Barriers to Adopting Privacy-preserving Data Mining

Richard Huebner
Norwich University

ABSTRACT

The primary issue examined in this research is that privacy-preserving data mining (PPDM) research has produced theoretical solutions and many peer-reviewed articles claiming to solve the problem. In order to gain any real benefit from the theoretical solutions, practitioners must attempt to convert that theory into practical software- and hardware- based solutions. This article begins with a review of data mining, privacy, and privacy-preserving data mining. It then reviews and analyzes the barriers that prevent widespread adoption of privacy-preserving data mining solutions. The article concludes by presenting recommendations and ideas for future work.

Keywords: data mining, privacy, privacy-preserving data mining, technology adoption

INTRODUCTION

Data mining can violate individual privacy (Yang, Zhong, & Wright, 2005). This is due to potential misuse of private or sensitive information inferred from data mining results (Vaidya, Clifton, & Zhu, 2006). Privacy preservation in data mining has emerged as a significant research field because of the ubiquity of demographic and sensitive data (Aggarwal, Pei, & Zhang, 2006). In addition, there is a need to extract knowledge from databases without revealing information about specific individuals (Vaidya, et al., 2006). This is especially true when sharing data across organizational boundaries (Xiong, Chitti, & Liu, 2007). When data mining results are presented to the user, inferences can be made about specific individuals (Zhu & Liu, 2004). It is these inferences that can be misused and therefore violate privacy (Vaidya, et al., 2006). The general concern is over the misuse of data mining results. Privacy-preserving data mining (PPDM) research strives to ensure that the privacy of each individual is maintained, yet present data mining results as accurately as possible. Data mining solutions that are privacy-aware should thus strive to provide highly accurate result sets while maintaining individual privacy.

The primary problem that will be examined in this article is that privacy-preserving data mining has generally not been adopted by industry (Clifton, Kantarcioglu, Vaidya, Lin, & Zhu, 2002; Vaidya, et al., 2006). There are “more papers than real-world solutions” (Clifton, et al., 2002, p. 28). Clifton et. al. (2002) noted that since there are so many solutions in PPDM, it is difficult to simplify the research to the point that the solutions can be developed and implemented. Privacy-preserving data mining algorithms have been published in the research community in leading computing journals, yet the most obvious problem is that PPDM-enabled tools have not been widely adopted. Some of the journals in which PPDM articles have appeared include: *ACM SIGMOD Record*, *Computers and Security*, *Ethics and Information Technology*, *IEEE Security and Privacy*, and *IEEE Transactions on Knowledge and Data Engineering*. Up to this point, research has focused on developing privacy-preserving algorithms that protect the confidentiality of individual information (Vaidya, et al., 2006). There is little evidence that these techniques have been adopted by industry. What are the barriers to successful development, implementation, and adoption of privacy-preserving data mining solutions?

BACKGROUND

Privacy-preserving data mining has emerged due to the following reasons. First, there are legal requirements for protecting data. Second, there are liabilities from inadvertent disclosure of data. Third, organizations need to share information with its partners, but do not want to provide certain types of data when they do so (Vaidya, et al., 2006). The Health Insurance Portability and Accountability Act (HIPAA), enacted by the U.S. Congress in 1996, required the establishment of national standards for electronic health care transactions and provides for the security and privacy of individually identifiable health information. The privacy discussed in the HIPAA rules refers to information privacy, or the prevention of disclosure of personal information (Moskop, Marco, Larkin, Geiderman, & Derse, 2005). In addition to HIPAA, there are requirements for protecting children’s online privacy. The Children’s Online Privacy Protection Act (COPPA) requires that organizations that collect or maintain personal information to 1) provide notice on the website of what information is collected, and 2) to obtain verifiable parental consent for its collection, use, or disclosure ("Children's Online Privacy Protection Act," 1998).

