Privacy-Preserving Data Mining on the Web: Foundations and Techniques

Stanley R. M. Oliveira[1,2]
[1]Embrapa Informática Agropecuária
André Tosello, 209 - Barão Geraldo
13083-886, Campinas, SP, Brasil

Osmar R. Zaïane[2]
[2]Department of Computing Science
University of Alberta
Edmonton, AB, Canada, T6G 1K7

Abstract

Privacy-preserving data mining (PPDM) on the Web is one of the newest trends in privacy and security research. It is driven by one of the major policy issues of the information era - *the right to privacy*. Although this research field is very new, we have already seen great interests in it: a) the recent proliferation in PPDM techniques is evident; b) the interest from academia and industry has grown quickly; and c) separate workshops and conferences devoted to this topic have emerged in the last few years.

This chapter describes the foundations for further research in PPDM on the Web. In particular, we discuss the problems in defining privacy and how privacy can be violated in data mining. Then, we describe the basis of PPDM including the historical roots, the definition of privacy preservation in data mining, and the general parameters that characterizes scenarios in PPDM. Subsequently, we analyze the implications of the Organization for Economic Cooperation and Development (OECD) data privacy principles in knowledge discovery. As a consequence, we suggest some policies for PPDM based on the OECD privacy guidelines. We also introduce a taxonomy of the existing PPDM techniques and a discussion on how these techniques are applicable on Web data. Finally, we suggest some privacy requirements that are related to industrial initiatives, and point to some technical challenges as future research trends in PPDM on the Web.

1 Introduction

In this section, we focus primarily on two important issues: the problem in defining which information is private in data mining and how privacy can be violated in data mining.

1.1 Problems in Defining Privacy

Analyzing what right to privacy means is a fraut with problems, such as the exact definition of privacy, whether it constitutes a fundamental right, and whether people are and/or should be concerned with it. Several definitions of privacy have been given, and they vary according to context, culture, and environment. For instance, in an 1890 paper [62], Warren & Brandeis defined privacy as "the right to be alone." Later, in a paper published in 1967 [63], Westin defined privacy 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". In [54], Schoeman 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." More recently, Garfinkel [20] stated that "privacy is about self-possession, autonomy, and integrity." On the other hand, Rosenberg argues that privacy may not be a right after all but a taste [52]: "If privacy is in the end a matter of individual taste, then seeking a moral foundation for it – beyond its role in making social institutions possible that we happen to prize – will be no more fruitful than seeking a moral foundation for the taste for truffles."

The above definitions suggest that, in general, privacy is viewed as a social and cultural concept. However, with the ubiquity of computers and the emergence of the Web, privacy has also become a digital problem [50]. With the Web revolution and the emergence of data mining, privacy concerns have posed technical challenges fundamentally different from those that occurred before the information era. 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 [9, 1, 24].

Clearly, the concept of privacy is often more complex than realized. In particular, in data mining, the definition of privacy preservation is still unclear, and there is very little literature related to this topic. A notable exception is the work presented in [6], in which PPDM is defined as "getting valid data mining results without learning the underlying data values." However, at this point, each existing PPDM technique has its own privacy definition. Our

primary concern about PPDM is that mining algorithms are analyzed for the side effects they incur in data privacy. Therefore, our definition for PPDM is close to those definitions in [54, 6] – *PPDM encompasses the dual goal of meeting privacy requirements and providing valid data mining results*. Our definition emphasizes the dilemma of balancing privacy preservation and knowledge disclosure.

1.2 Privacy Violation in Data Mining

Understanding privacy in data mining requires understanding how privacy can be violated and the possible means for preventing privacy violation. In general, one major factor contributes to privacy violation in data mining: *the misuse of data*.

Users' privacy can be violated in different ways and with different intentions. Although data mining can be extremely valuable in many applications (e.g. business, medical analysis, etc), it can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be violated if personal data are used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected.

One of the sources of privacy violation is called data magnets [50]. Data magnets are techniques and tools used to collect personal data. Examples of data magnets include explicitly collecting information through on-line registration, identifying users through IP addresses, software downloads that require registration, and indirectly collecting information for secondary usage. In many cases, users may or may not be aware that information is being collected or do not know how that information is collected [11, 35]. Worse is the privacy invasion occasioned by secondary usage of data when individuals are unaware of "behind the scenes" uses of data mining [25]. In particular, collected personal data can be used for secondary usage largely beyond the users' control and privacy laws. This scenario has led to an uncontrollable privacy violation not because of data mining itself, but fundamentally because of the misuse of data.

2 The Basis of Privacy-Preserving Data Mining

In this section, we describe the basis of PPDM including the historical roots, the definition of privacy preservation in data mining, and models of data miners in PPDM.

2.1 Historical Roots

The debate on PPDM has received special attention as data mining has been widely adopted by public and private organizations. We have witnessed three major landmarks that characterize the progress and success of this new research area: *the conceptive landmark*, *the deployment landmark*, and *the prospective landmark*. We describe these landmarks as follows:

The Conceptive landmark characterizes the period in which central figures in the community, such as O'Leary [39, 40], Fayyad, Piatetsky-Shapiro and Smith [18, 47], and others [33, 8], investigated the success of knowledge discovery and some of the important areas where it can conflict with privacy concerns. The key finding was that knowledge discovery can open new threats to informational privacy and information security if not done or used properly. They highlighted the success of knowledge discovery and some of important areas where it could conflict with privacy concerns. Since then, the debate on PPDM has gained momentum.

The Deployment landmark is the current period in which an increasing number of PPDM techniques have been developed and have been published in refereed conferences. The information available today is spread over countless papers and conference proceedings[1]. The results achieved in the last years are promising and suggest that PPDM will achieve the goals that have been set for it.

The Prospective landmark is a new period in which directed efforts toward standardization occur. At this stage, there is no consent about what privacy preservation means in data mining. In addition, there is no consensus on privacy principles, policies, and requirements as a foundation for the development and deployment of new PPDM techniques. The excessive number of techniques is leading to confusion among developers, practitioners, and others interested in this

[1]The Privacy Preserving Data Mining Site: http://www.cs.ualberta.ca/∼oliveira/psdm/psdm_index.html

technology. One of the most important challenges in PPDM now is to establish the groundwork for further research and development in this area.

