Chapter 3

Defining Privacy for Data Mining

Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya
Department of Computer Sciences, Purdue University
West Lafayette, IN 47907-2066

Abstract:
Privacy preserving data mining – getting valid data mining results without learning the underlying data values – has been receiving attention in the research community and beyond. It is unclear what *privacy preserving* means. This paper provides a framework and metrics for discussing the meaning of privacy preserving data mining, as a foundation for further research in this field.
Keywords: Privacy, Security

3.1 Introduction

There has recently been a surge in interest in privacy preserving data mining [Agrawal & Srikant2000, Agrawal & Aggarwal2001, Muralidhar, Sarathy, & Parsa2001, Kantarcioglu & Clifton2002, Vaidya & Clifton2002, Evfimievski *et al.*2002, Rizvi & Haritsa2002, Clifton & Estivill-Castro2002]. Even the popular press has picked up on this trend [Hamblen2002, Eisenberg2002]. However, the concept of what is meant by privacy isn't clear. In this paper we outline some of the concepts that are addressed

in this line of research, and provide a roadmap for defining and understanding privacy constraints.

Generally when people talk of privacy, they say "keep information about me from being available to others". This doesn't match the dictionary definition (Webster's), "freedom from unauthorized intrusion". It is this intrusion, or use of personal data in a way that negatively impacts someone's life, that causes concern. As long as data is not misused, most people do not feel their privacy has been violated. The problem is that once information is released, it may be impossible to prevent misuse. Utilizing this distinction – ensuring that a data mining project won't enable *misuse* of personal information – opens opportunities that "complete privacy" would prevent. To do this, we need technical and social solutions that ensure data will not be released.

The same basic concerns also apply to collections of data. Given a collection of data, it is possible to learn things that are not revealed by any individual data item. An individual may not be care about someone knowing their birth date, mother's maiden name, or social security number; but knowing all of them enables identity theft. This type of privacy problem arises with large, multi-individual collections as well. A technique that guarantees no individual data is revealed may still release information describing the collection as a whole. Such "corporate information" is generally the goal of data mining, but some results may still lead to concerns (often termed a secrecy, rather than privacy, issue.) The difference between such corporate privacy issues and individual privacy is not that significant. If we view disclosure of knowledge about an entity (information about an individual) as a potential individual privacy violation, then generalizing this to disclosure of information about a subset of the data captures both views.

In Section 3.4 we will give real-world examples of collection privacy problems. Section 3.3 discusses approaches to dealing with individual privacy. First, however, we give background on the two main classes of privacy preserving data mining.

3.2 Approaches to Privacy Preserving Data Mining

Two papers entitled *Privacy Preserving Data Mining* appeared in 2000. While both addressed a similar problem, constructing decision trees from private training data, the concepts of privacy were quite different. One was based on *data obscuration*, i.e., modifying the data values so real values are not disclosed [Agrawal & Srikant2000]. The other used Secure Multiparty Computation to "encrypt" data values [Lindell & Pinkas2000], ensuring that no party learns anything about another's data values. We first describe Secure Multiparty Computation, then give additional background on data obscuration. We also discuss a problem that has received little attention: How do we constrain data mining if it is possible that the results alone violate privacy?

3.2.1 Secure multiparty computation

The idea of Secure Multiparty Computation (SMC) [Yao1986, Goldreich, Micali, & Wigderson1987] is that the parties involved learn nothing but the results. Informally,

imagine we have a trusted third party to which all parties give their input. The trusted party computes the output and returns it to the parties.

SMC enables this without the trusted third party. There may be considerable communication between the parties to get the final result, but the parties don't learn anything from this communication. The computation is secure if given just one party's input and output from those runs, we can *simulate* what would be seen by the party. In this case, to simulate means that the distribution of what is actually seen and the distribution of the simulated view over many runs are computationally indistinguishable. We may not be able to exactly simulate every run, but over time we cannot tell the simulation from the real runs.

Since we could simulate the runs from knowing only our input and output, it makes sense to say that we don't learn anything from the run other than the output. This seems like a strong guarantee of privacy, and has been used in privacy preserving data mining work [Lindell & Pinkas2000, Du & Atallah2001a, Du & Atallah2001b].

