Survey on Privacy Preserving in Data Mining Tasks

Dr. Ahmad Haj Yasiens
Faculty of Engineering
Al- Ahlyya Amman University

Dr . Abdelraouf Ishtaiwi
Faculty of Sciences
Petra University

Abstract. Data mining technology that reveals knowledge in large databases could compromise the information that individuals or parties regard as private (sensitive). The aim is to find the right balance between maximiz-
ing analysis results (that are useful for the common good) and keeping the inferences that disclose sensitive information about parties or indi- viduals at a minimum. This field is called privacy preserving data mining. Many algorithms and techniques have been proposed to solve this prob- lem but the challenge is still open for new algorithms. better algorithms and even better understanding and standardization for the problem of privacy preserving data mining. This paper presents a survey of privacy preserving in data mining tasks.

Keywords: Privacy preserving data mining. secure multi-party computation. sanitizing algorithms. data sanitization. rule sanitization. sensitive knowl- edge.

تاريخ استلام البحث:2/4/ 2008 تاريخ قبول البحث:23 / 12 / 2008

مسح في المحافظات على السرية في مهمات التنقيب عن البيانات

ملخص:

تكنولوجيا التنقيب في البيانات التي تكشف معاومات

من قواعد بيانات ضخمة ربما تعرض الاشخاص أو المؤسسات لمخاطر أو مشاكل لأن المعلومات المكتشفة تعتبر سرية ولا ينبغي معرفتها .

الهدف هو توازن مابين كشف المعلومات العامة التي قد

تفيد جميع الأطراف وإخفاء المعلومات التي قد تضر بسرية أي طرف ، لقد اقترحت العديد من الخوارزميات لحل هذه المشكلة وإيجاد هذا التوازن ولكن المجال ما يزال متاحاً للمزيد من الخوارزميات الجديدة أو تحسين القديمة بالإضافة الى المسح في المحافظة على السرية في مهمات التنقيب عن البيانات ، يقدم هذا البحث تصنيفاً جديداً للمشاكل والحلول ويظهر بعض المشاكل في التصنيفات القديمة .

1 Introduction

The amount of data kept in computer files is growing at a phenomenal rate. The data mining field offers to discover unknown knowledge. Data mining is often defined as the process of discovering meaningful, new correlation patterns and trends through non trivial extraction of implicit, previously unknown information from large amount of data stored in repositories using pattern recognition as well as statistical and mathematical techniques [20]. A SQL query is usually stated or written to retrieve specific data while data miners might not even be exactly sure of what they require.

If the data is personal or corporate data, data mining offers the potential to reveal what others regard as sensitive (private). In some cases, it may be of mutual benefit for two parties (even competitors) to share their data for an analysis task. However, they would like to ensure their own data remains private. In other words, there is a need to protect sensitive knowledge during a data mining process. This problem is called Privacy Preserving Data Mining (PPDM).

Most organizations may be very clear about what constitutes examples of sensitive knowledge. However, what is challenging is to identify what is non-sensitive knowledge because there are many inference channels available to adversaries. It may be possible that making some knowledge public (because perceived as not sensitive), allows an adversary to infer sensitive knowledge. In fact, part of the challenge is to identify the largest set of non-sensitive knowledge that can be disclosed under all inference channels. However, what complicates matter further is that knowledge may be statements with possibility of truth, certainty or confidence. Thus, the only possible avenue is to ensure that the adversary will learn the statements with very little certainty.

This paper is organized as follows. In the next section, we present a background where we introduce crucial topics that participate to wards better understanding of this paper. In Section 3, we present a new taxonomy of where privacy preserving data mining techniques can be applied. In Section 4, we introduce a new look at the state-of-the-art in privacy preserving data mining. We classify all the problems and solutions (to the best of our knowledge) in the PPDM field under our three levels discussed in the taxonomy. Finally, Section 5 presents a summary and conclusion.

2 Background

2.1 Defining Privacy

In the literature there are many definitions for privacy. An 1890 paper [48] defined privacy as "the right to be alone". Later, in a paper published in 1967 [49], privacy was defined as "the desire of people to choose freely under what circumstances and to what extent they will expose themselves, their attitude, and their behavior to others". Also, Schoeman [43] defined privacy as "the right to determine what (personal) information is communicated to others" or "the control an individual has over information about himself or herself". One of the most recent definitions of privacy was presented by Garfinkel [22] stated that "privacy is about self-possession, autonomy, and integrity".

We notice that the above definitions are based on social and cultural concepts. However, with data explosion and Internet and web technology emerging, privacy is now a problem in the technology field [39].

In the information technology era, privacy refers to the right of users to conceal their personal information and have some degree of control over the use of any personal information disclosed to others [10, 2, 28, 1]. Clifton et al. presented a definition for privacy preserving data mining [8], "getting valid data mining results without learning the underlying data values". Recently, Oliveira et al. defined privacy preserving data mining based on [43, 8]as follows "PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results". we believe that this last definition could be the best general definition for privacy preserving data mining.

Oliveira et al. [35] mentions that privacy preservation occurs in two major dimensions.

- Individual privacy preservation: The essential goal of data privacy is to protect personal data. No one should be able to link directly or indirectly any public data to individuals. That is why private attributes are usually eliminated from database before the mining process. Miners in this case are pushed to learn general patterns rather than specific characteristics of individuals.
- Collective privacy preservation: Usually, it is not enough to protect personal data. We need to protect against learning sensitive knowledge representing the activities of a group. The aim here is very similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, should prevent disclosure of confidential information about individuals. However, the goal here is not only to protect personally identifiable information but also some patterns and trends that are not supposed to be discovered.

We believe that the above categorization where the privacy preserving data mining occurs is not very clear. In our opinion, privacy preserving data mining occurs in two dimensions inside each level of the three levels presented earlier in Section 3 and these two dimensions are:

- Individuals: This dimension involves implementing a privacy preserving data mining technique to protect the privacy of one individual or more whose their data is going to be published to the public. An example of this dimension is patients records or the census.
- PPDMSMC (Privacy Preserving Data Mining in Secure Multi-party Computation): This dimension involves protecting the privacy of two or more parties who want to perform a data mining task on the union of their private data. An example of this dimension is two parties who want to cluster the union of their private data.

Before presenting our classification, we want to discuss the secure multi-party computation problem which will help to understand many of the privacy preserving data mining algorithms and techniques presented under our classification.

2.2 Secure Multi-Party Computation

In the literature, the problem of parties who want to perform a computation on the union of their private data but do not trust each other and each party wants to hide its data from the other parties is referred to as the secure multi-party computation (SMC). Goldwasser defined the SMC problem as a problem that deals with computing any function on any input, in a distributed network where each participant holds one of the inputs, while ensuring that no more information is revealed to a participant in the computation than can be inferred from that participant's input and output [25].

There have been many solutions and algorithms to perform a secure multi-party computation and solve the problem. Most of the solutions assume that one of the parties should be trusted with the inputs somehow - usually encrypted or

modified in a way that will not affect the final results- and then that party will do the computation and distribute the results to the other parties.