2.2 Defining Privacy for Data Mining

In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. We refer to the former as *individual privacy preservation* and the latter as *collective privacy preservation*, which is related to corporate privacy in [6].

Individual privacy preservation: The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual.

Collective privacy preservation: Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to preserve strategic patterns that are paramount for strategic decisions, rather than minimizing the distortion of all statistics (e.g. bias and precision). In other words, the goal here is not only to protect personally identifiable information but also some patterns and trends that are not supposed to be discovered.

In the case of collective privacy preservation, organizations have to cope with some interesting conflicts. For instance, when personal information undergoes analysis processes that produce new facts about users' shopping patterns, hobbies, or preferences, these facts could be used in recommender systems to predict or affect their future shopping patterns. In general, this scenario is beneficial to both users and organizations. However, when organizations share data in a collaborative project, the goal is not only to protect personally identifiable information but also some strategic patterns. In the business world, such patterns are described as the knowledge that can provide competitive advantages, and therefore must be protected [57]. More challenging is to protect the knowledge discovered from confidential information (e.g. medical, financial, and crime information). The absence of privacy safeguards can equally compromise individuals' privacy. While violation of individual privacy is clear, violation of collective privacy can lead to violation of individual's privacy.

2.3 Characterizing Scenarios of Privacy Preservation on the Web

Before we describe the general parameters for characterizing scenarios in PPDM, let us consider two real-life motivating examples where PPDM poses different constraints:

Scenario 1: Suppose we have a server and many clients in which each client has a set of sold items (e.g. books, movies, etc). The clients want the server to gather statistical information about associations among items in order to provide recommendations to the clients. However, the clients do not want the server to know some strategic patterns (also called restrictive association rules). In this context, the clients represent companies and the server is a recommendation system for an e-commerce application, for example, fruit of the clients collaboration. In the absence of rating, which is used in collaborative filtering for automatic recommendation building, association rules can be effectively used to build models for on-line recommendation. When a client sends its frequent itemsets or association rules to the server, it must protect the restrictive itemsets according to some specific policies. The server then gathers statistical information from the non-restrictive

itemsets and recovers from them the actual associations. How can these companies benefit from such collaboration by sharing association rules while preserving some restrictive rules?

Scenario 2: Two organizations, an Internet marketing company and an on-line retail company, have datasets with different attributes for a common set of individuals. These organizations decide to share their data for clustering to find the optimal customer targets so as to maximize return on investments. How can these organizations learn about their clusters using each other's data without learning anything about the attribute values of each other?

Note that the above scenarios describe different privacy preservation problems. Each scenario poses a set of challenges. For instance, scenario 1 is a typical example of collective privacy preservation, while scenario 2 refers to individual's privacy preservation.

How can we characterize scenarios in PPDM? One alternative is to describe them in terms of general parameters. In [7], some parameters are suggested as follows:

Outcome: Refers to the desired data mining results. For instance, someone may look for association rules identifying relationships among attributes, or relationships among customers' buying behaviors as in scenario 1, or may even want to cluster data as in scenario 2.

Data Distribution: How are the data available for mining: are they centralized or distributed across many sites? In the case of data distributed throughout many sites, are the entities described with the same schema in all sites (horizontal partitions), or do different sites contain different attributes for one entity (vertical partitions)?

Privacy Preservation: What are the privacy preservation requirements? If the concern is solely that values associated with an individual entity not be released (e.g. personal information), techniques must focus on protecting such information. In other cases, the notion of what constitutes "sensitive knowledge" may not be known in advance. This would lead to human evaluation of the intermediate results before making the data available for mining.

3 Principles and Policies for PPDM

World-wide, privacy legislation, policies, guidelines, and codes of conduct have been derived from the set of principles established in 1980 by the OECD[2]. They represent the primary components for the protection of privacy and personal data, comprising a commonly understood reference point. A number of countries have adopted these principles as statutory law, in whole or in part. In this section, we analyze the OECD principles in the context of PPDM. We then suggest some policies for PPDM based on the OECD principles.

3.1 The implications of the OECD Privacy Guidelines in PPDM

We now analyze the implications of the OECD principles in PPDM. Then we suggest which principles should be considered absolute principles in PPDM.

1. **Collection Limitation Principle:** This principle states that some very sensitive data should not be held at all. Collection limitation is too general in the data mining context incurring in two grave consequences: a) the notion of "very sensitive" is sometimes unclear and may differ from country to country, leading to vague definitions; b) limiting the collection of data may make the data useless for knowledge discovery. Thus, this principle seems to be unenforceable in PPDM.

2. **Data Quality Principle:** This principle is related to the pre-processing stage in data mining in which data cleaning routines are applied to resolve inaccuracy and inconsistencies. Somehow, this principle is relevant in the pre-processing stage of knowledge discovery. However, most PPDM techniques assume that the data are already in an appropriate form to mine.

3. **Purpose Specification Principle:** This principle is the fundamental basis of privacy. Individuals should be informed of the purposes for which the information collected about

[2]Privacy Online – OECD Guidance on Policy and Practice. http://www.oecd.org/dataoecd/33/43/2096272.pdf

them will be used, and the information must be used solely for that purpose. In other words, restraint should be exercised when personal data are collected. This principle is extremely relevant in PPDM.

4. **Use Limitation Principle:** This principle is closely related to the purpose specification principle. Use limitation is perhaps the most difficult principle to address in PPDM. This principle states that the purpose specified to the data subject (consumer) at the time of the collection restricts the use of the information collected, unless the data subject has provided consent for additional uses. This principle is also fundamental in PPDM.

5. **Security Safeguards Principle:** This principle is basically irrelevant in the case of data privacy, but relevant for database security. Security safeguards principle is typically concerned with keeping sensitive information (e.g personal data) out of the hands of unauthorized users, which ensures that the data is not modified by users who do not have permission to do so. This principle is unenforceable in the context of PPDM.

