database administrator needs to choose which is the most suitable PPDM algorithm for a target real database, it is necessary to also take into account the above aspects. To achieve this goal we propose to use Data Quality in order to assess how and if these aspects are preserved after a data hiding sanitization.

5.1 The Information Quality Model

In order evaluate DQ it is necessary to provide a formal description that allow us to magnify the aggregate information of interest for a target database and the relevance of DQ properties for each aggregate information (AI) and for each attribute involved in the AI. The Information Quality Model (IQM) proposed here addresses this requirement. In the following, we give a formal definition for an Attribute Class (AC), a Data Model Graph (DMG) (used to represent the attributes involved in an aggregate information and their constraints) and an Aggregation Information Schema (AIS).Before giving the definition of DMG, AIS and ASSET we introduce some preliminary concepts.
Definition 3 ***4:*** An Attribute Class is defined as the tuple
$AT_C =< name, AW, AV, CW, CV, CSV, Slink >$ where:

- *Name* is the attribute id
- AW is the accuracy weigh for the target attribute
- AV is the accuracy value
- CW is the completeness weigh for the target attribute
- CV is the completeness value
- CSV is the consistency value
- *Slink* is list of simple constraints.

Definition 4 ***5:*** A Simple Constraint Class is defined as the tuple $SC_C =< name, Constr, CW, Clink, CSV >$ where:

- *Name* is the constraint id
- *Constraint* describes the constraint using some logic expression
- CW is the weigh of the constraint. It represents the relevance of this constraint in the AIS
- CSV is the number of violations to the constraint
- *Clink* it is the list of complex constraints defined on SC_C.

Definition 5 ***6:*** A Complex Constraint Class is defined as the tuple $CC_C =< name, Operator, CW, CSV, SC_C\_link >$ where:

- *Name* is the Complex Constraint id
- *Operator* is the "Merging" operator by which the simple constraints are used to build the complex one.
- CW is the weigh of the complex constraint
- CSV is the number of violations
- SC_Clink is the list of all the SC_C that are related to the CC_C.

Let D a database, we are able now to define the DMG, AIS and ASSET on D.

Definition 5 ***7 :*** A **DMG** (Data Model Graph) is an oriented graph with the following features:

- A set of nodes N_A where each node is an Attribute Class
- A set of nodes SC_C where each node describes a Simple Constraint Class
- A set of nodes CC_C where each node describes a Complex Constraint Class
- A set of direct edges $L_{Nj,Nk}$: $L_{Nj,Nk} \in ((N_A X SC_C) \cup (SC_C X CC_C) \cup (SC_C X N_A) \cup (CC_C X N_A))$.

Definition 6 ***8:*** An **AIS** ϕ is defined as a tuple $< \gamma, \xi, \lambda, \vartheta, \varpi, W_A IS >$ where: γ is a name, ξ is a DMG, λ is the accuracy of AIS, ϑ is the completeness of AIS, ϖ is the consistency of AIS and W_{AIS} represent the relevance of AIS in the database.

We are now able to identify as **ASSET** (Aggregate information Schema Set) as the collection of all the relevant AIS of the database.

The DMG completely describes the relations between the different data items of a given AIS and the relevance of each of these data respect to the data quality parameter. It is the "road map" that is used to evaluate the quality of a sanitized AIS

5.2 Data Quality Evaluation of AIS

By adopting the IQM scheme, now we are able to evaluate the data quality at the attribute level. By recalling *Definition (1,2)*, we define the *Accuracy lack* of an attribute k for an AIS A as the proportion of non accurate items in a database SD. Ate the same way, the *Completeness lack* of an attribute k is defined as the proportion of non complete items in SD. The accuracy lack index for an AIS can the be evaluated as follows:

$$ACL = \sum_{i=0}^{i=n} DMG.N_i.AV * DMG.N_i.AW \tag{5.1}$$

where $DMG.N_i.AW$ is the accuracy weight associated with the attribute identified by the node N_i. Similarly the completeness lack of an AIS can be measured as follows:

$$CML = \sum_{i=0}^{i=n} DMG.N_i.CV * DMG.N_i.CW \tag{5.2}$$

Finally the consistency lack index associated with an AIS is given by number of constraint violations occurred in all the sanitized transaction multiplied by the weight associated with every constraints (simple or complex).

$$CSL = \sum_{i=0}^{i=n} DMG.SC_i.csv * DMG.SC_i.cw + \sum_{j=0}^{j=m} DMG.CC_j.csv * DMG.CC_j.cw \tag{5.3}$$

5.3 The Evaluation Algorithm

In this section we present the methodology we have developed to evaluate the data quality of the AIS. This methodology is organized in two main phases:

- **Search**: in this phase all the tuples modified in the sanitized database are identified. The primary keys of all these transaction (we assume that the sanitization process does not change the primary key), are stored in a set named *evalset*. This set is the input of the Evaluation phase.

- **Evaluation**: in this phase the accuracy, the consistency and the completeness associated with the DMG and the AIS are evaluated using information on the accuracy and completeness weight associated with the DMG and related to the transactions in Evalset.