The SMC problem literature is extensive, having been introduced by Yao [51] and expanded by Goldreich, Micali, and Wigderson [24] and others [21]. Yao introduced the secure multi-party computation with a problem of two millionaire's who want to know who is richer without disclosing their net worth to each other [52]. Goldreich also proved that for any function, there is a secure multi-party computation solution [23].

Kantarcioglu et al. [29] mentions that secure multi-party computation has concentrated on two security models. The first one is the semi honest model, which assumes that each party follows the rules of the protocol but during the execution of the protocol, each party has the freedom to use what it sees to compromise security. A detailed discussion of the theorems and proofs of the semi-honest model can be found in [23]. The second security model is the malicious model, which assumes that any party could be deceiving or cheating but that will not affect the computation result.

They also introduced a third security model: preserving privacy with non-colluding parties where a malicious party might be able to corrupt the data but will not be able to learn the private data of other parties without cooperating with another party. So the security model of the authors in [29] assumes that all the parties are sincere and want to get right results out of the computation and will not compromise this goal by providing incorrect data.

Depending on the number of inputs, the computation of data can be classified into two models: the single input computation model and the multi-input computation model. The secure multi-party computation usually has at least two inputs. The transformation of the input to a secure multi-party computation can be divided into three transformations. The first one is the transformation of a multi-input computation model to a secure multi-party computation model as shown in Figure 1.

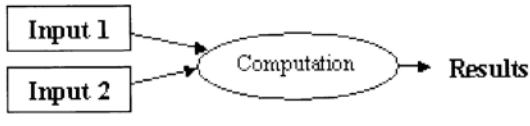

Fig. 1. Transformation of a Multi-Input Computation Model to a Secure Multi-Party Computation Model

The second is the transformation of a single-input computation model to a homogeneous secure multi-party computation model as shown in Figure 2.

The third one is the transformation of a single-input computation model to a heterogeneous secure multi-party computation as shown in Figure 3.

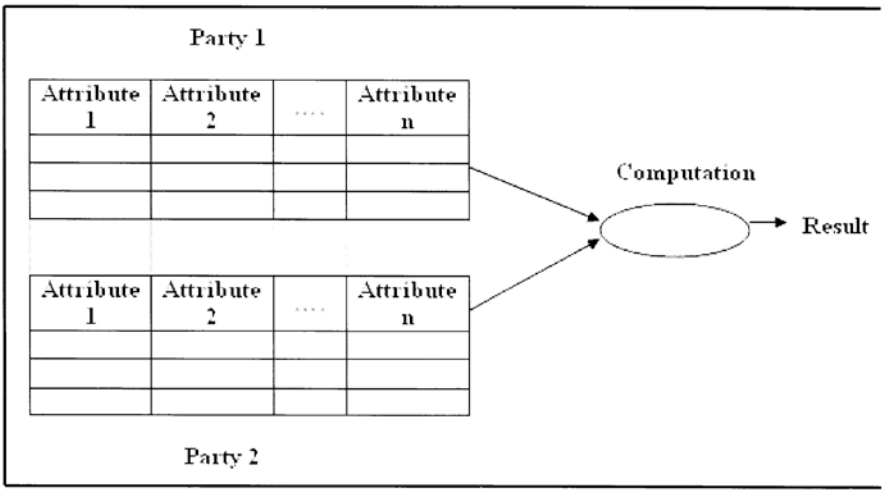

Fig. 2. Transformation of a Single-Input Computation Model to a Homogeneous Secure Multi-Party Computation Model

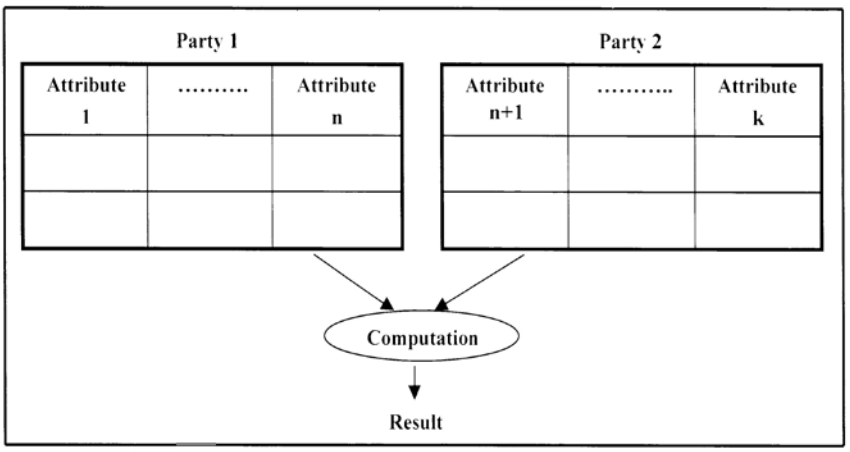

Fig. 3. Transformation of a Single-Input Computation Model to a Heterogeneous Secure Multi-Party Computation Model

Looking at these three transformations, the problem still how can we do the computation on the union of the data of different parties without each party disclosing its data to the others.

According to Du and Attallah [15], there are specific secure multi-party computation problems. we will present the problems related to privacy preserving data mining. In my state-of-the-art classification, we call this problem privacy preserving data mining in secure multi-party computation (PPDMSMC). The privacy preserving data mining problem in the context of SMC (PPDMSMC) was divided into the following problems:

- the Classification Problem: "Alice has a private structured database D_1, and Bob has another private structured database D_2; both of the structured database are comprised of attribute-value pairs. Each row of the database is a transaction and each column is an attribute taking on different values. One of the attributes in the database is designated as the class attribute. How could Alice and Bob build a decision tree based on the $D_1 \cup D_2$ without disclosing the content of the database to the other party?" we mentioned before that Lindel and Pinkas proposed a solution to this problem in [32].
- the Data Clustering Problem: "Alice has a private database D_1, and Bob has a private database D_2. They want to jointly perform data clustering on the union of D_1 and D_2". Vaidya et al. [46] and Estivill-Castro [18] proposed a solution to the Clustering problem. There work will be discussed under Level 2 and the PPDMSMC dimension later.
- the Association Rule Mining Problem: "Alice has a private database D_1, and Bob has a private database D_2. They want to jointly identify association rules in the union of D_1 and D_2". "For example, country A's intelligence agents have observed the activities $X = (x_1, \ldots, x_n)$ for a period of time, and Country B's intelligence agents have observed the activities $Y = (y_1, \ldots, y_m)$ for the same period of time. They want to collaboratively find out whether the activities in Y has any correlation with the activities in X. The results of collaboration could help both countries to understand the trend of the behaviors of the target, such as the behaviors of some suspected terrorism organization, the military movement of a dangerous country, etc. However, neither A or B is willing to disclose its observation to the other countries because they do not fully trust each other. It is possible that B might use A's intelligence information (or sell it to the target) to uncover A's agents, and thus causing damage to A's intelligence agents".
- the Data Generalization, Summarization, and Characterization Problem: "Alice has a private database D_1, and Bob has a private database D_2. They want to generalize, summarize or characterize the union of these two databases".