6. **Openness Principle:** This principle, also called transparency, states that people have the right to know what data about them have been collected, who has access to the data, and how the data are being used. In other words, people must be aware of the conditions under which their information is being kept and used. However, data mining is not an open and transparent activity requiring analysts to inform individuals about particular derived knowledge, which may inhibit the use of data. This principle is equally important in PPDM.

7. **Individual Participation Principle:** This principle suggests that data subjects should be able to challenge the existence of information gained through data mining applications. Since knowledge discovery is not openly apparent to data subjects, the data subjects are not aware of knowledge discoveries related to them. While debatably collected individual information could belong to individuals, one can argue that collective information mined from databases belongs to organizations that hold such databases. In this case, the impli-

cations of this principle for PPDM should be carefully weighed; otherwise, it could be too rigid in PPDM applications.

8. **Accountability Principle:** This principle states that data controllers should inform data subjects of the use and findings from knowledge discovery. In addition, data controllers should inform individuals about the policies regarding knowledge discovery activities, including the consequences of inappropriate use. Some countries (e.g. the UK, Japan, Canada) that have adopted the OECD privacy principles do not consider this principle since it is not limited in scope, area, or application. Thus, the accountability principle is too general for PPDM.

Our analysis above suggests that the OECD privacy principles can be categorized into three groups according to their influence on the context of PPDM: *Group 1* is composed of those principles that should be considered as absolute principles in PPDM, such as Purpose Specification, Use Limitation, and Openness. *Group 2* consists of some principles that somehow impact PPDM applications, and their full implications should be understood and carefully weighed depending on the context. The principles that fall into this category are Data Quality and Individual Participation. *Group 3* encompasses some principles that are too general or unenforceable in PPDM. This group includes Collection Limitation, Security Safeguards, and Accountability. Clearly, the principles categorized in groups 1 and 2 are relevant in the context of PPDM and are fundamental for further research, development, and deployment of PPDM techniques.

3.2 Adopting PPDM Policies from the OECD Privacy Guidelines

A privacy policy is a statement made by a data controller that one's private information will only be used for certain stated purposes. In general, individuals are expected to review the policies over time since computer-based services, business and competitive environment change constantly.

One fundamental point to be considered when designing some privacy policies is that too

many restrictions could seriously hinder the normal functioning of business and governmental organizations. The worst thing is that restrictions, if not carefully weighed, could make PPDM results useless.

Given these facts, we suggest some policies for PPDM based on the OECD privacy principles. We try to find a good compromise between privacy requirements and knowledge discovery. We describe the policies as follows:

1. **Awareness Policy:** When a data controller collects personally identifiable information, the data controller shall express why the data are collected and whether such data will be used for knowledge discovery.

2. **Limit Retention Policy:** A data controller shall take all reasonable steps to keep only personal information collected that is accurate, complete, and up to date. In the case of personal information that is no longer useful, it shall be removed and not subjected to analysis to avoid unnecessary risks, such as wrong decision making which may incur liability.

3. **Forthcoming Policy:** Policies regarding collecting, processing, and analyzing that produce new knowledge about individuals shall be communicated to those about whom the knowledge discovered pertains, in particular when the discovered knowledge is be disclosed or shared.

4. **Disclosure Policy :** Data controllers shall only disclose discovered knowledge about an individual for purposes for which the individual consents and the knowledge discovered about individuals shall never be disclosed inadvertently or without consent.

4 A Taxonomy of Existing PPDM techniques

In this section, we highlight the main idea behind the existing PPDM techniques in the literature. We classify these techniques into three major categories: data partitioning, data modification, and data restriction, as can be seen in Figure 1.

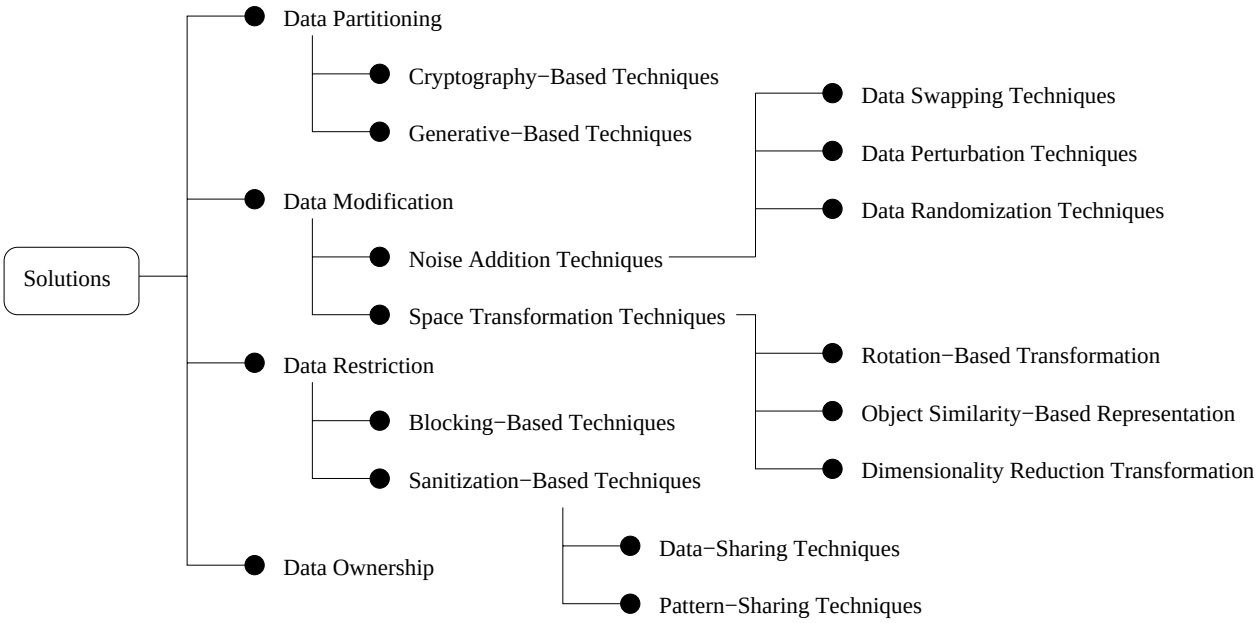

Figure 1: A taxonomy of PPDM techniques

4.1 Data Partitioning Techniques

Data partitioning techniques have been applied to some scenarios in which the databases available for mining are distributed across a number of sites, with each site only willing to share data mining results, not the source data. In these cases, the data are distributed either horizontally or vertically [7]. In a horizontal partition, different entities are described with the same schema in all partitions, while in a vertical partition the attributes of the same entities are split across the partitions. The existing solutions can be classified into *Cryptography-Based Techniques* and *Generative-Based Techniques*.