The algorithms for these phases are reported in Figures 1 and 2. Once the evaluation process is completed, a set of values is associated with each AIS that gives the balanced level of accuracy, completeness and consistency. However, this set may not be enough. A simple average of the different AIS's values could not be significant, because even in this case not all the AIS's in the ASSET have the same relevance. For this reason, a weight is associated with each AIS that represents the importance of the high level information represented by the AIS

```
INPUT:    Original database OD, Sanitized database SD
OUTPUT: a set Evalset of primary keys
Begin
   Foreach t_i ∈ OD do
      {j=0;
         While (s_jk ≠ t_ik)and(j < |SD|)do j++;
         l=0;
         While (s_jl = til)and(l < n) do l++;
         If(l < n)Then Evalset = Evalset ∪ t_ik
      }
End
```

Figure 5.1: Search algorithm

in the target context.The accuracy, the completeness and the consistency of the ASSET for each PPDM algorithm candidate are then evaluated as follows:

$$Accuracy_{\text{Asset}} = \frac{\sum_{i=0}^{i=|Asset|} AIS_i.accuracy * AIS_i.W}{|Asset|} \tag{5.4}$$

$$Completeness_{\text{Asset}} = \frac{\sum_{i=0}^{i=|Asset|} AIS_i.completeness * AIS_i.W}{|Asset|} \tag{5.5}$$

$$Consistency_{\text{Asset}} = \frac{\sum_{i=0}^{i=|Asset|} AIS_i.consistency * AIS_i.W}{|Asset|} \tag{5.6}$$

where $AIS_i.W$ represents the weight (relevance) associated with the i-th AIS.

INPUT: the original database OD, the sanitized database SD,Evalset, IQM.
OUTPUT: the IQM containing a data quality evaluation
Begin
Foreach IES **in** IQM **do**
{$DMG = IQM.IES.link$;
$avet = 0; cvet = 0$;
Foreach ($t_{ik} \in Evalset$)**do**
For($j = 0; j < n; j++$) **do**
If ($t_{ij} \neq s_{ij}$) **Then**
{
If $s_{ij} = NULL$ **Then** $cvet[j]++$;
Else $avet[j]++$;
validate_constr(IES,DMG,j)
}
For($m = 0; m < n; m++$)**do**
{ $DMG.N_m.AV = \frac{avet[m]}{|SD|}$; $DMG.N_m.CV = \frac{cvet[m]}{|SD|}$;}
$IQM.IES.AV = \sum_{i=0}^{i=n}(DMG.N_i.AV * DMG.N_i.AW)$;
$IQM.IES.CV = \sum_{i=0}^{i=n}(DMG.N_i.CV * DMG.N_i.CW)$;
$IQM.IES.CSV = \sum_{i=0}^{i=n} DMG.SC_i.CSV * DMG.SC_i.CW + \sum_{j=0}^{j=m} DMG.CC_j.CSV * DMG.CC_j.CW$;}
End
Procedure validate_constr(IES,DMg,j) **Begin**
N_A=AIS.DMG.j
For k=1; $k < |N_A.slink|$**; k++**
{ $N_C = N_A.Slink[k]$
if$N_C.Clink == NULL$ then{if !(N_C.constr(s_{ij})) then $N_C.CSV++$;}
else{ N_O=$N_C.Clink$; globalconstr=composeconstr(N_O, s_{ij})
if !(globalconstr) then $N_O.CSV++$}}
End

Figure 5.2: Evaluation Algorithm

Chapter 6

Evaluation Framework

As shown by the approaches reported Chapter 2, and especially by the approach by Bertino et al. [10], in many real world applications, it is necessary to take into account even other parameters that are not directly related to DQ. On the other hand we believe that DQ should represent the invariant of a PPDM evaluation and should be used to identify the best algorithm within a set of previously selected "Best Algorithms". To preselect this "best set", we suggest to use some parameters as discriminant to select the algorithms that have an acceptable behavior under some aspects generally considered relevant especially in "production environments" (efficiency, scalability, hiding failure and level of privacy). In order to understand if these four parameters are sufficient to identify an acceptable set of candidates, we performed an evaluation test. We identified a starting set of PPDM algorithms for Association Rules Hiding (the algorithms presented in [97] and a new set of three algorithms based on data fuzzification [10]). Then, by using the IBM synthetic data generator[1]we generated a categoric database representing an hypothetical Health Database storing the different therapies associated with the patients. We also built the associated DMG. On this database, we applied the different algorithms and then we measured the previous parameters. Once we built the "Best Set" we discovered that some algorithms that performed less changes to the database, which in some way indicates a better quality, are not in this set. A reason is for example a low efficiency. For this reason we believe that even in the preselection phase a "coarse" DQ parameter must be introduced. In our opinion, the *Coarse DQ Measure* depends on the specific class of PPDM algorithms. If the algorithms adopt a perturbation or a blocking technique, the coarse DQ can be measured by the dissimilarity between the original dataset D and the sanitized one D' by measuring, for example, in the case of transactional datasets, the difference between the item frequencies of the two datasets before and after the sanitization. Such dissimilarity can be estimated by the following expression:

$$Diss(D, D') = \frac{\sum_{i=1}^{n} |f_D(i) - f_{D'}(i)|}{\sum_{i=1}^{n} f_D(i)} \tag{6.1}$$

[1]http://www.almaden.ibm.com/software/quest/Resources/datasets/syndata.html

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'. The same method can be used, extending the previous formula, also in the case of blocking techniques. If the data modification consists of aggregating some data values, the coarse DQ is given by the loss of detail in the data. As in the case of the $k-$Anonymity algorithm [92], given a database DB with N_A attributes and N transactions, if we identify as generalization scheme a domain generalization hierarchy GT with a depth h, it is possible to measure the coarse quality of a sanitized database SDB as:

$$Quality(SDB) = 1 - \frac{\sum_{i=1}^{i=N_A} \sum_{j=1}^{i=N} \frac{h}{|GT_{Ai}|}}{|DB| * |N_A|} \tag{6.2}$$

where $\frac{h}{|GT_{Ai}|}$ represent the detail loss for each cell sanitized.