2.3 Privacy Preserving Distributed Data Mining

A Distributed Data Mining (DDM) model assumes that the data sources are distributed across multiple sites. The challenge here is: how can we mine the

data across the distributed sources securely or without any party disclose its data to the others? Most of the algorithms developed in this field do not take privacy in account because the focus is on efficiency. A simple approach to mine private data over multiple sources is to run existing data mining tools at each site independently and combine the results [5] [38]. However, this approach failed to give valid results for the following reasons:

- Values for a single entity may be split across sources. Data mining at individual sites will be unable to detect cross-site correlations.
- The same item may be duplicated at different sites, and will be over-weighted in the results.
- Data at a single site is likely to be from a homogeneous population. Important geographic or demographic distinctions between that population and others cannot be seen on a single site.

Recently, research has addressed classification using Bayesian Networks in vertically partitioned data [7], and situations where the distribution is itself interesting with respect to what is learned [50]. Shenoy et al. proposed an efficient algorithm for vertically mining association rules [44]. Finally, data mining algorithms that partition the data into subsets have been developed [40]. However, non of this work directly addressed privacy issues and concerns.

2.4 The Balance

How much non-sensitive knowledge must be hidden to block as many inference channels as possible? The problem of protection against inference has been addressed in the literature of statistical databases since 1979 [13, 14]. Figure 4 describes the inference problem and shows that publishing the non-sensitive data might not be enough to block inference channels.

Some researchers refer to the process of protection against inference as data sanitization [3]. Data sanitization is defined as the process of making sensitive information in non-production databases safe for wider visibility [17]. Others [36] advocate a solution based on collaborators mining independently their own data and then sharing some of the resulting patterns. This second alternative is called rule sanitization [36]. In this later case a set of association rules is processed to block inference of so called sensitive rules.

Keeping the protection against inference problem in mind, how much data must be exposed for a useful and beneficial mining process, and at the same time protecting the privacy of individuals or parties? If a party hides too much, we might end up with useless results. On the contrary, if a party exposed more than a specific limit, we might face the inference problem and as a result jeopardize the privacy of that party.

The following questions are examples from real life of why the balance is important:

- How could planning decisions be taken if census data were not collected?

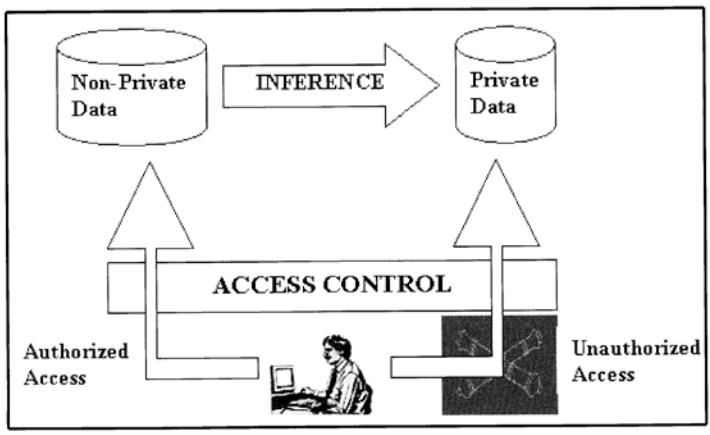

Fig. 4. The Inference Problem

– How could epidemics be understood if medical records were not analyzed?

Privacy advocates face considerable opposition, since data mining brings collective benefits in many contexts. Data mining has been also instrumental in detecting money laundering operations, telephone fraud, and tax evasion schemes. In such domains, it can be argued that privacy issues are secondary in the light of an important common good.

Data mining can be a powerful means of extracting useful information from data. As more and more digital data becomes available, the potential for misuse of data mining grows. The data mining field provides an interesting and varied set of challenges. PPDM is one of these challenges and can be achieved with a cost. There is a challenge on minimizing this cost while ensuring that the properties of private computation remains. A fundamental goal is to develop privacy and security models and protocols appropriate for data mining and to ensure that next generation data mining systems are designed from the ground up to employ these models and protocols.

3 Taxonomy of PPDM Techniques

Researchers usually classify PPDM problems based on the techniques used to protect sensitive data. When the classification is based on privacy preserving technique, we call this "classification by the what". We believe that classification by the what can be divided into two distinguished categories. First, hiding data or showing it exactly. Secure multi-party computation for example falls under this category. Other solutions that fall under this category are like: limiting access, augment the data, swapping, and auditing. Usually, the approaches under this category have less privacy but better accuracy in terms of results.

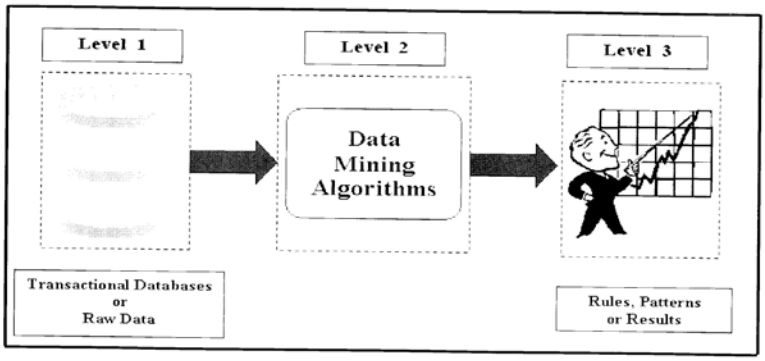

Fig. 5. Three Major Levels Where Privacy Preserving Data Mining Can Be Attempted

Second, perturbing the data which means changing attributes values with new values. This can be accomplished by adding noise, replacing selected values with a question mark (blocking). Approaches under this category have greater privacy but less accuracy in terms of results.

We are going to present a new categorization that is based on "classification by the where". We believe our classification is general, comprehensive and gives better understanding to the field of PPDM in terms of lying each problem under the right category. The new classification is as follows: PPDM can be attempted at three levels as shown in Figure 5. The first level is raw data or databases where transactions reside. The second level is data mining algorithms and techniques that ensure privacy. The third level is the output of different data mining algorithms and techniques.

At Level 1, researchers have applied different techniques to raw data or databases for the sake of protecting privacy of individuals (by preventing data miners from getting sensitive data or sensitive knowledge), or protecting privacy of two or more parties who want to perform some analysis on the combination of their data without disclosing their data to each other.

At Level 2, privacy preserving techniques are embedded in the data mining algorithms or techniques and may allow skillful users to enter specific constraints before or during the mining process.

Finally, at Level 3, researches have applied different techniques to the output of data mining algorithms or techniques for the same purpose of Level 1.

Most of the research in PPDM problems has been performed at Level 1. Few researchers applied privacy preserving techniques at Level 2 or Level 3. In our categorization, PPDM occurs in two dimensions under each level of the three levels mentioned above and these are:

- Individuals: This dimension involves implementing a PPDM technique to protect the privacy of an individual or more whose their data is going to be

published to the public. An example of this dimension is patients records or the census.

- PPDMSMC (Privacy Preserving Data Mining in Secure Multi-party Computation): This dimension involves protecting the privacy of two or more parties who want to perform a data mining task on the union of their sensitive data. An example of this dimension is two parties who want to cluster the union of their sensitive data without any party disclosing its data to the other.