We must be careful when using Secure Multiparty Computation to define privacy. For example, suppose we use a SMC technique to build a decision tree from databases at two sites [Lindell & Pinkas2000], classifying people into high and low risk for a sensitive disease. Assume that the non-sensitive data is public, but the sensitive data (needed as training data to build the classifier) cannot be revealed. The SMC computation won't reveal the sensitive data, but the resulting classifier will enable all parties to estimate the value of the sensitive data. It isn't that the SMC was "broken", but that the result itself violates privacy.

3.2.2 Obscuring data

Another approach to privacy is to obscure data: making private data available, but with enough noise added that exact values (or approximations sufficient to allow misuse) cannot be determined. One approach, typically used in census data, is to aggregate items. Knowing the average income for a neighborhood is not enough to determine the actual income of a resident of that neighborhood. An alternative is to add random noise to data values, then mine the distorted data. While this lowers the accuracy of data mining results, research has shown that the loss of accuracy can be small relative to the loss of ability to estimate an individual item. We can reconstruct the original distribution of a collection of obscured numeric values, enabling better construction of decision trees [Agrawal & Srikant2000, Agrawal & Aggarwal2001]. This would enable data collected from a web survey to be obscured at the source – the correct values would never leave the respondent's machine – ensuring that exact (misusable) data doesn't exist. Techniques have also been developed for association rules, enabling valid rules to be learned from data where items have been randomly added to or removed from individual transactions [Evfimievski *et al.*2002, Rizvi & Haritsa2002].

3.2.3 Perfect privacy

One problem with the above is the tradeoff between privacy and accuracy of the data mining results. Secure Multiparty Computation does better, but at a high computational and communication cost. In the "web survey" example, the respondents could engage

in a secure multiparty computation to obtain the results, and reveal no information that is not contained in the results. However, getting thousands of respondents to participate synchronously in a complex protocol is impractical. While useful in the corporate model, it is not appropriate for the web model. Here we present a solution based on moderately trusted third parties – the parties are not trusted with exact data, but trusted only not to collude with the "data receiver".

Assume the existence of k *untrusted, noncolluding* sites.

- *Untrusted* implies that none of these sites should be able to gain any useful information from any of the inputs of the local sites.
- *Noncolluding* implies that none of these sites should collude with any other sites to obtain information beyond the protocol.

Then, all of the local parties can split their local inputs into k random shares which are then split across the k untrusted sites. Each of these random shares are meaningless information by themselves. However, if any of the parties combined their data, they would gain some meaningful information from the combined data. For this reason, we require that the sites be noncolluding. We believe this assumption is not unrealistic. Each site combines the shares of the data it has received using a secure protocol to get the required data mining result.

The following is a brief description of this approach. Every party is assumed to have a single bit of information x_i, identified by some key i. Each party locally generates a random number r_i and then sends $(i, \bar{x_i} = x_i \oplus r_i)$ to one site and (i, r_i) to the second site. Note that neither site will be able to predict the x_i. Due to the xor operation $\oplus$, the input they see is indistinguishable from any uniformly generated random sequence. Given any data mining task(f) defined on $X = [x_1, x_2, \ldots, x_n]$, it suffices to evaluate $f(\bar{X} \bar{\oplus} R) = f(X)$ since $R = [r_1, r_2, \ldots, r_n]$ and $\bar{X} \bar{\oplus} R = [\bar{x_1} \oplus r_1, \bar{x_2} \oplus r_2, \ldots, \bar{x_n} \oplus r_n]$. It is a known fact that with the assumption of existence of trapdoor permutations (RSA is assumed to be a trapdoor permutation), any functionality g, $(g : \{0,1\}^* \times \{0,1\}^* \mapsto \{0,1\}^* \times \{0,1\}^*)$ can be evaluated privately in the semi-honest model [Goldreich, Micali, & Wigderson1987]. Since the initial xor operation can be easily represented as a circuit, given functionality f, we can define a functionality $g(X, R) = f(\bar{X} \bar{\oplus} R)$. Thus, any data mining functionality can be evaluated privately without revealing any information other than the final result. (For a more complete treatment, see [Kantarcioglu & Vaidya2002].)