Once we have identified the *Best Set* we are able to apply our DQ-driven evaluation.

We now present a three steps Evaluation Framework based on the previous concepts.

1. A set of "Interesting" PPDM's is selected. These algorithms are tested on a generic database and evaluated according the general parameters (Efficiency, Scalability, Hiding failure, Coarse Data Quality, Level of privacy). The result of this step is a restricted set of *Candidate algorithms*

2. A test database with the same characteristics of the target database is generated. An IQM schema with the AIS and the related DMG is the result of this step.

3. The Information Driven DQ Evaluation Algorithm is applied in order to identify the algorithm that finally will be applied.

As it is probably obvious to the readers, the most "time consuming" step in terms of required user interactions is step 2. The design a good IQM is the core of our evaluation framework. We believe that a top down approach is, in this cases, the most appropriate. More in detail, the first task should be the identification of the high level information that is relevant and for which we are interested in measuring the impact of PPDM algorithms. It could also can be useful to involve in this task some authorized users (e.g. in case of Health DBA's, doctors, etc.) in order to understand all the possible uses of the database and the relevance of the retrieved information. The use of datamining tools could be useful to identify non-evident aggregate information.
A second task would then, given the high level information, determine the different constraints (both simple and complex) and evaluate their relevance. Also in this case, discussions with authorized users and DB designers, and the use of DM tools (e.g. discover association rules) could help to build a good IQM. Finally it is necessary, by taking into account all the previous information, to rate the relevance of the attributes involved. This top down analysis is useful

not only for the specific case of PPDM evaluation, but, if well developed, is a powerful tool to understand the real information contents, its value and the relation between the information stored in a given database. In the context of an "Information Society" this is a non negligible added value.

6.1 A three-steps framework for the evaluation of PPDM algorithms

As shown by the approaches reported in the state of the art, and especially by the approach by Bertino et al. [10], in many real world applications, it is necessary to take into account also other parameters that are not directly related to DQ. On the other hand we believe that DQ should represent the invariant of a PPDM evaluation and should be used to identify the best algorithm within a set of previously selected "Best Algorithms". To preselect this "best set", we suggest to use the *Operational Parameters* presented before as discriminant to select the algorithms that have an acceptable behavior under some aspects generally considered relevant especially in "production environments" (efficiency, scalability, hiding failure and level of privacy).

In order to understand if these four parameters are sufficient to identify an acceptable set of candidates, we performed an evaluation test. We identified a starting set of PPDM algorithms for Association Rules Hiding (the algorithms presented in [97] and a new set of three algorithms based on data fuzzification [10]). Then, by using the IBM synthetic data generator[2] we generated a categoric database representing an hypothetical Health database storing the different therapies associated with the patients. We also built the associated DMG. On this database, we applied the different algorithms and then we measured the previous parameters. Once we built the "Best Set" we discovered that some algorithms that performed less changes to the database, which in some way indicates a better quality, are not in this set. A reason is for example a low efficiency. For this reason we believe that even in the preselection phase a "coarse" DQ parameter must be introduced. In our opinion, the *Coarse DQ Measure* depends on the specific class of PPDM algorithms. If the algorithms adopt a perturbation or a blocking technique, the coarse DQ can be measured by the dissimilarity between the original dataset D and the sanitized one D' by measuring, for example, in the case of transactional datasets, the difference between the item frequencies of the two datasets before and after the sanitization. Such dissimilarity can be estimated by the following expression:

$$Diss(D, D') = \frac{\sum_{i=1}^{n} |f_D(i) - f_{D'}(i)|}{\sum_{i=1}^{n} f_D(i)} \tag{6.3}$$

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'. The same method can be used, extending the previous formula, also in the case of blocking techniques. If the data modification consists

[2] http://www.almaden.ibm.com/software/quest/Resources/datasets/syndata.html

of aggregating some data values, the coarse DQ is given by the loss of detail in the data. As in the case of the $k-$Anonymity algorithm [92], given a database DB with N_A attributes and N transactions, if we identify as generalization scheme a domain generalization hierarchy GT with a depth h, it is possible to measure the coarse quality of a sanitized database SDB as:

$$Quality(SDB) = 1 - \frac{\sum_{i=1}^{i=N_A} \sum_{j=1}^{i=N} \frac{h}{|GT_{Ai}|}}{|DB| * |N_A|} \tag{6.4}$$

where $\frac{h}{|GT_{Ai}|}$ represent the detail loss for each cell sanitized. Once we have identified the *Best Set* we are able to apply our DQ-driven evaluation.