4.1.1 Cryptography-Based Techniques

In the context of PPDM over distributed data, cryptography-based techniques have been developed to solve problem of the following nature: two or more parties want to conduct a computation based on their private inputs. The issue here is how to conduct such a computation so that no party knows anything except its own input and the results. This problem is referred to as the Secure Multi-party Computation (SMC) problem [21, 13, 48]. The techniques

proposed in [36, 31] address privacy-preserving classification, while the techniques proposed in [30, 58] address privacy-preserving association rule mining, and the technique in [59] addresses privacy-preserving clustering.

4.1.2 Generative-Based Techniques

Generative-based techniques are designed to perform distributed mining tasks. In this approach, each party shares just a small portion of its local model which is used to construct the global model. The existing solutions are built over horizontally partitioned data. The solution presented in [60] addresses privacy-preserving frequent itemsets in distributed databases, whereas the solution in [37] addresses privacy-preserving distributed clustering using generative models.

4.2 Data Modification Techniques

These techniques modify the original values of a database that needs to be shared, and in doing so, privacy preservation is ensured. The transformed database is made available for mining and must meet privacy requirements without losing the benefit of mining. In general, data modification techniques aim at finding an appropriate balance between privacy preservation and knowledge disclosure. Methods for data modification include *noise addition techniques* and *space transformation techniques*.

4.2.1 Noise Addition Techniques

In statistical databases, noise addition techniques are used to protect individuals' privacy but at the expense of allowing partial disclosure, providing information with less statistical quality, and introducing biases into query responses [56]. In data mining, the major requirement of a security control mechanism (in addition to protect privacy) is not to ensure precise and bias-free statistics but rather to preserve the high-level descriptions of knowledge discovered from large databases [5, 14]. Thus, the idea behind noise addition techniques for PPDM is that some noise (e.g. information not present in a particular tuple or transaction) is added to the original data

to prevent the identification of confidential information relating to a particular individual. In other cases, noise is added to confidential attributes by randomly shuffling the attribute values to prevent the discovery of some patterns that are not supposed to be discovered.

We categorize noise addition techniques into three groups: (1) data swapping techniques that interchange the values of individual records in a database [14]; (2) data distortion techniques that perturb the data to preserve privacy, and the distorted data maintain the general distribution of the original data [4, 2, 64]. A different line of work is investigated in [32], in which the authors question the utility of the random value distortion techniques in privacy preservation; and (3) data randomization techniques which allow one to perform the discovery of general patterns in a database with error bound, while protecting individual values. Like data swapping and data distortion techniques, randomization techniques are designed to find a good compromise between privacy protection and knowledge discovery [17, 16, 51, 66, 65].

4.2.2 Space Transformation Techniques

Space transformation techniques are specifically designed to address privacy-preserving clustering. These techniques are designed to protect the underlying data values subjected to clustering without jeopardizing the similarity between objects under analysis. Thus, a space transformation technique must not only meet privacy requirements but also guarantee valid clustering results.

We categorize space transformation techniques into three major groups: (1) *Rotation-Based Transformation* makes the original attribute values difficult to perceive or understand and preserves all the information for clustering analysis [44]. The idea behind this technique is that the attributes of a database are split into pairwise attributes selected randomly. One attribute can be selected and rotated more than once, and the angle θ between an attribute pair is also selected randomly. This data transformation can be seen as a technique on the border with obfuscation. Obfuscation techniques aim at making information highly illegible without actually changing its inner meaning [10]; (2) *object similarity-based representation* relies on the idea

behind the similarity between objects, i.e., a data owner could share some data for clustering analysis by simple computing the dissimilarity matrix (matrix of distances) between the objects and then sharing such a matrix with a third party. Many clustering algorithms in the literature operate on a dissimilarity matrix [22]. This solution is simple to be implemented but requires a high communication cost since its complexity is of the order $O(m^2)$, where m is the number of objects under analysis [45]; (3) *dimensionality reduction-based transformation* can be used to address privacy-preserving clustering when the attributes of objects are available either in a central repository or split across many sites. By reducing the dimensionality of a dataset to a sufficiently small value, one can find a trade-off between privacy, communication cost, and accuracy. Once the dimensionality of a database is reduced, the released database preserves (or slightly modifies) the distances between data points. In tandem with the benefit of preserving the similarity between data points, this solution protects individuals' privacy since the attribute values of the objects in the transformed data are completely different from those in the original data [45].

4.3 Data Restriction Techniques

Data restriction techniques focus on limiting the access to mining results through either generalization or suppression of information (e.g., items in transactions, attributes in relations), or even by blocking the access to some patterns that are not supposed to be discovered. Such techniques can be divided into two groups: *Blocking-based techniques* and *Sanitization-based techniques*.

4.3.1 Blocking-Based Techniques

Blocking-based techniques aim at hiding some sensitive information when data are shared for mining. The private information includes restrictive association rules and classification rules that must remain private. Before releasing the data for mining, data owners must consider how much information can be inferred or calculated from large databases, and must look for ways to minimize the leakage of such information. In general, blocking-based techniques are

feasible to recover patterns less frequent than originally since sensitive information is either suppressed or replaced with unknowns to preserve privacy. The techniques in [26, 27] address privacy preservation in classification, while the techniques in [28, 53] address privacy-preserving association rule mining.