While this solution is not especially efficient, indeed not even necessarily very practical for large quantities of data, it does demonstrate a method of maintaining perfect privacy while computing the required data mining function.

3.2.4 Limitations on results

How can we constrain the results of data mining? There has been work in this area, addressing specific problems such as hiding specific association rules [Atallah *et al.*1999, Saygin, Verykios, & Clifton2001] or limiting confidence in *any* data mining [Clifton2000]. While these provide some specific techniques, the means available to constrain results are limited. What is needed is a general way to specify what is and is not allowed.

One possible approach is constraint-based data mining [Bayardo2002]. This line of research is concerned with improving the efficiency of algorithms and understandability of results through providing up-front constraints on what results would be of interest. Would the languages used to describe these constraints also serve to define what results are *acceptable* from a privacy standpoint? While the current methods do not *enforce* that nothing outside the constraints can be learned, they could provide a starting point for further research.

The rest of this paper provides some specific suggestions for methods to specify privacy constraints in ways that still allow data mining. We start with a discussion of individual privacy. We then discuss corporate privacy, or constraining what is disclosed about subsets of the entire data. We conclude with several orthogonal metrics for defining and measuring privacy.

3.3 Individual Privacy

Most legal efforts have been directed to protecting data of the individual. For example, the European Community regulates *personal data* [EC 95/461995]:

> 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;

and specifies that data can be

> kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use.

The key element here is "identifiable": As long as the data cannot be traced to an individual, the regulations do not apply.

The U.S. HIPAA rules [HIPAA2001] are similar – they apply to *protected health information*, defined as individually identifiable health information:

> *Individually identifiable health information* is information that is a subset of health information, including demographic information collected from an individual, and:
>
> 1. Is created or received by a health care provider, health plan, employer, or health care clearinghouse; and
> 2. Relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual; and

(a) That identifies the individual; or

(b) With respect to which there is a reasonable basis to believe the information can be used to identify the individual.

Any information that cannot be traced to a specific individual falls outside the scope of most privacy laws. This provides one solution to privacy and data mining projects: As long as the data used is not individually identifiable, there should be no problems.

Another factor in individual data is how the data is collected and held. The U.S. HIPAA rules assume that data is created and held by a healthcare provider. This gives a *corporate* model – individually identifiable information first "appears" within a collection held by someone other than the individual. An alternative is the "world wide web" model, where individuals provide the data in electronic form themselves. These models lead to different solutions. We first look at some generally applicable solutions, then discuss some that are only relevant in the corporate model.

Data obscuration is effective both in the web and corporate model. Obscuration can be done by the individual (if the receiver isn't trusted), or by the holder of data (to reduce concerns about breached security.) However, obscuring data falls into a legal gray area. Rules such as EC 95/46 and HIPAA would probably view individually identifiable data with values obscured as protected, even if the exact values are unknowable. However, obscuration could be as or more effective at protecting actual data values than the aggregation methods used on publicly available census data. Demonstrating the effectiveness of data obscuration in comparison with census data could improve public acceptance, and lead to changes in legal standards.

Data obscuration techniques could also be used to ensure that otherwise identifiable data isn't individually identifiable. Re-identification experiments have shown that data that might be viewed as non-identifiable, such as birth date and postal code, can in combination allow identification of an individual [Sweeney2001]. Obscuring the data could make re-identification impossible, thus meeting both the spirit and letter of privacy laws.

3.4 Collection Privacy

Protecting individual data items may not be enough – we may need to protect against learning about subsets of a collection. Such issues are common in a data warehousing environment, where data from multiple sources is combined for analysis (see Figure 3.1). This requires that the warehouse be trusted to maintain the privacy of all parties - since it knows the source of data, it learns site-specific information as well as global results. Even techniques that prevent disclosure of individual data items may reveal rules, trends, or patterns about individual sites. This may reveal trade secrets, or embarrassing or damaging information. In a sense this is a scaled-up version of the individual privacy problem, however it is an area where the Secure Multiparty Computation approach is more likely to be applicable.