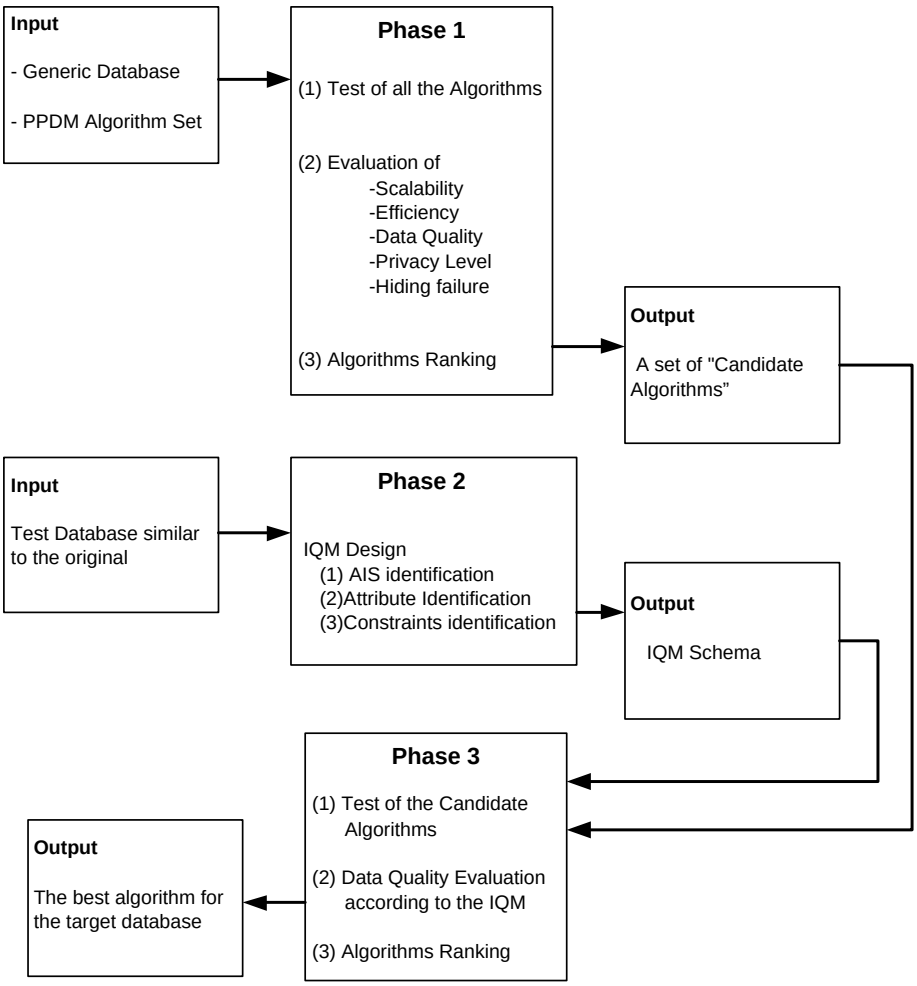

Figure 6.1: The Evaluation Framework

We now present a three steps Evaluation Framework based on the previous concepts.

1. A set of "Interesting" PPDM's is selected. These algorithms are tested on a generic database and evaluated according the general parameters (Efficiency, Scalability, Hiding failure, Coarse Data Quality, Level of privacy). The result of this step is a restricted set of *Candidate algorithms*

2. A test database with the same characteristics of the target database is generated. An IQM schema with the AIS and the related DMG is the result of this step.

3. The Information Driven DQ Evaluation Algorithm is applied in order to

identify the algorithm that finally will be applied.

As it is probably obvious to the readers, the most "time consuming" step in terms of required user interactions is step 2. The design of a good IQM is the core of our evaluation framework. We believe that a top down approach is, in this cases, the most appropriate. More in detail, the first task should be the identification of the high level information that is relevant and for which we are interested in measuring the impact of PPDM algorithms. It could also can be useful to involve in this task some authorized users (e.g. in case of Health DBA's, doctors, etc.) in order to understand all the possible uses of the database and the relevance of the retrieved information. The use of datamining tools could be useful to identify non-evident aggregate information.
A second task would then be, given the high level information, to determine the different constraints (both simple and complex) and evaluate their relevance. Also in this case, discussions with authorized users and DB designers, and the use of DM tools (e.g. discover association rules) could help to build a good IQM. Finally it is necessary, by taking into account all the previous information, to rate the relevance of the attributes involved. This top down analysis is useful not only for the specific case of PPDM evaluation, but, if well developed, is a powerful tool to understand the real information contents, its value and the relation between the information stored in a given database. In the context of an "Information Society" this is a non negligible added value.

Chapter 7

Conclusions

The existence of a complete framework allowing to evaluate in a general and complete way a PPDM algorithm, is a mandatory condition if we want to really use these algorithms in the real world over real applications. In this report we have presented the last results obtained in this field.

A PPDM algorithm is conceived to protect sensitive information from an unauthorized disclosure. This automatically implies that the PPDM Algorithms are strongly related to the concept of Privacy. For this reason we explored this concept, and we developed a new approach based on the Shannon Entropy in order to give a measure of the Privacy introduced by the use of a PPDM algorithm. Moreover we gave an instantiation of this measure in the context of Association Rule Hiding and in the context of Classification Hiding. We plan to extend this study in the direction of the cluster hiding.

It is, however, not possible to measure the PPDM algorithm only in term of the performance (i.e. scalability, efficiency etc.). Due to the intrinsic information driven nature of the PPDM algorithms, in fact, their impact on the different databases, may be extremely different. For this reason, in order to be an instrument useful to select the right algorithm for a particular situation, such a framework must be able to magnify the behavior of the PPDM algorithms over a target database taking into consideration even its impact on the DQ of the sanitized database.

In this report, we have proposed an approach based on the concept of DQ as main discriminant in the PPDM Algorithm evaluation. We introduced some algorithms in order to perform this task and we presented a three-steps methodology allowing to take in consideration even the other type of parameters more related to the "pure performance" provided. Also in this case we gave an implementation of this concept conceived for association rule hiding, but, in this case, the algorithms can be easily adapted for the evaluation of every type of PPDM techniques.