4.3.2 Sanitization-Based Techniques

Unlike blocking-based techniques that hide sensitive information by limiting or replacing some items or attribute values with unknowns, sanitization-based techniques hide sensitive information by strategically suppressing some items in transactional databases, or even by generalizing information to preserve privacy in classification. These techniques can be categorized into two major groups: (1) data-sharing techniques in which the sanitization process acts on the data to remove or hide the group of restrictive association rules that contain sensitive knowledge. To do so, a small number of transactions that contain the restrictive rules have to be modified by deleting one or more items from them or even adding some noise, i.e., new items not originally present in such transactions [61, 12, 41, 42, 43]; and (2) pattern-sharing techniques in which the sanitizing algorithm acts on the rules mined from a database, instead of the data itself. The existing solution removes all restrictive rules before the sharing process and blocks some inference channels [46]. In the context of predictive modeling, a framework was proposed in [23] for preserving the anonymity of individuals or entities when data are shared or made publicly.

4.4 Data Ownership Techniques

Data ownership techniques can be applied to two different scenarios: (1) to protect the ownership of data by people about whom the data were collected [19]. The idea behind this approach is that a data owner may prevent the data from being used for some purposes and allow them to be used for other purposes. To accomplish that, this solution is based on encoding permissions on the use of data as theorems about programs that process and mine the data. Theorem proving techniques are then used to guarantee that these programs comply with the permissions; and

(2) to identify the entity that receives confidential data when such data are shared or exchanged [38]. When sharing or exchanging confidential data, this approach ensures that no one can read confidential data except the receiver(s). It can be used in different scenarios, such as statistical or research purposes, data mining, and on-line business-to-business (B2B) interactions.

4.5 Are These Techniques Applicable to Web Data?

After describing the existing PPDM techniques, we now move on to analyze which of these techniques are applicable to Web data. Our goal here is to identify real scenarios in which these techniques can be used to discover usage patterns from Web data. To do so, hereinafter we use the following notation:

- **WDT**: these techniques are designed essentially to support Web usage mining, i.e., the techniques address Web data applications only. We refer to these techniques as Web Data Techniques (WDT).
- **GPT**: these techniques can be used to support both public data release and Web-based applications. We refer to these techniques as General Purpose Techniques (GPT).

Cryptography-Based Techniques: these techniques can be used to support business collaboration on the Web since these solutions assume that the data are split across many sites. Scenario 2 (Section 2.3) is a typical example of Web-based application which can be addressed by cryptography-based techniques. Other applications related to e-commerce can be found in [55, 48, 34]. Therefore, such techniques are classified as WDT.

Generative-Based Techniques: these techniques can be applied to scenarios in which the goal is to extract useful knowledge from large, distributed data repositories. In these scenarios, the data cannot be directly centralized or unified as a single file or database either due to legal, proprietary or technical restrictions. In general, generative-based techniques are designed to support distributed Web-based applications.

Noise Addition Techniques: these techniques can be categorized as GPT. For instance, data swapping and data distortion techniques are used for public data release, while data randomization could be used to build models for on-line recommendations [65]. Scenario 1 (Section 2.3) is a typical example of an on-line recommendation system.

Space Transformation Techniques: these are general purpose techniques (GPT). In [45], the authors show that space transformation techniques could be used to promote social benefits as well as to address applications on the Web. An example of social benefit occurs, for instance, when a hospital shares some data for research purposes (e.g., cluster of patients with the same diseases). Space transformation techniques can also be used when the data mining process is outsourced or even when the data are distributed across many sites.

Blocking-Based Techniques: in general, these techniques are applied to protect sensitive information in databases. They could be used to simulate an access control in a database in which some information is hidden from users who do not have the right to access it. However, these techniques can also be used to suppress confidential information before the release of data for mining. We classify such techniques as GPT.

Sanitization-Based Techniques: Like blocking-based techniques, sanitization-based techniques can be used by statistical offices who publish sanitized version of data (e.g., census problem). In addition, sanitization-based techniques can be used to build models for on-line recommendations as described in Scenario 1 (Section 2.3).

Data Ownership Techniques: Although these techniques can be used to general purposes, the most evident applications of such techniques are related to Web mining and on-line business-to-business (B2B) interactions.

Table 1 shows a summary of the PPDM techniques and their relationship with Web data applications.

PPDM Techniques	Category
Cryptography-Based Techniques	WDT
Generative-Based Techniques	WDT
Noise Addition Techniques	GPT
Space Transformation Technique	GPT
Blocking-Based Techniques	GPT
Sanitization-Based Techniques	GPT
Data Ownership Techniques	GPT

Table 1: A summary of the PPDM techniques and their relationship with Web data.

5 Requirements for Technical Solutions

In this section, we suggest some requirements that are related to industrial initiatives. These requirements are essential for the development and deployment of technical solutions.

5.1 Requirements for the development of technical solutions

Ideally, a technical solution for a PPDM scenario would enable us to enforce privacy safeguards and to control the sharing and use of personal data. However, such a solution raises some crucial questions:

- What levels of effectiveness are in fact technologically possible and what corresponding regulatory measures are needed to achieve these levels?
- What degrees of privacy and anonymity must be sacrificed to achieve valid data mining results?

These questions cannot have "yes-no" answers, but involve a range of technological possibilities and social choices. The worst response to such questions is to ignore them completely and not pursue the means by which we can eventually provide informed answers.

Technology alone cannot address all of the concerns surrounding PPDM scenarios [3]. The above questions can be to some extent addressed if we provide some key requirements to guide the development of technical solutions.

The following key words are used to specify the extent to which an item is a requirement for the development of technical solutions to address PPDM:

- **Must:** this word means that the item is an absolute requirement;
- **Should:** this word means that there may exist valid reasons not to treat this item as a requirement, but the full implications should be understood and the case carefully weighed before discarding this item.

Independence: A promising solution for the problem of PPDM, for any specific data mining task (e.g. association rules, clustering, classification), *should* be independent of the mining task algorithm.

Accuracy: When it is possible, an effective solution *should* do better than a trade-off between privacy and accuracy on the disclosure of data mining results. Sometimes a trade-off *must* be found as in scenario 2, in Section 2.3.