Preventing individual information disclosure has become increasingly important due to the number and size of data breaches during the last five years. There are countless examples of inadvertent disclosure of data. For example, in May, 2006, the Social Security numbers of about 26.5 million U.S. veterans were stolen in a random burglary from a VA employee's house where a laptop was stolen (Torres, 2007). In Edmonton, Canada, a security breach occurred where children had their medical information stolen. The medical information of 270 children was stored on a small flash drive (also known as a memory stick or thumb drive) and had been placed in an employee's purse, which subsequently was stolen. The flash drive contained children's personal health numbers, names, dates of service, and diagnoses (Unknown, 2007). As many as 200,000 credit and debit card numbers were compromised due to a security breach at TJX Companies, a Framingham, Massachusetts-based company (Abelson, 2007). This particular breach has resulted in multiple cases of fraudulent activity with the stolen numbers as well as at least one case of identity theft (Abelson, 2007).

Outsourcing and information sharing have become commonplace due to advances in distributed computing and Internet technologies (Xiong, et al., 2007). Xiong et al. noted that with the increasing need to share data, protecting that data has also become important because sharing data with organizations in countries that have lesser privacy and security standards creates additional challenges. Organizations also put themselves at risk when they outsource their data processing activities to third-party vendors (Xiong, et al., 2007).

Despite the pervasiveness of information sharing, one study showed that control over one's own individual data is a central concern for consumers (Han & Maclaurin, 2002). In this study, consumers were much more concerned about their own privacy than the organizational benefits of data mining. Consumers felt that organizations would use information in ways that exploited privacy. However, the study also found that those organizations that posted privacy policies or notices on their web sites were able to put consumers more at ease when shopping online. Consumers were generally skeptical of organizational use of data mining due to their concerns about privacy issues. Organizations could do a better job of putting customers at ease of they: 1) explained the benefits of data mining to consumers and be up-front with consumers on how data is going to be used; 2) control the amount of outbound calls and e-mail that use personal information in conjunction with an e-mail marketing campaign; 3) focus on building trust online by including secure shopping cart technology and encrypting credit card information when completing the check-out process; 4) develop separate privacy notices for each specific customer segments (Han et al., 2002).

Why should organizations be concerned about protecting individual information privacy? First, organizations should establish trust between itself and its customers (Marcella & Stucki, 2003). However, privacy and trust are difficult to define since it means something different to each person. A review of the literature shows that there are many different definitions of privacy. One definition of privacy is the extent to which others have limited access to personal and sensitive information about one's intimacies, thoughts, or body (Persson & Hansson, 2003). Not all information is private and privacy need not refer to information about a person's body or thoughts, but could also refer to possessions. For example, the fact that a person owns a certain product may be considered private information.

Privacy issues are universal and not limited to the United States. In 1948, the United Nations ratified a worldwide principle of privacy that states no one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or

attacks. Since the United States belongs to the United Nations, one can assume that the U.S. holds these principles to be correct and proper. In addition to universal privacy rights, there are personal data privacy and related data collection and retention issues.

In 2000, the Federal Trade Commission (FTC) published a set of five principles of Fair Information Practices (FIP). The principles include the following: notice, choice, access, security, and enforcement. Under the *notice* principle, organizations must disclose their information practices before collecting personal information. The *choice* principle states that a consumer must be given an option of how personal information is collected and used, especially the use of personal information beyond the purpose for which it was first collected. The *access* principle states that a consumer must be able to look at and change any information that may be inaccurate. Furthermore, this principle supports the fact that a consumer has a right to know what data is being collected about them. The *security* principle states that an organization must take reasonable precautions making sure that collected data is accurate and secure from unauthorized use or intrusion. The *enforcement* principle addresses accountability. These FIP standards are based on international standards initially developed by the Organization for Economic Cooperation and Development (OECD) (Peslak, 2006). Pelak's analysis of 100 international web sites found that the five principles of notice, access, choice, security, and enforcement were not being followed. This is clearly an opportunity for improvement to follow these basic privacy principles.