Addressing these issues requires understanding the reasons behind them. We now discuss two issues that lead to privacy concerns in collections of data, and ways to understand those that enable data mining to proceed.

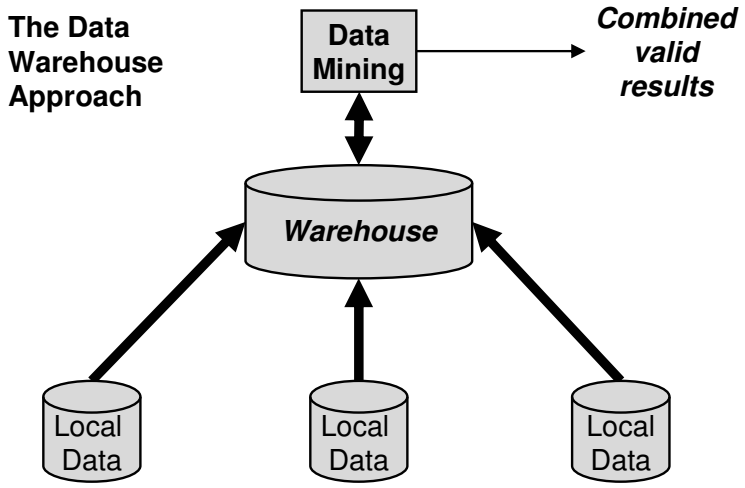

Figure 3.1: The Data Warehouse approach to mining distributed sources.

3.4.1 Secrecy

Individual privacy concerns can lead to corporate privacy concerns. The holder of a collection of individual data may be trusted by those individuals, but if that data is revealed, this trust (often protected contractually) is broken. The collection holder may be willing to participate in a distributed data mining project, but only if it can ensure that its own data items are not revealed. Secure Multiparty Computation would seem to provide a solution to this, however the problem of results revealing private information still remains.

Another issue is protecting the data holder. Even if we assume that

- Individual data items can be disclosed, or are protected by the privacy preserving methods; and
- Global data mining results do not violate the privacy/secrecy concerns

problems may still arise. Knowledge about a subset of the combined data set (as separable from the global results) may reveal secrets of one of the data holders.

As an example, a medical study may want to use data mining to establish overall trends from hospital data. Even if the techniques used protect patient privacy, they may reveal hospital-specific information. Rules establishing conditions that lead to a high complication rate for a particular operation would be useful study results. However, if these conditions are tied to a particular hospital, there may be liability or public relations implications. Such implications may limit willingness to participate in such a study. We can develop efficient techniques for data mining that protect such study [Kantarcioglu & Clifton2002].

In other cases, participants in the data mining project may have specific secrets they wish to withhold, such as trade secrets. An industry consortium may want to mine data

to find ways that all members can use to improve their processes. However, learning a particular member's trade secret, and sharing that with the rest of the consortium, is inappropriate. The ability to protect secrets while otherwise participating in a data mining exercise will expand the applicability of data mining.

3.4.2 Limitations on collaboration

At other times, it may not be the participants that are concerned about sharing data, but external parties. As an example, U.S. antitrust regulations limit the ability of companies to collaborate. The basic premise is protecting the consumer: collaborations that harm the consumer (e.g., price-fixing cartels) are illegal. The problem is establishing that collaboration *is* to the consumers' benefit. If the CEOs of Ford and General Motors meet privately, the public doesn't know if there is illegal collaboration, likely triggering an investigation. Privacy preserving data mining techniques provide a solution to this. If we can prove that no information is shared other than the results, it is easier to justify that the collaboration is for legal purposes. As we discuss in the next section, this leads to a "need to know" criteria for determining acceptable information sharing.

3.5 Measures of Privacy Preservation

We have discussed some issues imposing privacy constraints on data mining. How do we translate these into solutions that address the issues? The key is an ability to measure privacy. Since privacy has many meanings, we require a set of metrics. Several suggestions are given in this section.