Finally, sometimes PPDM techniques are not limited to one level; in other words, PPDM techniques can start at a level and extend to the next level or levels.

4 State-of-the-art in Privacy Preserving Data Mining - A New Look

In the literature of PPDM field, researchers presented different privacy preserving data mining problems based on the classifications of the authors. These classifications are good but we believe that from the targeted people (individuals and parties who want to protect their sensitive data) point of view, it is difficult to understand. We believe these people are interested in the answer of the following questions: Can this privacy preserving algorithm or technique protect our sensitive data at Level 1, Level 2 or at Level 3? The second important question is: Which dimension does this algorithm or technique fall under, is it the individuals or the PPDMSMC? In the following we review the work that has been done under each level to the best of our knowledge

4.1 Level 1 (Raw Data or Databases)

Individuals Dimension In 1996, Clifton et al. [9] presented a number of ideas to protect the privacy of individuals at Level 1. These include the following:

- Limiting access: we can control access to data so users can have access only to a sample of the data. We can lower the confidence of any mining that is attempted on the data. In other words, the control of access to data stops users from obtaining large chunks and varied samples of the database.
- Fuzz the data: altering the data by forcing, for example, aggregation to the daily transactions instead of individual transactions, prevents useful mining and at the same time allows the desired use of data. This approach is used by the U.S. Census Bureau.
- Eliminate unnecessary data: unnecessary data that could lead to private information. For example, the first three digit in social security number present the office that issued the number which can be used to reveal the location of the owner of that number. Another example, if an organization assigns phone numbers to employees based on their location; one can mine the company

phone book to find employees who for instance work on the same project. A solution to this problem is to give unique identifiers randomly to such records to avoid meaningful groupings based on these identifiers.

- Augment the data: which means adding value to the data without altering its usefulness. This added data is usually misleading and serve the purpose of securing the privacy of the owner of this data. For example, adding fictitious people to the phone book will not affect the retrieval of information about individuals but will affect queries that for instance try to retrieve all individuals who work in a certain company.
- Audit: to publish data mining results inside an organization rather than the world, we can use auditing to detect misuse so that administrative or criminal disciplinary action may be initiated.

Despite of the potential success of the previous solutions to restrict access to or perturb data. The challenge is to block the inference channels. Data blocking has been used for association rule confusion [6]. This approach of blocking data is implemented by replacing sensitive data with question mark instead of replacing data with false incorrect values. This is usually desirable for medical applications. An approach that applies blocking to the association rule confusion has been presented in [42].

In 2001, Saygin et al. [41] proposed an approach for hiding rules by replacing selected values or attributes with unknowns instead of replacing them with false values. In the following we discuss the approach:

Using Unknowns to Prevent Discovery of Association Rules

This technique depends on the assumption that in order to hide a rule $A \rightarrow B$ either the support of the itemset $A \cup B$ should be decreased below the minimum support threshold (MST) or the confidence of the rule should be decreased below the minimum confidence threshold (MCT). Based on the above, we might have the following cases for itemset A which contains a sensitive association rule:

- A remains sensitive when $\text{minsup}(A) \geq \text{MST}$
- A is not sensitive when $\text{maxsup}(A) < \text{MST}$
- A is sensitive with a degree of uncertainty when $\text{minsup}(A) \leq \text{MST} \leq \text{maxsup}(A)$

According to [41] the only way to decrease the support of a rule $A \rightarrow B$ is to replace 1s by ?s for the items in $A \cup B$ in the database. In this process, the minimum support value will be changed while the maximum support value will stay the same. Also, the confidence of a rule $A \rightarrow B$ can be decreased by replacing both 1s and 0s by ?s.

In the same year(2001), Dasseni et al. [11] proposed another approach that is based on perturbing support and/or confidence to hide association rules.

Hiding Association Rules by Using Confidence and Support

The work in [11] proposed a method to hide a rule by either decreasing its support or confidence. This is done by decreasing the support or the confidence one unit at a time by modifying the values of one transaction at a time. Since conf($A \rightarrow B$) = supp(AB) / supp(A), there are two strategies for decreasing the confidence of a rule:

- increasing the support of A in transactions not supporting B.
- Decreasing the support of B in transactions supporting both A and B.

Also to decrease the support for a rule $A \rightarrow B$, we can decrease the support for the itemset(AB). An example mentioned by the people who proposed the method in [11] clarifies the method as follows; let us suppose that $s = 20\%$ and $c = 80\%$.

Let us suppose that we have the database in the table shown in Figure 6. With the values for s and c above, we can deduce that we have two rules $AB \rightarrow C$ and $BC \rightarrow A$ with 100% confidence.

TID	Items
T1	ABC
T2	ABC
T3	A C
T4	A
T5	B

AR	Conf.
AB → C	100%
BC → A	100%

Fig. 6. $AB \rightarrow C$ and $BC \rightarrow A$ with 100% Confidence

Let us suppose that we want to hide the rule $AB \rightarrow C$ by increasing support of AB. Let us do that by turning to 1 the item B in transaction T_4 so the database becomes as shown in Figure 7.

Notice that the confidence for the rule $AB \rightarrow C$ was decreased to 66%. Having in mind that $c = 80\%$, we were successful in hiding the rule $AB \rightarrow C$. We can also hide the rule $AB \rightarrow C$ by decreasing the support of C by turning to 0 the item C in T_1 as Figure 8 shows.

Notice that the confidence for the rule was decreased to 50% which means that we were successful in hiding the rule $AB \rightarrow C$. Finally we can also hide the rule $AB \rightarrow C$ by decreasing the support of ABC by turning to 0 the item B in T_1 and turning to 0 the item C in T_2 as Figure 9 shows.

TID	Items
T1	ABC
T2	ABC
T3	A C
T4	AB
T5	B

AR	Conf.
AB → C	66%
BC → A	100%

Fig. 7. Hide the Rule $AB \rightarrow C$ by Increasing Support of AB

TID	Items
T1	AB
T2	ABC
T3	A C
T4	A
T5	B

AR	Conf.
AB → C	50%
BC → A	100%

Fig. 8. Hide the Rule $AB \rightarrow C$ by Decreasing the Support of C

TID	Items
T1	A C
T2	AB
T3	A C
T4	A
T5	B

AR	Conf.
AB → C	0%
BC → A	0%

Fig. 9. Hide the Rule $AB \rightarrow C$ by Decreasing the Support of ABC

Notice that the confidence for the rule $AB \rightarrow C$ was decreased to 0% this time, again we were successful in hiding the rule.

In 2006, HajYasien et al. [27] proposed a new heuristic algorithm called the QIBC algorithm that improves the privacy of sensitive knowledge (as itemsets) by blocking more inference channels. In another work [26], they presented two techniques based on item-restriction that hide sensitive itemsets.