Privacy Level: This is also a fundamental requirement in PPDM. A technical solution *must* ensure that the mining process does not violate privacy up to a certain degree of security.

Attribute Heterogeneity: A technical solution for PPDM *should* handle heterogeneous attributes (e.g. categorical and numerical).

Versatility: A versatile solution to address the problem of PPDM *should* be applicable to different kinds of information repository, i.e., the data could be centralized, or even distributed horizontally or vertically.

Communication Cost: When addressing data distributed across many sites, a technical solution *should* consider carefully issues of communication cost.

5.2 Requirements to guide the deployment of technical solutions

Information technology vendors in the near future will offer a variety of products which claim to help protect privacy in data mining. How can we evaluate and decide whether what is

being offered is useful? The nonexistence of proper instruments to evaluate the usefulness and feasibility of a solution to address a PPDM scenario challenge us to identify the following requirements:

Privacy Identification: We should identify what information is private. Is the technical solution aiming at protecting individual privacy or collective privacy?

Privacy Standards: Does the technical solution comply with international instruments that state and enforce rules (e.g. principles and/or policies) for use of automated processing of private information?

Privacy Safeguards: Is it possible to record what has been done with private information and be transparent with individuals about whom the private information pertains?

Disclosure Limitation: Are there metrics to measure how much private information is disclosed? Since privacy has many meanings depending on the context, we may require a set of metrics to do so. What is most important is that we need to measure not only how much private information is disclosed, but also measure the impact of a technical solution on the data and on valid mining results.

Update Match: When a new technical solution is launched, two aspects should be considered: a) the solution should comply with existing privacy principles and policies; b) in case of modifications to privacy principles and/or policies that guide the development of technical solutions, any release should consider these new modifications.

6 Future Research Trends

Preserving privacy on the Web has an important impact on many Web activities and Web applications. In particular, privacy issues have attracted a lot of attention due to the growth of e-commerce and e-business. These issues are further complicated by the global and self-regulatory nature of the Web.

Privacy issues on the Web are based on the fact that most users want to maintain strict anonymity on Web applications and activities. The ease access to information on the Web, coupled with the ready availability of personal data, also made it easier and more tempting for interested parties (e.g. businesses and governments) to willingly or inadvertently intrude on individuals' privacy in unprecedented ways.

Clearly, privacy issues on Web data is an umbrella that encompasses many Web applications such as e-commerce, stream data mining, multimedia mining, among others. In this work, we focus on issues toward foundation for further research in PPDM on the Web because these issues will certainly play a significant role in the future of this new area. In particular, a common framework for PPDM should be conceived, notably in terms of definitions, principles, policies, and requirements. The advantages of a framework of that nature are as follows: (a) a common framework will avoid confusing developers, practitioners, and many others interested in PPDM on the Web; (b) adoption of a common framework will inhibit inconsistent efforts in different ways, and will enable vendors and developers to make solid advances in the future of research in PPDM on the Web.

The success of a framework of this nature can only be guaranteed if it is backed up by a legal framework, such as the Platform for Privacy Preferences (P3P) Project [29, 49]. This project is emerging as an industry standard providing a simple, automated way for users to gain more control over the use of personal information on Web sites they visit.

The European Union has taken a lead in setting up a regulatory framework for Internet Privacy and has issued a directive which sets guidelines for processing and transfer of personal data [15].

7 Summary

It is commonly agreed upon that data privacy is the onus of the custodian of the data. However, to whom encumbers the responsibility for the privacy of the patterns discovered from the data

is still unclear and unstipulated.

In this chapter, we have laid down the foundations for further research in the area of Privacy-Preserving Data Mining (PPDM) on the Web. Although our work described in this chapter is preliminary and conceptual in nature, it is a vital prerequisite for the development and deployment of new techniques. In particular, we described the problems we face in defining what information is private in data mining, and discussed how privacy can be violated in data mining. We described the basis of PPDM including the historical roots, the definition of privacy preservation in data mining, and the general parameters that characterizes scenarios in PPDM. We then analyzed the implications of the Organization for Economic Cooperation and Development (OECD) data privacy principles in knowledge discovery. As a consequence, we suggested some policies for PPDM based on the OECD privacy guidelines. We also introduced a taxonomy of the existing PPDM techniques and a discussion on how these techniques are applicable on Web data. Subsequently, we suggested some desirable privacy requirements that are related to industrial initiatives. These requirements are essential for the development and deployment of technical solutions. Finally, we pointed to standardization issues as a technical challenge for future research trends in PPDM on the Web.

References

[1] M. Ackerman, L. Cranor, and J. Reagle. Privacy in E-Commerce: Examining User Scenarios and Privacy Preferences. In *Proc. of the ACM Conference on Electronic Commerce*, pages 1–8, Denver, Colorado, USA, November 1999.

[2] D. Agrawal and C. C. Aggarwal. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In *Proc. of ACM SIGMOD/PODS*, pages 247–255, Santa Barbara, CA, May 2001.

[3] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. Hippocratic Databases. In *Proc. of the 28th Conference on Very Large Data Bases*, Hong Kong, China, August 2002.

[4] R. Agrawal and R. Srikant. Privacy-Preserving Data Mining. In *Proc. of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 439–450, Dallas, Texas, May 2000.

[5] L. Brankovic and V. Estivill-Castro. Privacy Issues in Knowledge Discovery and Data Mining. In *Proc. of Australian Institute of Computer Ethics Conference (AICEC99)*, Melbourne, Victoria, Australia, July 1999.

[6] C. Clifton, M. Kantarcioğlu, and J. Vaidya. Defining Privacy For Data Mining. In *Proc. of the National Science Foundation Workshop on Next Generation Data Mining*, pages 126–133, Baltimore, MD, USA, November 2002.

[7] C. Clifton, M. Kantarcioğlu, J. Vaidya, X. Lin, and M. Y. Zhu. Tools For Privacy Preserving Distributed Data Mining. *SIGKDD Explorations*, 4(2):28–34, December 2002.