Bibliography

[1] N. Adam and J. Worthmann, *Security-control methods for statistical databases: a comparative study.* ACM Comput. Surv., Volume 21(4), pp. 515-556, year 1989, ACM Press.

[2] D. Agrawal and C. C. Aggarwal, *On the Design and Quantification of Privacy Preserving Data Mining Algorithms.* In Proceedings of the 20*th* ACM Symposium on Principle of Database System, pp. 247-255, year 2001, ACM Press.

[3] R. Agrawal, T. Imielinski and A. Swami, *Mining Association Rules between Sets of Items in Large Databases.* Proceedings of ACM SIGMOD, pp. 207-216, May 1993, ACM Press.

[4] R. Agrawal and R. Srikant, *Privacy Preserving Data Mining.* In Proceedings of the ACM SIGMOD Conference of Management of Data, pp. 439-450, year 2000, ACM Press.

[5] R. Agrawal and R. Srikant, *Fast algorithms for mining association rules.* In Proceeding of the 20*th* International Conference on Very Large Databases, Santiago, Chile, June 1994, Morgan Kaufmann.

[6] G. M. AMDAHL, *Validity of the Single-Processor Approach to Achieving Large Scale Computing Capabilities.* AFIPS Conference Proceedings(April 1967),pp. 483-485, Morgan Kaufmann Publishers Inc.

[7] M. J. Atallah, E. Bertino, A. K. Elmagarmid, M. Ibrahim and V. S. Verykios, *Disclosure Limitation of Sensitive Rules.* In Proceedings of the IEEE Knolwedge and Data Engineering Workshop, pp. 45-52, year 1999, IEEE Computer Society.

[8] D. P. Ballou, H. L. Pazer, *Modelling Data and Process Quality in Multi Input, Multi Output Information Systems.* Management science, Vol. 31, Issue 2, pp. 150-162, (1985).

[9] Y. Bar-Hillel, *An examination of information theory.* Philosophy of Science, volume 22, pp.86-105, year 1955.

[10] E. Bertino, I. Nai Fovino and L. Parasiliti Provenza, *A Framework for Evaluating Privacy Preserving Data Mining Algorithms*. Data Mining and Knowledge Discovery Journal, year 2005, Kluwert.

[11] E.Bertino and I.Nai Fovino, *Information Driven Evaluation of Data Hiding Algorithms. 7th* International Conference on Data Warehousing and Knowledge Discovery. Copenaghen, August 2005, Springer-Verlag.

[12] N. M. Blachman, *The amount of information that y gives about X.* IEEE Truns. Inform. Theon. vol. IT-14, pp. 27-31. Jan. 1968, IEEE Press.

[13] L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification of Regression Trees.* Wadsworth International Group, year 1984.

[14] S. Brin, R. Motwani, J. D. Ullman and S. Tsur, *Dynamic itemset counting and implication rules for market basket data.* In Proc. of the ACM SIGMOD International Conference on Management of Data, year 1997, ACM Press.

[15] L. Chang and I. S. Moskowitz, *Parsimonious downgrading and decision trees applied to the inference problem.* In Proceedings of the 1998 New Security Paradigms Workshop, pp.82-89, year 1998, ACM Press.

[16] P. Cheeseman and J. Stutz, *Bayesian Classification (AutoClass): Theory and Results.* Advances in Knowledge Discovery and Data Mining, AAAI Press/MIT Press, year 1996.

[17] M. S. Chen, J. Han and P. S. Yu, *Data Mining: An Overview from a Database Perspective.* IEEE Transactions on Knowledge and Data Engineering, vol. 8 (6), pp. 866-883, year 1996, IEEE Educational Activities Department.

[18] F. Y. Chin and G. Ozsoyoglu, *Auditing and inference control in statistical databases.* IEEE Trans. Softw. Eng. SE-8, 6 (Apr.), pp. 574-582, year 1982, IEEE Press.

[19] F. Y. Chin and G. Ozsoyoglu, *Statistical database design.* ACM Trans. Database Syst. 6, 1 (Mar.), pp. 113-139, year 1981, ACM Press.

[20] L. H. Cox, *Suppression methodology and statistical disclosure control.* J. Am. Stat. Assoc. 75, 370 (June), pp. 377-385, year 1980.

[21] E. Dasseni, V. S. Verykios, A. K. Elmagarmid and E. Bertino, *Hiding Association Rules by using Confidence and Support.* in proceedings of the *4th* Information Hiding Workshop, pp. 369383, year 2001, Springer-Verlag.

[22] D. Defays, *An efficient algorithm for a complete link method.* The Computer Journal, 20, pp. 364-366, 1977.

[23] D. E. Denning and J. Schlorer, *Inference control for statistical databases.* Computer 16 (7), pp. 69-82, year 1983 (July), IEEE Press.

[24] D. Denning, *Secure statistical databases with random sample queries.* ACM TODS, 5, 3, pp. 291-315, year 1980.

[25] D. E. Denning, *Cryptography and Data Security.* Addison-Wesley, Reading, Mass. 1982.

[26] V. Dhar, *Data Mining in finance: using counterfactuals to generate knowledge from organizational information systems.* Information Systems, Volume 23, Number 7, pp. 423-437(15), year 1998.

[27] J. Domingo-Ferrer and V. Torra, *A Quantitative Comparison of Disclosure Control Methods for Microdata.* Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 113-134, P. Doyle, J. Lane, J. Theeuwes, L. Zayatz ed., North-Holland, year 2002.