PPDMSMC Dimension In the context of privacy preserving data mining, the following problem [15] was introduced by Du and Atallah in 2001. Alice and Bob have two sensitive structured databases $D1$ and $D2$ respectively. Both of the databases are comprised of attribute-value pairs. Each row in the database represents a transaction and each column represents an attribute with different domains. Each database includes a class attribute. How could Alice and Bob build a decision tree based on $D1 \cup D2$ without disclosing the content of their databases to each other? In the same context, the problem could be that Alice and Bob want to perform any data mining technique on the union of $D1$ and $D2$. In 2002, Vaidya et al. [45] proposed an algorithm for efficiently discovering frequent itemsets in vertically partitioned data between two parties without any party discloses its data to the others. In 2003, Kantarcioglu et al. [29] addressed the question: Can we apply a model (to mine the data) without revealing it? The paper presents a method to apply classification rules without revealing either the data or the rules.

A protocol for private controlled classification was proposed with three phases: the encryption phase, the prediction phase and the verification phase. The problem can be stated formally as follows:

Given an instance x from site D with v attributes, we need to classify x according to a rule set R provided by site G. It is assumed that each attribute of x has n bits, and x_i denotes the i^{th} attribute of x. it is also assumed that each given classification rule $r \in R$ is of the form $(L_1 \wedge L_2 \wedge ... \wedge L_v) \rightarrow C$ where C is the predicted class if $(L_1 \wedge L_2 \wedge ... \wedge L_v)$ evaluates to true. Each L_i is either $x_i = a$, or a don't care (always true). In addition, D has a set F of rules that are not allowed to be used for classification. In other words, D requires $F \cap R = \emptyset$. The goal is to find the class value of x according to R while satisfying the following conditions:

- D will not be able to learn any rules in R,
- D will be convinced that $F \cap R = \emptyset$ holds, and
- G will only learn the class value of x and what is implied by the class value.

In summary, the first two phases perform the correct classification without revealing the rules or the data. For the sites that are preforming the classification, phase three insures that the intersection between their rules and the classification rules is $\emptyset$.

4.2 Level 2 (Data Mining Algorithms and Techniques)

Individuals Dimension The work under the individuals dimension presented by Srikant et al., Ng et al., Lakshmanan, and Boulicaut in 1997, 1998, 1999, and 2000 respectively did not address privacy preserving data mining directly. They applied techniques to impose constraints during the mining process to limit the number of rules to what they call "interesting rules".

PPDMSMC Dimension Under the PPDMSMC dimension, there have been some cryptography based algorithms to solve the SMC problem. In 2000, Lindell and Pinkas have used the ID3 algorithm over horizontally partitioned data to introduce a secure multi-party technique for classification [32]. In 2002, Du et al. proposed a privacy preserving algorithm based on cryptographic protocol that implemented the ID3 algorithm over vertically partitioned data [16]. Lin et al. proposed a secure way for clustering using the EM algorithm [12] over horizontally partitioned data [31]. In 2003, Vaidya proposed a privacy preserving k-means algorithm [46] that requires three non-colluding sites. These sites could be among the sites holding the data or could be external sites. A permutation algorithm is presented that enhances the security of the calculations. Formally, the problem can be described with two parties A and B. B has an n-dimensional vector $X = (x_1, \ldots, x_n)$, and A has an n-dimensional vector $Y = (y_1, \ldots, y_n)$. A also has a permutation π of the n numbers. The aim is for B to receive the result $\pi(X + Y)$, without disclosing anything else. In other words, neither A nor B can learn the vector of the other and B does not learn π. The Y is used to hide the permutation of the other vector. It is a vector of random numbers from a uniform random distribution. The solution makes use of a tool known as Homomorphic Encryption. An encryption function $H : R \rightarrow S$ is called additively homomorphic if there is an efficient algorithm $Plus$ to compute $H(x+y)$ from $H(x)$ and $H(y)$ that does not reveal x or y. Examples that include such systems can be found in Benaloh [4], Naccache and Stern [33], Okamoto and Uchiyama [34], and Paillier [37]. This allows to perform addition of encrypted data without decrypting it. In addition to the secure multi-party computation discussion, the paper gives definitions and proofs and also calculations of the cost of the algorithms presented. In 2004, Estivill-Castro [18] proposed a solution to construct a representative-based clustering algorithms under the scenario that the dataset is partitioned into at least two sections owned by parties who do not trust each other. A protocol presented in the paper that allows parties to carry this task under the k-medoids algorithm. That was an improvement over the previous algorithm proposed by Vaidya because clustering with medoids (medians or other loss functions) is a more robust alternative than clustering with k-means (a method that is statistically biased and statistically inconsistent with very low robustness to noise).

4.3 Level 3 (Output of Data Mining Algorithms and Techniques)

Privacy preserving data mining at this level provides more security since no raw data or databases are shared here. The output of a data mining process is shared.

Individuals Dimension Under this dimension, parties share the knowledge after removing what is sensitive or share the rules as a set and no party knows which knowledge in particular belongs to which party. In doing so, parties avoid sharing databases or raw data which will reduce the hazards of inferring any sensitive knowledge. The challenge here is that the release of all the patterns that are not sensitive is not enough.

In 2004, Oliveira et al. presented a sanitization technique [36] to block what is called "Forward Inference Attack" and "Backward Inference Attack". We believe this is a good start and opens the door for future work at this level. However, the paper introduced only one form of data mining outputs (that is itemsets) and ignored other forms of outputs. The following is a discussion of the secure association rule sharing mentioned above.

Secure Association Rule Sharing

Sharing association rules is usually beneficial especially for the industry sector but requires privacy protection. A party might decide to release only part of the knowledge and hide strategic patterns which are called restrictive rules [36]. Restrictive rules must be hidden before sharing to protect the privacy of involved parties. The challenge here is that removing the restrictive rules from the set of rules is not enough. Removing more rules might be important to protect against the inference. Many researchers have addressed the problem of protection against the inference through sanitizing the knowledge be it data or rules. The existing sanitizing algorithms can be classified into two major classes: Data-Sharing approach and Pattern-Sharing approach. We mentioned before that the algorithms of data-sharing techniques are classified into three categories. Item Restriction-Based, Item Addition-Based, and Item obfuscation-Based. The previous categories sanitizes the transactions while the pattern sharing approach sanitizes a set of restrictive rules and blocking some inference channels. This is called secure association rule sharing.

The secure association rule sharing problem can be defined formally as follows [47]: "Let D be a database, R be the set of rules mined from D based on a minimum support threshold σ, and R_R be a set of restrictive rules that must be protected according to some security/privacy policies. The goal is to transform R into R', where R' represents the set of non-restrictive rules. In this case, R' becomes the released set of rules that is made available for sharing. Ideally, $R' = R - R_R$. However, there could be a set of rules r in R' from which one could derive or infer a restrictive rule in R_R. So in reality, $R' = R-(R_R+R_{SE})$, where R_{SE} is the set of non-restrictive rules that are removed as side effect of the sanitization process to avoid recovery of R_R". Figure 10 illustrates the problems that occur during the rule sanitization process.