3.5.1 Bounded knowledge

The data obscuration approach leads to a *bounded knowledge* metric. Bounded knowledge implies that some information about a protected attribute may be revealed, but the actual value can only be estimated. This may be based on hard bounds (e.g., by adding noise from a random variable uniformly distributed within the bounds), or probabilistic estimates (e.g., by adding noise from a gaussian distribution).

A good measure for quantifying privacy based on such bounded estimation is given by [Agrawal & Aggarwal2001]. They propose a measure based on the *differential entropy* of a random variable. Specifically, if we add noise from a random variable A, the privacy is:

$$\Pi(A) = 2^{-\int_{\Omega_A} f_A(a) log_2 f_A(a) da}$$

where Ω_A is the domain of A.

This metric has several nice features. It is intuitively satisfying in simple cases. For noise generated from A, a uniformly distributed random variable between 0 and a, $\Pi(A) = a$. Thus this privacy metric is exactly the width of the unknown region. Furthermore, if a sequence of random variables A_i converges to B, then $\Pi(A_i)$ converges to $\Pi(B)$. For most random variables, e.g., a gaussian, the notion of width of the unknown region does not make sense. However, we can calculate Π, and the above

properties allow us to make the case that the privacy is equivalent to having no knowledge of the value except that it is within a region of width Π.

The authors extend this definition to *conditional privacy*, capturing the possibility that the inherent privacy from obscuring data may be reduced by what we can learn from a collection. They show how this can be applied to measure the actual privacy after reconstructing distributions of the original data to improve the accuracy of decision trees build on the obscured data [Agrawal & Srikant2000, Agrawal & Aggarwal2001]. Similar analyses on other data obscuration techniques would provide an effective way to compare those techniques.

3.5.2 Need to know

The *Need to know* concept is well established in controlling access to data. In the U.S., access to classified data requires both a security clearance and a justification of why the data should be accessed. The same concept appears in the EC95/46 directive:

> Member States shall provide that personal data may be processed only if:
> (a) the data subject has unambiguously given his consent; or
> (b) processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract; or ...

Note clause (b): It is acceptable to use individual data if it is needed to achieve a result requested by the individual.

This same standard also applies to corporate privacy problems. A need to know standard can be used to decide if collaboration between companies falls afoul of antitrust regulations. For example, airlines in the United States make their fares available on shared reservation systems used by travel agents (Sabre, Apollo). This allows easy comparison of prices by both consumers and the airlines – and a quick check will show that the airlines generally offer the same prices on the same routes. Airlines used to put out a notice of proposed prices, and if other airlines didn't match the price the notice would be removed. This gives the appearance of colluding to fix prices, illegal under U.S. antitrust law. The airlines were found to engage in such price fixing, and ordered to stop.

Operations today are similar – the shared reservations systems exist and prices are the same on the same routes. The only change is that the proposed prices no longer exist. Instead, an airline must actually change its prices, then see if other airlines go along. Why is this allowed, even though the end result is the same?

The key is that the current system provides two key benefits to consumers:

1. Price competition among the airlines – they are *allowed* to offer as low a price as they want; and

2. Easy comparison shopping – consumers can check prices, and take the lowest.

The first benefit cannot be accomplished without allowing airlines to change their prices. The second, giving consumers the ability to see and compare prices, also enables the airlines to see and compare prices. Sharing currently available prices thus

meets the *need to know* standard. The previous "notice of proposed pricing" approach did not, as this information was not shared with or useful to consumers. The information sharing that did not meet the "need to know" standard, the notice of proposed pricing, was found to be illegal.

Secure Multiparty Computation [Yao1986, Goldreich, Micali, & Wigderson1987] provides a basis for data mining that meets a need to know standard. If the results of the data mining are required to accomplish an allowable task, then learning those results should be allowed under the "need to know" standard under which most privacy regulations allow release of information. Data mining approaches that are Secure Multiparty Computations can be proven not to disclose anything except the results. However, Secure Multiparty Computation based techniques are not sufficient. We still need to ensure that the data mining results do fall under the need to know standard. Complicating matters is the fact that we do not know the results in advance, so we must demonstrate that any possible data mining results fall under need to know. Alternatively we can define a limited subset of data mining results that are acceptable, and ensure that the computation terminates without releasing information (other than the lack of results) if the results do not fall in this subset. This requires that the computation be customized for each data mining project.