[28] P. Domingos and M. Pazzani, *Beyond independence: Conditions for the optimality of the simple Bayesian classifier.* Proceedings of the Thirteenth International Conference on Machine Learning, pp. 105-112, San Francisco, CA, year 1996, Morgan Kaufmann.

[29] P. Drucker, *Beyond the Information Revolution.* The Atlantic Monthly, 1999.

[30] P. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* Wiley, year 1973, New York.

[31] G. T. Duncan, S. A. Keller-McNulty and S. L. Stokes, *Disclosure risks vs. data utility: The R-U confidentiality map.* Tech. Rep. No. 121. National Institute of Statistical Sciences. 2001

[32] C. Dwork and K. Nissim, *Privacy preserving data mining in vertically partitioned database.* In Crypto 2004, Vol. 3152, pp. 528-544.

[33] D. L. EAGER, J. ZAHORJAN and E. D. LAZOWSKA, *Speedup Versus Efficiency in Parallel Systems.* IEEE Trans. on Computers, C-38, 3 (March 1989), pp. 408-423, IEEE Press.

[34] L. Ertoz, M. Steinbach and V. Kumar, *Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data.* In Proceeding to the SIAM International Conference on Data Mining, year 2003.

[35] M. Ester, H. P. Kriegel, J. Sander and X. XU, *A density-based algorithm for discovering clusters in large spatial databases with noise.* In Proceedings of the 2nd ACM SIGKDD, pp. 226-231, Portland, Oregon, year 1996, AAAI Press.

[36] A. Evfimievski, *Randomization in Privacy Preserving Data Mining.* SIGKDD Explor. Newsl., vol. 4, number 2, year 2002, pp. 43-48, ACM Press.

[37] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, *Privacy Preserving Mining of Association Rules.* In Proceedings of the 8*th* ACM SIGKDDD International Conference on Knowledge Discovery and Data Mining, year 2002, Elsevier Ltd.

[38] S. E. Fahlman and C. Lebiere, *The cascade-correlation learning architecture.* Advances in Neural Information Processing Systems 2, pp. 524-532. Morgan Kaufmann, year 1990.

[39] R. P. Feynman, R. B. Leighton and M. Sands, *The Feynman Lectures on Physics, v I.* Reading, Massachusetts: Addison-Wesley Publishing Company, year 1963.

[40] S. Fortune and J. Wyllie, *Parallelism in Random Access Machines.* Proc. Tenth ACM Symposium on Theory of Computing(1978), pp. 114-118, ACM Press.

[41] W. Frawley, G. Piatetsky-Shapiro and C. Matheus, *Knowledge Discovery in Databases: An Overview.* AI Magazine, pp. 213-228, year 1992.

[42] S. P. Ghosh, *An application of statistical databases in manufacturing testing.* IEEE Trans. Software Eng. 1985. SE-11, 7, pp. 591-596, IEEE press.

[43] S.P.Ghosh, *An application of statistical databases in manufacturing testing.* In Proceedings of IEEE COMPDEC Conference, pp. 96-103,year 1984, IEEE Press.

[44] S. Guha, R. Rastogi and K. Shim, *CURE: An efficient clustering algorithm for large databases.* In Proceedings of the ACM SIGMOD Conference, pp. 73-84, Seattle, WA. 1998, ACM Press.

[45] S. Guha, R. Rastogi and K. Shim, *ROCK: A robust clustering algorithm for categorical attributes.* In Proceedings of the 15*th* ICDE, pp. 512-521, Sydney, Australia, year 1999, IEEE Computer Society.

[46] J. Han and M. Kamber, *Data Mining: Concepts and Techniques.* The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers, August 2000.

[47] J. Han, J. Pei and Y. Yin, *Mining frequent patterns without candidate generation.* In Proceeding of the 2000 ACM-SIGMOD International Conference on Management of Data, Dallas, Texas, USA, May 2000, ACM Press.

[48] M. A. Hanson and R. L. Brekke, *Workload management expert system - combining neural networks and rule-based programming in an operational application.* In Proceedings Instrument Society of America, pp. 1721-1726, year 1988.

[49] J. Hartigan and M. Wong, *Algorithm AS136: A k-means clustering algorithm.* Applied Statistics, 28, pp. 100-108, year 1979.

[50] A. Hinneburg and D. Keim, *An efficient approach to clustering large multimedia databases with noise.* In Proceedings of the 4*th* ACM SIGKDD, pp. 58-65, New York, year 1998, AAAI Press.

[51] T. Hsu, C. Liau and D.Wang, *A Logical Model for Privacy Protection.* Lecture Notes in Computer Science, Volume 2200, Jan 2001, pp. 110-124, Springer-Verlag.

[52] IBM Synthetic Data Generator. http://www.almaden.ibm.com/software/quest/Resources/datasets/syndata.html

[53] M. Kantarcioglu and C. Clifton, *Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data.* In Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 24-31, year 2002, IEEE Educational Activities Department.

[54] G. Karypis, E. Han and V. Kumar, *CHAMELEON: A hierarchical clustering algorithm using dynamic modeling.* COMPUTER, 32, pp. 68-75, year 1999.

[55] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley and Sons, New York, year 1990.

[56] W. Kent, *Data and reality.* North Holland, New York, year 1978.

[57] S. L. Lauritzen, *The em algorithm for graphical association models with missing data.* Computational Statistics and Data Analysis, 19 (2), pp. 191-201, year 1995, Elsevier Science Publishers B. V.

[58] W. Lee and S. Stolfo, *Data Mining Approaches for Intrusion Detection.* In Proceedings of the Seventh USENIX Security Symposium (SECURITY '98), San Antonio, TX, January 1998.