Rules are usually derived from frequent itemsets. Frequent itemsets derived from a database can be represented in the form of a directed graph. A frequent

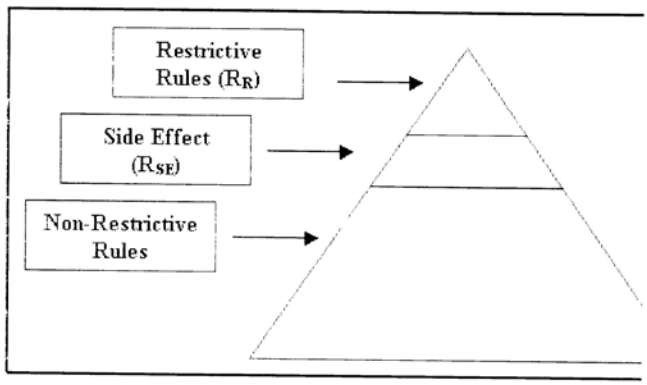

Fig. 10. The Inference Problem in Association Rules

itemset graph, denoted by $G = (C, E)$, is a directed graph which nonempty set of frequent itemsets C, a set of edges E that are orde of the elements of C, such that $\forall u, v \in C$ there is an edge from u to v and if $|v| - |u| = 1$ where $|x|$ is the size of itemset x.

The frequent itemset graph includes one or more levels which defined as follows: Let $G = (C, E)$ be a frequent itemset graph. The the length of the maximum path connecting an 1-itemset u to any v, such that $u, v \in C$ and $u \subset v$ [36].

In general, top-down traversal of G constrained by a minim threshold σ is used to discover the itemsets in G. It is an iterati which k-itemsets are used to explore $(k + 1)$-itemsets.

Based on the definitions above we can present the so called a sanitized rules [36]. If someone mines a sanitized set of rules and or more restrictive rules, that is called an attack against sanitize attacks against sanitized rules are identified in [36] as follows:

Forward Inference Attacks: Suppose we want to remove restrictive rules derived from the itemset ABC as shown in Figur not be enough to remove the itemset ABC because a miner can ABC is frequent when she find that AB, BC and AC are frequent. is referred to as forward inference attack. To avoid the inference frequent we need to remove one of its subsets in level 1 in Figure 1 of deeper graph, the removal is done recursively up to level 1.

Backward Inference Attack: Suppose we want to sanitize an from the itemset AC, it is straightforward to infer that AC is freq we still have the frequent itemsets ABC and ACD which from ei can be inferred. To block this attack, we must remove any superset AC. In this particular case, ABC and ACD must be removed as w

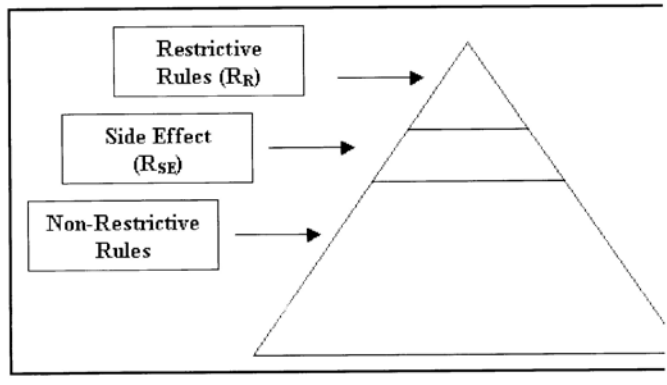

Fig. 10. The Inference Problem in Association Rules

itemset graph, denoted by $G = (C, E)$, is a directed graph which c nonempty set of frequent itemsets C, a set of edges E that are ordere of the elements of C, such that $\forall u, v \in C$ there is an edge from u to v i and if $|v| - |u| = 1$ where $|x|$ is the size of itemset $|x|$.

The frequent itemset graph includes one or more levels which ar defined as follows: Let $G = (C, E)$ be a frequent itemset graph. The l the length of the maximum path connecting an 1-itemset u to any otl v, such that $u, v \in C$ and $u \subset v$ [36].

In general, top-down traversal of G constrained by a minimun threshold σ is used to discover the itemsets in G. It is an iterative which k-itemsets are used to explore $(k + 1)$-itemsets.

Based on the definitions above we can present the so called atta sanitized rules [36]. If someone mines a sanitized set of rules and d or more restrictive rules, that is called an attack against sanitized attacks against sanitized rules are identified in [36] as follows:

Forward Inference Attacks: Suppose we want to remove (sa restrictive rules derived from the itemset ABC as shown in Figure not be enough to remove the itemset ABC because a miner can d ABC is frequent when she find that AB, BC and AC are frequent. S is referred to as forward inference attack. To avoid the inference th frequent we need to remove one of its subsets in level 1 in Figure 11. of deeper graph, the removal is done recursively up to level 1.

Backward Inference Attack: Suppose we want to sanitize any r from the itemset AC, it is straightforward to infer that AC is freque we still have the frequent itemsets ABC and ACD which from eith can be inferred. To block this attack, we must remove any superset th AC. In this particular case, ABC and ACD must be removed as well

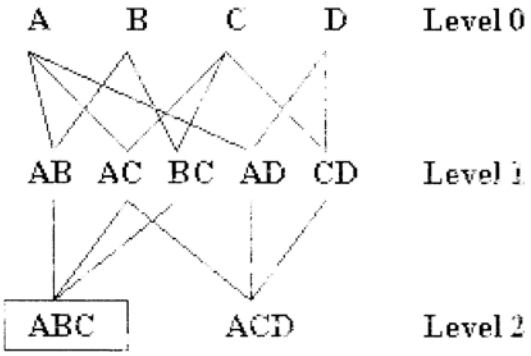

Fig. 11. An Example of Forward Inference

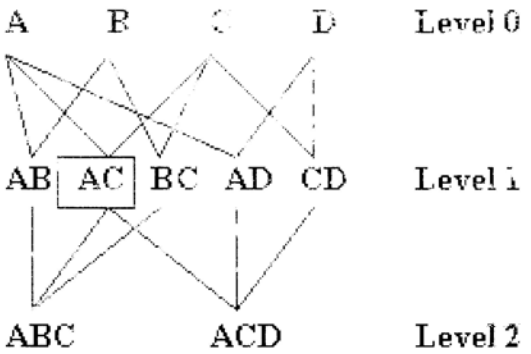

Fig. 12. An Example of Backward Inference

PPDMSMC Dimension In 2002, Kantarcloglu et al. [30] proposed a privacy preserving method for mining association rules on horizontally partitioned data. Their method is based on two phases. Phase 1, uses commutative encryption. An encryption is commutative if the following two equations hold for any encryption key $k1, k2, ...kn \in K$, any element $m \in M$ and permutations of i, j: $\forall m1, m2 \in M$ such that $m1 = m2$:

$$E_{ki1}(...E_{kin}(m)...) = E_{kj1}(...E_{kjn}(m)...) \quad (1)$$

and for any given k,

$$Pr(E_{ki1}(...E_{kin}(m)...) = E_{kj1}(...E_{kjn}(m)...)) \quad (2)$$

Each party encrypts its own items then the (already encrypted) itemsets of every other party. Then these will be passed around with every party decrypting to obtain the complete set. Figure 13 illustrates phase 1. In the second phase, an initiating party passes its support for each of the itemsets in the complete set, plus a random value, to its neighbor. The neighbor adds its support value for each itemset in the complete set and passes them on. The final party with the initiating party compute if the final results are greater than the threshold plus the random value through a secure comparison.