Need to know is interesting in that it is a binary measure. A result is either acceptable (it is required to accomplish the end goal), or unacceptable. It is also difficult in that it can apply to a collection of results rather than a single result; two different result sets may each be sufficient to achieve a goal (either alone meets the need to know standard), but knowing both result sets exceeds need to know. Formal definitions of this criteria are probably domain specific, however the legal and social acceptability of this "measure" suggests that further research is warranted.

3.5.3 Protected from disclosure

Another problem is when we have specific items we want to protect. This may be individual data items, specific rules we don't want disclosed, or even general classes of knowledge that must be protected. The database security community has developed effective techniques for *inference prevention* [Delugach & Hinke1996, Hinke, Delugach, & Wolf1997, Yip & Levitt1998]. This work is for "provable facts", or inferences that are always true. Data mining finds inferences that are interesting, but do not always hold. Methods have been proposed to alter data to bring the support or confidence of specific rules below a threshold [Atallah *et al.*1999, Saygin, Verykios, & Clifton2001], but choice of an appropriate threshold is still difficult. Alternatively, we may want to protect against association rules that involve a particular outcome (e.g., any rules that pertain to equipment failures), but the problem of defining when a rule is considered strong enough to violate privacy concerns is still a problem.

One option is to use classification as a measure. Many data mining problems can be expressed in terms of classification, e.g., association rules can be used as decision rules. If the ability to classify is limited, many other types of data mining will be limited as well. This idea has been used to evaluate the risk posed by data mining when the knowledge to be protected is not known [Clifton2000]. However, this assumes the goal is prevention of *any* data mining. Use of classification as a metric to prohibit

learning specific facts is ripe for exploration. For example, we could state "it should be impossible to learn a classifier from the data that can predict a person having AIDS with $P(false\ hit) < .9$" – any classifier suggesting someone had AIDS would be wrong at least 90% of the time. This would alleviate concerns that even if individually identifiable data didn't contain the protected attribute (has AIDS), the data might enable someone to *learn* that attribute.

3.5.4 Anonymity

While protecting against learning a particular attribute may address some concerns, it requires that we know the potentially misused attribute in advance. It may also over-constrain the problem – learning statistics that apply to an entire population allows us to better predict those statistics for an individual, but need not violate the standards of use of "individually identifiable data". What is needed is a result-independent method of stating that data mining results do not violate individual privacy, even if they allow us to learn something that can be applied to an individual.

One method for protecting individual privacy is k-anonymity [Samarati & Sweeney1998]. The goal of k-anonymity is to only release data where for all possible queries, at least k results will be returned. To achieve this result, generalization and suppression techniques are used. In generalization techniques, some attributes are replaced with more general values so that k people will be found with any attribute value. For example, exact ages are replaced by some age ranges. In suppression techniques, data points that may cause too much generalization may be eliminated or a column that has identifying information can be deleted. Although this approach works well for individual data, it is not directly applicable to restricting privacy-preserving data mining results.

We propose the following definition of individual privacy that maintains the spirit of k-anonymity, but protects against data mining *results* that can be used to predict information about an individual.

Definition 3.5.1 *Two records that belonging to different individuals I_1, I_2 are p-indistinguishable given data X if for every polynomial-time function $f : I \mapsto \{0, 1\}$*

$$|Pr\{f(I_1) = 1|X\} - Pr\{f(X_2) = 1|X\}| \leq p$$

where $0 < p < 1$.

Informally this means that given a set of (privacy preserving) data, we are unable to learn any classifier that would distinguish between two individuals based on the data we've seen about those individuals. This doesn't necessarily mean we can't learn a good classifier, only that we can't use it to distinguish between the individuals.