[8] C. Clifton 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.

[9] S. Cockcroft and P. Clutterbuck. Attitudes Towards Information Privacy. In *Proc. of the 12th Australasian Conference on Information Systems*, Coffs Harbour, NSW, Australia, December 2001.

[10] C. Collberg, C. Thomborson, and D. Low. A Taxonomy of Obfuscating Transformations. Technical report, TR–148, Department of Computer Science, University of Auckland, New Zealand, July 1997.

[11] M. J. Culnan. How Did They Get My Name?: An Exploratory Investigation of Consumer Attitudes Toward Secondary Information. *MIS Quartely*, 17(3):341–363, September 1993.

[12] E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino. Hiding Association Rules by Using Confidence and Support. In *Proc. of the 4th Information Hiding Workshop*, pages 369–383, Pittsburg, PA, April 2001.

[13] W. Du and M. J. Atallah. Secure Multi-Party Computation Problems and their Applications: A Review and Open Problems. In *Proc. of 10th ACM/SIGSAC 2001 New Security Paradigms Workshop*, pages 13–22, Cloudcroft, New Mexico, September 2001.

[14] V. Estivill-Castro and L. Brankovic. Data Swapping: Balancing Privacy Against Precision in Mining for Logic Rules. In *Proc. of Data Warehousing and Knowledge Discovery DaWaK-99*, pages 389–398, Florence, Italy, August 1999.

[15] European Comission. The directive on the protection of individuals with regard of the processing of personal data and on the free movement of such data, 1998. Available at http://www2.echo.lu.

[16] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting Privacy Breaches in Privacy Preserving Data Mining. In *Proc. of the 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2003)*, San Diego, CA, USA, June 2003.

[17] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy Preserving Mining of Association Rules. In *Proc. of the 8th ACM SIGKDD Intl. Conf. on Knowlegde Discovery and Data Mining*, pages 217–228, Edmonton, AB, Canada, July 2002.

[18] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In *Advances in Knowledge Discovery and Data Mining. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R. Uthurusamy (eds.)*, pages 1–34, MIT Press, Cambridge, MA, 1996.

[19] A. P. Felty and S. Matwin:. Privacy-Oriented Data Mining by Proof Checking. In *Proc. of the 6th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 138–149, Helsinki, Finland, August 2002.

[20] S. Garfinkel. *Database Nation: The Death of the Privacy in the* 21^{st} *Century.* O'Reilly & Associates, Sebastopol, CA, USA, 2001.

[21] O. Goldreich, S. Micali, and A. Wigderson. How to Play Any Mental Game - A Completeness Theorem for Protocols with Honest Majority. In *Proc. of the 19th Annual ACM Symposium on Theory of Computing*, pages 218–229, New York City, USA, May 1987.

[22] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA, 2001.

[23] V. S. Iyengar. Transforming Data to Satisfy Privacy Constraints. In *Proc. of the 8th ACM SIGKDD Intl. Conf. on Knowlegde Discovery and Data Mining*, pages 279–288, Edmonton, AB, Canada, July 2002.

[24] P. Jefferies. Multimedia, Cyberspace & Ethics. In *Proc. of International Conference on Information Visualisation (IV2000)*, pages 99–104, London, England, July 2000.

[25] G. H. John. Behind-the-Scenes Data Mining. *Newletter of ACM SIG on KDDM*, 1(1):9–11, June 1999.

[26] T. Johnsten and V. V. Raghavan. Impact of Decision-Region Based Classification Mining Algorithms on Database Security. In *Proc. of 13th Annual IFIP WG 11.3 Working Conference on Database Security*, pages 177–191, Seattle, USA, July 1999.

[27] T. Johnsten and V. V. Raghavan. Security Procedures for Classification Mining Algorithms. In *Proc. of 15th Annual IFIP WG 11.3 Working Conference on Database and Applications Security*, pages 293–309, Niagara on the Lake, Ontario, Canada, July 2001.

[28] T. Johnsten and V. V. Raghavan. A Methodology for Hiding Knowledge in Databases. In *Proc. of the IEEE ICDM Workshop on Privacy, Security, and Data Mining*, pages 9–17, Maebashi City, Japan, December 2002.

[29] R. Joseph and C. L. Faith. The Platform for Privacy Preferences. 42(2):48-55, 1999.

[30] M. Kantarcioğlu and C. Clifton. Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. In *Proc. of The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, Madison, Wisconsin, June 2002.

[31] M. Kantarcioğlu and J. Vaidya. Privacy Preserving Naïve Bayes Classifier for Horizontally Partitioned Data. In *Proc. of the IEEE ICDM Workshop on Privacy Preserving Data Mining*, pages 3–9, Melbourne, FL, USA, November 2003.

[32] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the Privacy Preserving Properties of Random Data Perturbation Techniques. In *Proc. of the 3rd IEEE International Conference on Data Mining (ICDM'03)*, pages 99–106, Melbourne, Florida, USA, November 2003.

[33] W. Klösgen. KDD: Public and Private Concerns. *IEEE EXPERT*, 10(2):55–57, April 1995.

[34] W. Kou and Y. Yesha. *Electronic Commerce Technology Trends: Challenges and Opportunities*. IBM Press Alliance Publisher: David Uptmor, IIR Publications, Inc., 2000.

[35] K. C. Laudon. Markets and Privacy. *Communication of the ACM*, 39(9):92–104, September 1996.

[36] Y. Lindell and B. Pinkas. Privacy Preserving Data Mining. In *Crypto 2000, Springer-Verlag (LNCS 1880)*, pages 36–54, Santa Barbara, CA, August 2000.

[37] S. Meregu and J. Ghosh. Privacy-Preserving Distributed Clustering Using Generative Models. In *Proc. of the 3rd IEEE International Conference on Data Mining (ICDM'03)*, pages 211–218, Melbourne, Florida, USA, November 2003.

[38] A. Mucsi-Nagy and S. Matwin. Digital Fingerprinting for Sharing of Confidential Data. In *Proc. of the Workshop on Privacy and Security Issues in Data Mining*, pages 11–26, Pisa, Italy, September 2004.