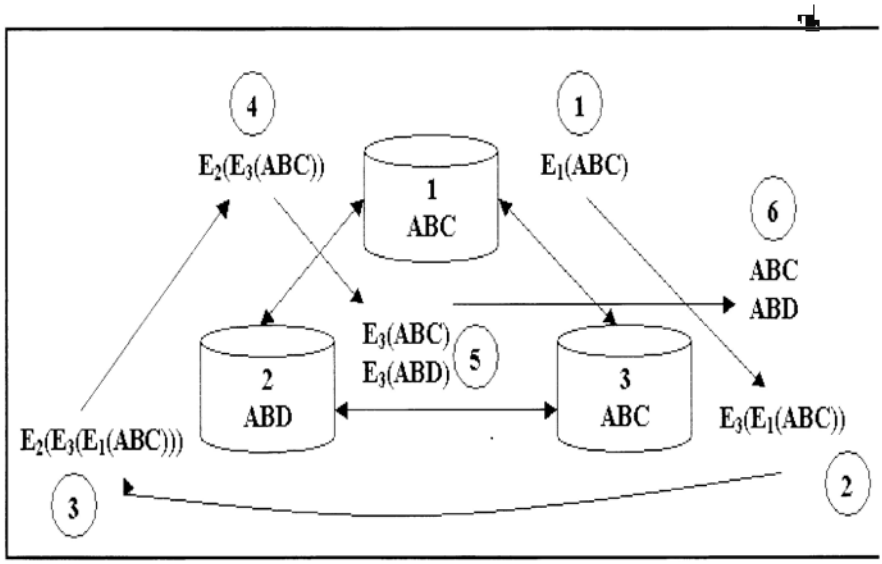

Fig. 13. Steps for Determining Global Candidate Itemsets

In 2007, Estivill-Castro et al. [19] proposed a flexible and easy-to-implement protocol for privacy-preserving data sharing based on a a public-key cryptosystem. They showed that their protocol is efficient, and especially more efficient than the protocol presented Kantarcloglu et al. [30].

5 Summary and Conclusion

Organizations and people like their data to be strictly protected against any unauthorized access. For them data privacy and security is a priority. At the same time, it might be necessary for them to share data for the sake of getting beneficial results. The problem is how can these individuals or parties compare their data or share it without revealing the actual data to each other. It is also always assumed that the parties who want to compare or share data results do not trust each other and/or compete with each other.

The concept of data mining has been around for almost 30 years but it took the innovative computing technology and software the last decade for it to develop into the powerful tool it is today. Data mining is a powerful tool but like all powerful things is subject to abuse, misuse and ethical considerations. To ensure the integrity of its use, and therefore the confidence of the users, research must adequately regulate itself concerning privacy issues. Failure to do so will increase the hesitation of individuals as well as organizations from releasing or exchanging data which will affect the performance of these organizations and limit their ability to take steps for the future. Not to mention that the release of sensitive data will invite security and intervention of the authorities, which will create its own set of problems.

We have presented a new classification for the privacy preserving data mining problems. The new classification is better because individuals or parties who want to protect their data usually look at the privacy preserving tools as a black box with input and output. Usually the focus is not whether the privacy preserving data mining technique is based on cryptography techniques or based on heuristic techniques. What is important is, we want to protect the privacy of our data at Level 1, Level 2 or at Level 3. Finally, we can conclude from the review that most of the research in privacy preserving data mining has been done under Level 1 and few research under Level 2 and Level 3.

References

1. Special section on privacy and security. *SIGKDD Explorations*, 4(2):i–48, January 2003.
2. Ackerman, M. L. Cranor, and J. Reagle. Privacy in e-commerce: Examining user scenarios and privacy preferences. In *Proceedings of the ACM Conference on Electronic Commerce*, pages 1–8, Denver, Colorado, USA, November 1999. ACM Press.
3. Atallah,M. E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios. Disclosure limitation of sensitive rules. In *Proceedings of 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, pages 45–52, Chicago, Illinois, USA, November 1999. IEEE Computer Society.

4. Benaloh.J. Dense probabilistic encryption. In *Proceedings of the Workshop on Selected Areas of Cryptography*, pages 120–128, Kingston, Ontario, Canada, May 1994.
5. Chan. P. An extensible meta-learning approach for scalable and accurate inductive learning. *PhD Thesis, Department of Computer Science, Columbia University, New York, NY, USA*, 1996.
6. Chang L. and I. S. Moskowitz. An integrated framework for database inference and privacy protection. In B. M. Thuraisingham, R. P. van de Riet, K. R. Dittrich, and Z. Tari, editors, *Proceedings of the 14th Annual Working Conference on Database Security*, pages 161–172, The Netherlands, August 2000. Kluwer Academic Publishers.
7. Chen, R. K. Sivakumar, and H. Kargupta. Distributed web mining using bayesian networks from multiple data streams. In N. Cercone, T. Young Lin, and X. Wu, editors, *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM'01)*, pages 75–82, San Jose, Clifornia, USA, November 2001. IEEE Computer Society.
8. Clifton, C. M. Kantarcioglu, and J. Vaidya. Defining privacy for data mining. In H. Kargupta, A. Joshi, and K. Sivakumar, editors, *Proceedings of the National Science Foundation Workshop on Next Generation Data Mining*, pages 126–133, Baltimore, MD, USA, November 2002.
9. Clifton C. and D. Marks. Security and privacy implications of data mining. In *Workshop on Data Mining and Knowledge Discovery*, pages 15–19, Montreal, Canada, February 1996. University of British Columbia, Department of Computer Science.
10. Cockcroft S. and P. Clutterbuck. Attitudes towards information privacy. In *Proceedings of the 12th Australasian Conference on Information Systems*, Australia, 2001.
11. Dasseni, E. V. S. Verykios, A. K. Elmagarmid, and E. Bertino. Hiding association rules by using confidence and support. In I. S. Moskowitz, editor, *Proceedings of the 4th Information Hiding Workshop*, volume 2137, pages 369–383, Pittsburg, PA, USA, April 2001. Springer Veralg Lecture Notes in Computer Science.
12. Dempster,A. P. N. M. Laird, and D. B. Rubin. Maximum liklihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B–39(1):1–38, July 1977.
13. Denning D. E. and P. J. Denning. Data security, volume 11, pages 227–249. ACM Press, 1979.
14. Denning.D. E. R. *Cryptography and Data Security (book)*. Addison-Wesley, 2nd edition, 1982.
15. Du W. and M. J. Atallah. Secure multi-party computation problems and their applications: A review and open problems. In V. Raskin, S. J. Greenwald, B. Timmerman, and D. M. Kienzle, editors, *Proceedings of the New Security Paradigms Workshop*, pages 13–22, Cloudcroft, New Mexico, USA, September 2001. ACM Press.
16. Du W. and Z. Zhan. Building decision tree classifier on private data. In C. Clifton and V. Estivill-Castro, editors, *Proceedings of the IEEE international conference on Privacy, security and data mining*, volume 14, pages 1–8, Maebashi City, Japan, December 2002. ACS.
17. Edgar.D. Data sanitization techniques. In *Database Knowledge Base*. White Papers, October 2004.