From this, we can define a privacy preserving data mining process as one that does not enable us to distinguish between any individuals based solely on that data mining process.

Definition 3.5.2 *A data mining process said to be p-individual privacy preserving if using all of the information X seen during the data mining process, any two individual records are p-indistinguishable.*

Example: Assume that we are using the model described in Section 3.2.3 to find the count of particular attribute value, e.g., we want to know the number of people that have a particular cancer type where this information is represented as binary attribute. Each user i sends $(i, X_i \bigoplus r_i)$ to one noncolluding site and (i, r_i) to a second noncolluding site, where r_i is a random bit and $\bigoplus$ is the xor operation. Clearly the data gathering process is individual privacy preserving, because what both sites see about individuals is indistinguishable from random data. So the probability of predicting that each individual has the cancer is the same. Now let us find the count of the individuals who have the cancer. To do this, assume the noncolluding sites add $\frac{d(1-p)}{p}$ noisy data items to actual data (d is number of actual data entries). The sites then count the support including the noise items. (Details of this procedure can be found in [Kantarcioglu & Vaidya2002].) Due to potential re-use of the data actual id's i is replaced by random permutation and fake data is randomly distributed among the original data. When one site receives the count, it subtracts the amount of fake support and learns the actual result. A third noncolluding site joins the $(i_j, X_i \bigoplus r_i)$ with (i_j, r_i). Clearly, the first two noncolluding sites do not learn anything from this process. After the process, the probability that any given individual has the cancer is the same, even though there may be huge difference between prior and posterior probabilities.

We now show that p-indistinguishability holds for any two data entries on the third (join) site. Assume that the join site compares two randomly ordered data items. Since with probability p any item compared is fake, the probability that any statement is true is less than p. In other words, $Pr\{f(X_{org}) = 1|X_{seen}\} = p * Pr\{f(X_{org}) = 1|X_{seen}$ is true$\} + (1-p) * 0 \leq p$. We can see that $|Pr\{f(X_1) = 1|X_1$ is seen$\} - Pr\{f(X_2) = 1|X_2$ is seen$\} \leq p$. (In the worst case, one probability will be zero and one will be p.) Since p-indistinguishability is satisfied for any given data item pairs on the join site, we can conclude that the data mining process is individual privacy preserving assuming no collusion.

3.6 Conclusions

Privacy preserving data mining has the potential to increase the reach and benefits of data mining technology. However, we must be able to justify that privacy is preserved. For this, we need to be able to communicate what we mean by "privacy preserving". The current mixture of definitions, with each paper having its own definition of what privacy is maintained, will lead to confusion among potential adopters of the technology.

We have presented some suggestions for defining, measuring, and evaluating privacy preservation. We showed how these relate to both privacy policy and practice in the wider community, and to techniques in privacy preserving data mining. The key point to remember is that privacy preserving data mining is possible. Technology has been, and is being, developed to allow data mining without disclosing private information. There are legal and historical definitions of privacy that can be used to justify that this technology does preserve privacy.

This is by no means the definitive word on the subject. While some measures, such as the differential entropy metric of [Agrawal & Aggarwal2001], have clear mathe-

matical foundations and applications, others (such as using classification accuracy as a means of protecting rules from disclosure) have strong potential for further development. Adopting a common framework for discussion of privacy preservation will enable next generation data mining technology to make substantial advances in alleviating privacy concerns.

Bibliography

[Agrawal & Aggarwal2001] Agrawal, D., and Aggarwal, C. C. 2001. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, 247–255. Santa Barbara, California, USA: ACM.

[Agrawal & Srikant2000] Agrawal, R., and Srikant, R. 2000. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*, 439–450. Dallas, TX: ACM.

[Atallah *et al.*1999] Atallah, M.; Bertino, E.; Elmagarmid, A.; Ibrahim, M.; and Verykios, V. 1999. Disclosure limitation of sensitive rules. In *Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, 25–32.

[Bayardo2002] 2002. Special issue on constraints in data mining. *SIGKDD Explorations* 4(1).

[Clifton & Estivill-Castro2002] Clifton, C., and Estivill-Castro, V., eds. 2002. *IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, volume 14. Maebashi City, Japan: Australian Computer Society.