[59] A. V. Levitin and T. C. Redman, *Data as resource: properties, implications and prescriptions.* Sloan Management review, Cambridge, Vol. 40, Issue 1, pp. 89-101, year 1998.

[60] Y. Lindell and B. Pinkas, *Privacy Preserving Data Mining.* Journal of Cryptology, vol. 15, pp. 177-206, year 2002, Springer Verlag.

[61] R. M. Losee, *A Discipline Independent Definition of Information.* Journal of the American Society for Information Science 48 (3), pp. 254-269, year 1997.

[62] M. Masera, I. Nai Fovino, R. Sgnaolin *A Framework for the Security Assessment of Remote Control Applications of Critical Infrastructure* 29th ESReDA Seminar "Systems Analysis for a More Secure World", year 2005

[63] G. MClachlan and K. Basford, *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York, year 1988.

[64] M. Mehta, J. Rissanen and R. Agrawal, *MDL-based decision tree pruning.* In Proc. of KDD, year 1995, AAAI Press.

[65] G. L. Miller, *Resonance, Information, and the Primacy of Process: Ancient Light on Modern Information and Communication Theory and Technology.* PhD thesis, Library and Information Studies, Rutgers, New Brunswick, N.J., May 1987.

[66] I. S. Moskowitz and L. Chang, *A decision theoretical based system for information downgrading.* In Proceedings of the 5*th* Joint Conference on Information Sciences, year 2000, ACM Press.

[67] S. R. M. Oliveira and O. R. Zaiane, *Toward Standardization in Privacy Preserving Data Mining.* ACM SIGKDD 3rd Workshop on Data Mining Standards, pp. 7-17, year 2004, ACM Press.

[68] S. R. M. Oliveira and O. R. Zaiane, *Privacy Preserving Frequent Itemset Mining.* Proceedings of the IEEE international conference on Privacy, security and data mining, pp. 43-54, year 2002, Australian Computer Society, Inc.

[69] S. R. M. Oliveira and O. R. Zaiane, *Privacy Preserving Clustering by Data Transformation.* In Proceedings of the 18*th* Brazilian Symposium on Databases, Manaus, Amazonas, Brazil, pp. 304-318, year 2003.

[70] K. Orr, *Data Quality and System Theory.* Comm. of the ACM, Vol. 41, Issue 2, pp. 66-71, Feb. 1998, ACM Press.

[71] M. A. Palley and J. S. Simonoff, *The use of regression methodology for compromise of confidential information in statistical databases.* ACM Trans. Database Syst. 12,4 (Dec.), pp. 593-608, year 1987.

[72] J. S. Park, M. S. Chen and P. S. Yu, *An Effective Hash Based Algorithm for Mining Association Rules.* Proceedings of ACM SIGMOD, pp. 175-186, May, 1995, ACM Press.

[73] Z. Pawlak, *Rough SetsTheoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, 1991.

[74] G. Piatetsky-Shapiro, *Discovery, analysis, and presentation of strong rules.* Knowledge Discovery in Databases, pp. 229-238, AAAI/MIT Press, year 1991.

[75] A. D. Pratt, *The Information of the Image.* Ablex, Norwood, NJ, 1982.

[76] J. R. Quinlan, *C4.5: Programs for Machine Learning.* Morgan Kaufmann, year 1993.

[77] J. R. Quinlan, *Induction of decision trees.* Machine Learning, vol. 1, pp. 81-106, year 1986, Kluwer Academic Publishers.

[78] R. Rastogi and S. Kyuseok, *PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning.* Data Mining and Knowledge Discovery, vol. 4, n.4, pp. 315-344, year 2000.

[79] R. Rastogi and K. Shim, *Mining Optimized Association Rules with Categorical and Numeric Attributes.* Proc. of International Conference on Data Engineering, pp. 503-512, year 1998.

[80] H. L. Resnikoff, *The Illusion of Reality.* Springer-Verlag, New York, 1989.

[81] S. J. Rizvi and J. R. Haritsa, *Maintaining Data Privacy in Association Rule Mining.* In Proceedings of the 28*th* International Conference on Very Large Databases, year 2003, Morgan Kaufmann.

[82] S. J. Russell, J. Binder, D. Koller and K. Kanazawa, *Local learning in probabilistic networks with hidden variables.* In International Joint Conference on Artificial Intelligence, pp. 1146-1152, year 1995, Morgan Kaufmann.

[83] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning internal representations by error propagation.* Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations pp. 318–362, Cambridge, MA: MIT Press, year 1986.

[84] G. Sande, *Automated cell suppression to reserve confidentiality of business statistics.* In Proceedings of the 2nd International Workshop on Statistical Database Management, pp. 346-353, year 1983.

[85] A. Savasere, E. Omiecinski and S. Navathe, *An efficient algorithm for mining association rules in large databases.* In Proceeding of the Conference on Very Large Databases, Zurich, Switzerland, September 1995, Morgan Kaufmann.

[86] J. Schlorer, *Information loss in partitioned statistical databases.* Comput. J. 26, 3, pp. 218-223, year 1983, British Computer Society.

[87] C. E. Shannon, *A Mathematical Theory of Communication.* Bell System Technical Journal, vol. 27,(July and October),1948, pp.379423, pp. 623-656.

[88] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, Ill. 1949.

[89] A. Shoshani, *Statistical databases: characteristics,problems, and some solutions.* Proceedings of the Conference on Very Large Databases (VLDB), pp.208-222, year 1982, Morgan Kaufmann Publishers Inc.