[39] D. E. O'Leary. Knowledge Discovery as a Threat to Database Security. In G. Piatetsky-Shapiro and W. J. Frawley (editors): Knowledge Discovery in Databases. AAAI/MIT Press, pages 507-516, Menlo Park, CA, 1991.

[40] D. E. O'Leary. Some Privacy Issues in Knowledge Discovery: The OECD Personal Privacy Guidelines. *IEEE EXPERT*, 10(2):48–52, April 1995.

[41] S. R. M. Oliveira and O. R. Zaïane. Privacy Preserving Frequent Itemset Mining. In *Proc. of the IEEE ICDM Workshop on Privacy, Security, and Data Mining*, pages 43–54, Maebashi City, Japan, December 2002.

[42] S. R. M. Oliveira and O. R. Zaïane. Algorithms for Balancing Privacy and Knowledge Discovery in Association Rule Mining. In *Proc. of the 7th International Database Engineering and Applications Symposium (IDEAS'03)*, pages 54–63, Hong Kong, China, July 2003.

[43] S. R. M. Oliveira and O. R. Zaïane. Protecting Sensitive Knowledge By Data Sanitization. In *Proc. of the 3rd IEEE International Conference on Data Mining (ICDM'03)*, pages 613–616, Melbourne, Florida, USA, November 2003.

[44] S. R. M. Oliveira and O. R. Zaïane. Achieving Privacy Preservation When Sharing Data For Clustering. In *Proc. of the Workshop on Secure Data Management in a Connected World (SDM'04) in conjunction with VLDB'2004*, pages 67–82, Toronto, Ontario, Canada, August 2004.

[45] S. R. M. Oliveira and O. R. Zaïane. Privacy-Preserving Clustering by Object Similarity-Based Representation and Dimensionality Reduction Transformation. In *Proc. of the Workshop on Privacy and Security Aspects of Data Mining (PSDM'04) in conjunction with the Fourth IEEE International Conference on Data Mining (ICDM'04)*, Brighton, UK, November 2004.

[46] S. R. M. Oliveira, O. R. Zaïane, and Y. Saygin. Secure Association Rule Mining. In *Proc. of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04)*, pages 74–85, Sydney, Australia, May 2004.

[47] G. Piatetsky-Shapiro. Knowledge Discovery in Personal Data vs. Privacy: A Mini-Symposium. *IEEE Expert*, 10(2):46–47, 1995.

[48] B. Pinkas. Cryptographic Techniques For Privacy-Preserving Data Mining. *SIGKDD Explorations*, 4(2):12–19, December 2002.

[49] Platform for Privacy Preferences (P3P) Project. Available at http://www.w3.org/P3P/.

[50] A. Rezgui, A. Bouguettaya, and M. Y. Eltoweissy. Privacy on the Web: Facts, Challenges, and Solutions. *IEEE Security & Privacy*, 1(6):40–49, Nov-Dec 2003.

[51] S. J. Rizvi and J. R. Haritsa. Maintaining Data Privacy in Association Rule Mining. In *Proc. of the 28th International Conference on Very Large Data Bases*, Hong Kong, China, August 2002.

[52] A. Rosenberg. Privacy as a Matter of Taste and Right. In E. F. Paul, F. D. Miller, and J. Paul, editors, The Right to Privacy, pages 68-90, Cambridge University Press, 2000.

[53] Y. Saygin, V. S. Verykios, and C. Clifton. Using Unknowns to Prevent Discovery of Association Rules. *SIGMOD Record*, 30(4):45–54, December 2001.

[54] F. D. Schoeman. Philosophical Dimensions of Privacy, Cambridge Univ. Press, 1984.

[55] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations*, 1(2):12–23, January 2000.

[56] P. Tendick and N. S. Matloff. Recent Results on the Noise Addition Method for Database Security. In *Proc. of the 1987 Joint Meetings, American Statistical Association / Institute of Mathematical Statistics (ASA/IMA)*, pages 406–409, Washington, DC, USA, 1987.

[57] E. Turban and J. E. Aronson. *Decision Support Systems and Intelligent Systems*. Prentice-Hall, New Jersey, USA, 2001.

[58] J. Vaidya and C. Clifton. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. In *Proc. of the 8th ACM SIGKDD Intl. Conf. on Knowlegde Discovery and Data Mining*, pages 639–644, Edmonton, AB, Canada, July 2002.

[59] J. Vaidya and C. Clifton. Privacy-Preserving K-Means Clustering Over Vertically Partitioned Data. In *Proc. of the 9th ACM SIGKDD Intl. Conf. on Knowlegde Discovery and Data Mining*, pages 206–215, Washington, DC, USA, August 2003.

[60] A. A. Veloso, W. Meira Jr., S. Parthasarathy, and M. B. Carvalho. Efficient, Accurate and Privacy-Preserving Data Mining for Frequent Itemsets in Distributed Databases. In *Proc. of the 18th Brazilian Symposium on Databases*, pages 281–292, Manaus, Brazil, October 2003.

[61] V. S. Verykios, A. K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni. Association Rule Hiding. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):434–447, 2004.

[62] S. D. Warren and L. D. Brandeis. The Right to Privacy. *Harvard Law Review*, 4(5):193–220, 1890.

[63] A. F. Westin. The Right to Privacy, Atheneum, 1967.

[64] C. W. Wu. Privacy Preserving Data Mining: A Signal Processing Perspective and a Simple Data Perturbation Protocol. In *Proc. of the IEEE ICDM Workshop on Privacy Preserving Data Mining*, pages 10–17, Melbourne, FL, USA, November 2003.

[65] N. Zang, S. Wang, and W. Zhao. A New Scheme on Privacy Preserving Association Rule Mining. In *Proc. of the 15th European Conference on Machine Learning (ECML) and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Pisa, Italy, September 2004.

[66] Y. Zhu and L. Liu. Optimal Randomization for Privacy Preserving Data Mining. In *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 761–766, Seattle, WA, USA, August 2004.