18. Estivill-Castro. V. Private representative-based clustering for vertically partitioned data. In R. Baeza-Yates, J. L. Marroquin, and E. Chavez, editors, *Proceedings of the 5th Mexican International Conference in Computer Science (ENC'04)*, volume 00, pages 160–167, Colima, Mxico, September 2004. IEEE Computer Society.
19. . Estivill-Castro V and A. HajYasien. Fast private association rule mining by a protocol securely sharing distributed data. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI 2007)*, New Brunswick, New Jersey, USA, May 2006. IEEE Computer Society Press.
20. Fayyad, U. M. G. Piatetsky-Shapiro, P. Smyth, and R. (eds.) Uthurusamy. In *Advances in Knowledge Discovery and Data Mining (Journal)*. AIII Press/MIT Press, March 1996.
21. Franklin,M. Z. Galil, and M. Yung. An overview of secure distributed computing. Technical Report TR CUCS-008-92, 1992.
22. Garfinkel. S. Database nation: The death of privacy in the 21st century. Sebastopol, CA, USA, 2001. O'Reilly and Associates.
23. Goldreich.O. Secure multi-party computation. Working Draft, Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel, June 1998.
24. Goldreich,O. S. Micali, and A. Wigderson. How to play any mental game. In *Proceedings of the 19th annual ACM Symposium on Theory of Computing*, pages 218–229, New York, New York, USA, 1987. ACM Press.
25. Goldwasser.S. Multi-party computations: Past and present. In *Proceedings of the 16th Annual ACM Symposium on the Principles of Distributed Computing*, pages 1–6, Santa Barbara, California, USA, 1997. ACM Press.
26. HajYasien A. and V. Estivill-Castro. Two new techniques for hiding sensitive itemsets and their empirical evaluation. In S. Bressan, J. Kng, and R. Wagner, editors, *Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2006)*, volume 4081, pages 302–311, Krakow, Poland, Septemper 2006. Springer Veralg Lecture Notes in Computer Science.
27. HajYasien,A. V. Estivill-Castro, and R. Topor. Sanitization of databases for refined privacy trade-offs. In A. M. Tjoa and J. Trujillo, editors, *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI 2006)*, volume 3975, pages 522–528, San Diego, USA, May 2006. Springer Veralg Lecture Notes in Computer Science.
28. Jefferies.P. Multimedia, cyberspace and ethics. In *Proceedings of the International Conference on Information Visualisation (IV'00)*, pages 99–104, London, England, July 2000. IEEE Computer Society.
29. . Kantarcioglu M and C. Clifton. Assuring privacy when big brother is watching. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 88–93, San Diego, California, USA, 2003. ACM Press.
30. . Kantarcioglu M and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In *IEEE Transactions on Knowledge and Data Engineering Journal*, volume 16(9), pages 1026–1037, Piscataway, NJ, USA, September 2004. IEEE Educational Activities Department.
31. Lin X. and C. Clifton. Privacy-preserving clustering with distributed EM mixture modeling. In *Knowledge and Information Systems Journal*, volume 8(1), pages 68–81, New York, NY, USA, July 2005. Springer Verlag.
32. . Lindell Y and B. Pinkas. Privacy preserving data mining. In *CRYPTO-00*, volume 1880, pages 36–54, Santa Barbara, California, USA, 2000. Springer Verlag Lecture Notes in Computer Science.

33. Naccache D. and J. Stern. A new public key cryptosystem based on higher residues. In *Proceedings of the 5th ACM conference on Computer and Communications Security*, pages 59–66, San Francisco, California, USA, 1998. ACM Press.
34. Okamoto T. and S. Uchiyama. A new public-key cryptosystem as secure as factoring. In *Advances in Cryptology (Eurocrypt'98)*, volume 1403, pages 308–318, Helsinki, Finland, June 1998. Springer Veralg Lecture Notes in Computer Science.
35. Oliveira S. R. M. and O. R. Zaiane. Toward standardization in privacy-preserving data mining. In *Proceedings of the 3rd Workshop on Data Mining Standards (DM-SSP 2004), in conjuction with KDD 2004*, pages 7–17, Seattle, WA, USA, August 2004.
36. Oliveira, S. R. M. O. R. Zaiane, and Y. Saygin. Secure association rule sharing. In H. Dai, R. Srikant, and C. Zhang, editors, *Proceedings of the 8th PAKDD Conference*, volume 3056, pages 74–85, Sydney, Australia, May 2004. Springer Veralg Lecture Notes in Computer Science.
37. Paillier.P. Public key cryptosystems based on composite degree residuosity classes. In *Advances in Cryptology (Eurocrypt'99)*, volume 1592, pages 223–238, Prague, Czech Republic, May 1999. Springer Veralg Lecture Notes in Computer Science.
38.. Prodromidis A and P. Chan. Meta-learning in distributed data mining systems: Issues and approaches, chapter 3. AAAI Press, 2000.
39. Rezgui,A. A. Bouguettaya, and M. Y. Eltoweissy. Privacy on the web: Facts, challenges, and solutions. *IEEE Security and Privacy*, 1(6):40–49, November 2003.
40.. Savasere, AE. Omiecinski, and S. B. Navathe. An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB'95)*, pages 432–444, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
41. Saygin, Y. V. S. Verykios, and C. Clifton. Using unknowns to prevent discovery of association rules. In *ACM SIGMOD Record*, volume 30(4), pages 45–54, New York, NY, USA, December 2001. ACM Press.
42. Saygin,Y. V. S. Verykios, and A. K. Elmagarmid. Privacy preserving association rule mining. In E. Lim M. Shan Z. Yanchun, A. Umar, editor, *Proceedings of the 12th International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems (RIDE'02)*, pages 151–158, San Jose, California, USA, February 2002. IEEE Computer Society.
43.. Schoeman. F. D Philosophical dimensions of privacy. Cambridge Univ. Press, 1984.
44. Shenoy, P. J. R. Haritsa, S. Sundarshan, G. Bhalotia, M. Bawa, and D. Shah. Turbo-charging vertical mining of large databases. In *ACM SIGMOD Record*, volume 29(2), pages 22–33, Dallas, Texas, USA, June 2000. ACM Press.
45. Vaidya J. and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *Proceedings the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 639–644, Edmonton, Alberta, Canada, July 2002. ACM Press.
46. Vaidya J. and C. Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 206–215, Washington, D.C., USA, August 2003. ACM Press.
47. Verykios V. S., A. K. Elmagarmid, B. Elisa, Y. Saygin, and D. Elena. Association rule hiding. In *IEEE Transactions on Knowledge and Data Engineering*, volume 16(4), pages 434–447, Los Alamitos, CA, USA, April 2004. IEEE Computer Society.
48. Warren S. D. and L. D. Brandeis. The right to privacy. *Harvard Law Review*, 4(5):193–220, 1890.