[90] R. SIBSON, *SLINK: An optimally efficient algorithm for the single link cluster method.* Computer Journal, 16, pp. 30-34, year 1973.

[91] P. Smyth and R. M. Goodman, *An information theoretic Approach to Rule Induction from databases.* IEEE Transaction On Knowledge And Data Engineering, vol. 3, n.4, August,1992, pp. 301-316, IEEE Press.

[92] L. Sweeney, *Achieving* k*-Anonymity Privacy Protection using Generalization and Suppression.* International Jurnal on Uncertainty, Fuzzyness and Knowledge-based System, pp. 571-588, year 2002, World Scientific Publishing Co., Inc.

[93] R. Srikant and R. Agrawal, *Mining Generalized Association Rules.* Proceedings of the 21*th* International Conference on Very Large Data Bases, pp. 407-419, September 1995, Morgan Kaufmann.

[94] G. K. Tayi, D. P. Ballou, *Examining Data Quality.* Comm. of the ACM, Vol. 41, Issue 2, pp. 54-58, year 1998, ACM Press.

[95] M. Trottini, *A Decision-Theoretic Approach to data Disclosure Problems.* Research in Official Statistics, vol. 4, pp. 722, year 2001.

[96] M. Trottini, *Decision models for data disclosure limitation.* Carnegie Mellon University, Available at `http://www.niss.org/dgii/TR/ThesisTrottini -final.pdf`, year 2003.

[97] University of Milan - Computer Technology Institute - Sabanci University *Codmine* IST project. 2002-2003.

[98] J. Vaidya and C. Clifton, *Privacy Preserving Association Rule Mining in Vertically Partitioned Data.* In Proceedings of the 8*th* ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639-644, year 2002, ACM Press.

[99] V. S. Verykios, E. Bertino, I. Nai Fovino, L. Parasiliti, Y. Saygin, Y. Theodoridis, *State-of-the-art in Privacy Preserving Data Mining.* SIGMOD Record, 33(1) pp. 50-57, year 2004, ACM Press.

[100] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, *Association Rule Hiding.* IEEE Transactions on Knowledge and Data Engineering, year 2003, IEEE Educational Activities Department.

[101] C. Wallace and D. Dowe, *Intrinsic classification by MML. The Snob program.* In the Proceedings of the 7*th* Australian Joint Conference on Artificial Intelligence, pp. 37- 44, UNE, World Scientific Publishing Co., Armidale, Australia, 1994.

[102] G. J. Walters, *Philosophical Dimensions of Privacy: An Anthology.* Cambridge University Press, year 1984.

[103] G. J. Walters, *Human Rights in an Information Age: A Philosophical Analysis.* chapter 5, University of Toronto Press, year 2001.

[104] Y. Wand and R. Y. Wang, *Anchoring Data Quality Dimensions in Ontological Foundations.* Comm. of the ACM, Vol. 39, Issue 11, pp. 86-95, Nov. 1996, ACM Press.

[105] R. Y. Wang and D. M. Strong, *Beyond Accuracy: what Data Quality Means to Data Consumers.* Journal of Management Information Systems Vol. 12, Issue 4, pp. 5-34, year 1996.

[106] L. Willenborg and T. De Waal, *Elements of statistical disclosure control.* Lecture Notes in Statistics Vol.155, Springer Verlag, New York.

[107] N. Ye and X. Li, *A Scalable Clustering Technique for Intrusion Signature Recognition.* 2001 IEEE Man Systems and Cybernetics Information Assurance Workshop, West Point, NY, June 5-6, year 2001, IEEE Press.

[108] M. J. Zaki, S. Parthasarathy, M. Ogihara and W. Li, *New algorithms for fast discovery of association rules* In Proceeding of the 8rd International Conference on KDD and Data Mining, Newport Beach, California, August 1997, AAAI Press.

European Commission

EUR 23069 EN– Joint Research Centre – Institute for the Protection and Security of the Citizen
Title: Privacy Preserving Data Mining, Evaluation Methodologies
Author(s): Igor Nai Fovino and Marcelo Masera
Luxembourg: Office for Official Publications of the European Communities
2008 – 58 pp. – 21 x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593

Abstract
Privacy is one of the most important properties an information system must satisfy. A relatively new trend shows that classical access control techniques are not sufficient to guarantee privacy when datamining techniques are used. Privacy Preserving Data Mining (PPDM) algorithms have been recently introduced with the aim of modifying the database in such a way to prevent the discovery of sensible information. Due to the large amount of possible techniques that can be used to achieve this goal, it is necessary to provide some standard evaluation metrics to determine the best algorithms for a specific application or context. Currently, however, there is no common set of parameters that can be used for this purpose. Moreover, because sanitization modifies the data, an important issue, especially for critical data, is to preserve the quality of data. However, to the best of our knowledge, no approaches have been developed dealing with the issue of data quality in the context of PPDM algorithms. This report explores the problem of PPDM algorithm evaluation, starting from the key goal of preserving of data quality. To achieve such goal, we propose a formal definition of data quality specifically tailored for use in the context of PPDM algorithms, a set of evaluation parameters and an evaluation algorithm. Moreover, because of the ``environment related'' nature of data quality, a structure to represent constraints and information relevance related to data is presented. The resulting evaluation core process is then presented as a part of a more general three step evaluation framework, taking also into account other aspects of the algorithm evaluation such as efficiency, scalability and level of privacy.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.