[Clifton2000] Clifton, C. 2000. Using sample size to limit exposure to data mining. *Journal of Computer Security* 8(4):281–307.

[Delugach & Hinke1996] Delugach, H. S., and Hinke, T. H. 1996. Wizard: A database inference analysis and detection system. *IEEE Transactions on Knowledge and Data Engineering* 8(1).

[Du & Atallah2001a] Du, W., and Atallah, M. J. 2001a. Privacy-preserving cooperative scientific computations. In *14th IEEE Computer Security Foundations Workshop*, 273–282.

[Du & Atallah2001b] Du, W., and Atallah, M. J. 2001b. Privacy-preserving statistical analysis. In *Proceeding of the 17th Annual Computer Security Applications Conference*.

[EC 95/461995] 1995. Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities* No I.(281):31–50.

[Eisenberg2002] Eisenberg, A. 2002. With false numbers, data crunchers try to mine the truth. *New York Times*.

[Evfimievski *et al.*2002] Evfimievski, A.; Srikant, R.; Agrawal, R.; and Gehrke, J. 2002. Privacy preserving mining of association rules. In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 217–228.

[Goldreich, Micali, & Wigderson1987] Goldreich, O.; Micali, S.; and Wigderson, A. 1987. How to play any mental game - a completeness theorem for protocols with honest majority. In *19th ACM Symposium on the Theory of Computing*, 218–229.

[Hamblen2002] Hamblen, M. 2002. Privacy algorithms: Technology-based protections could make personal data impersonal. *Computerworld*.

[Hinke, Delugach, & Wolf1997] Hinke, T. H.; Delugach, H. S.; and Wolf, R. P. 1997. Protecting databases from inference attacks. *Computers and Security* 16(8):687–708.

[HIPAA2001] 2001. Standard for privacy of individually identifiable health information. *Federal Register* 66(40).

[Kantarcioglu & Clifton2002] Kantarcioglu, M., and Clifton, C. 2002. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In *The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02)*, 24–31.

[Kantarcioglu & Vaidya2002] Kantarcioglu, M., and Vaidya, J. 2002. An architecture for privacy-preserving mining of client information. In Clifton, C., and Estivill-Castro, V., eds., *IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, volume 14, 37–42. Maebashi City, Japan: Australian Computer Society.

[Lindell & Pinkas2000] Lindell, Y., and Pinkas, B. 2000. Privacy preserving data mining. In *Advances in Cryptology – CRYPTO 2000*, 36–54. Springer-Verlag.

[Muralidhar, Sarathy, & Parsa2001] Muralidhar, K.; Sarathy, R.; and Parsa, R. A. 2001. An improved security requirement for data perturbation with implications for e-commerce. *Decision Science* 32(4):683–698.

[Rizvi & Haritsa2002] Rizvi, S. J., and Haritsa, J. R. 2002. Maintaining data privacy in association rule mining. In *Proceedings of 28th International Conference on Very Large Data Bases*, 682–693. VLDB.

[Samarati & Sweeney1998] Samarati, P., and Sweeney, L. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*.

[Saygin, Verykios, & Clifton2001] Saygin, Y.; Verykios, V. S.; and Clifton, C. 2001. Using unknowns to prevent discovery of association rules. *SIGMOD Record* 30(4):45–54.

[Sweeney2001] Sweeney, L. 2001. *Computational Disclosure Control: A Primer on Data Privacy Protection*. Ph.D. Dissertation, Massachusetts Institute of Technology.

[Vaidya & Clifton2002] Vaidya, J. S., and Clifton, C. 2002. Privacy preserving association rule mining in vertically partitioned data. In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 639–644.

[Yao1986] Yao, A. C. 1986. How to generate and exchange secrets. In *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science*, 162–167. IEEE.

[Yip & Levitt1998] Yip, R., and Levitt, K. 1998. The design and implementation of a data level database inference detection system. In *Proceedings of the Twelfth Annual IFIP WG 11.3 Working Conference on Database Security*.