DATA MINING CHALLENGES

Lee and Siau (2001) listed seven requirements and challenges associated with data mining. They noted that data mining can be a complicated and difficult process. For example, data mining must be able to handle different types of data. Data does not always exist in textual format. Multimedia data, spatial and hypertext data may also be mined, but specific mining techniques must be developed to handle those data types. Currently, most data mining techniques are designed for alphanumeric data only. Secondly, data mining algorithms must be able to handle data in an efficient and scalable manner. Data mining algorithms must be predictable regardless of the size of the dataset. Third, data mining must handle noisy or missing data within a dataset and still be able to produce an accurate representation of the data in the form of a model. There is a significant quality aspect involved and required when attempting to perform data mining activities. Next, end users must be able to perform data mining tasks without having an extensive knowledge of data mining algorithms. In other words, data mining tools should allow the user to explore the data on his or her own, without having to know exactly what he or she is looking for. In fact, much of data mining is exploratory in that the user does not necessarily know exactly what he or she is looking for. However, when data mining results are presented, they should be easily understood (Lee & Siau, 2001).

Data quality is an important aspect of data mining. High quality data that has been prepared specifically for data mining tasks will result in useful data mining models and output. Conversely, low quality data has a significant negative impact on the utility of data mining results. What is meant by the term 'data quality'? One might begin by outlining the characteristics of data that might be of *low* quality. For example, data that is inaccurate, incomplete, insecure, ambiguous, or outdated may be considered low quality. Turban et al. (2005) noted that the cause of many data problems has to do with data being entered or generated improperly. Furthermore, data problems emerge when data is tampered with or gathered

inconsistently (Turban, Aronson, & Liang, 2005). If not resolved, data quality problems can also delay data warehouse implementations and the utility if data mining results. Organizations gather so much data on a daily basis that the question is no longer how to store it, but rather how to gain useful information from large amounts of data.

Incomplete data has a significant impact on data mining results and impairs the data mining algorithms from providing an accurate representation of the underlying data. Incomplete data can occur due to several reasons. First, incomplete data may exist due to partial system failures. Second, incomplete data may exist because data was not entered accurately. Third, some individuals may be unwilling to provide values for specific attributes, potentially due to privacy concerns. Finally, values may not be available at the time the data was entered (Aggarwal & Parthasarathy, 2001). To address the concern of mining very large and incomplete datasets, Aggarwal and Parthasarathy (2001) introduced conceptual reconstruction. Conceptual reconstruction uses a correlation structure of the data and expresses data in terms of conceptual features rather than the dimensionality of the dataset. This approach is a data reconstruction technique where useful data mining results can still be obtained even with incomplete data. Principal component analysis is one such statistical technique used to discover the conceptual features of a dataset.

How does one improve data quality? Winkler (2004) outlined two strategies for improving data quality. Although these methods are not unique to data mining, they can easily be applied to datasets that are used in data mining projects. The first strategy includes imputation methods for inconsistent or missing data. This approach improves data quality by filling in missing data or editing existing data. Editing data in this way is not done in an arbitrary way. Instead, editing data can be done through the use of statistical models. The second strategy recommends a data cleaning approach that locates duplicate records within a dataset. Winkler's strategy proposed the use of a bridging file, which connects two data files together based on a common field (if one is available). The reason for the bridging file is to link two different data sources, in an attempt to improve the overall data quality of all the data. If a common field is not available, automated methods are used to determine how closely records from one file match those in another file. Machine learning methods can be used to determine how closely the records match (Winkler, 2004). Winkler also recommended that file linkage approaches (such as the bridging file) are applicable to data warehousing and may improve data quality within that environment.

One approach to improving data quality is through understanding the semantics of the data and its context. Madnick and Zhu (2006) developed an approach where they considered data quality issues to be data misinterpretation problems. That is, instead of calling the problem a data quality problem, they showed that sometimes the data results can be misinterpreted due to differences in context and meaning of the underlying data. This problem can occur when the user has inadequate knowledge of the underlying data and the methods used to produce aggregate outputs. One example included data collection for IBM's stock price on a given day. However, when several sources were consulted, each source had a different value for that same day. The question then becomes, *which source is correct*? Madnick and Zhu suggest the possibility that all the sources may be correct and that each could be interpreted incorrectly. The Context Interchange (COIN) technology concept shows that semantics and context can be captured in order to further one's understanding of the underlying data, which thereby improves the data quality (Madnick & Zhu, 2006).

An approach to data quality that is markedly different from Madnick and Zhu's approach is the one suggested by Shankaranarayanan and Cai (2006). Shankaranarayanan and Cai (2006) suggest that data quality can be managed in both an objective and context-dependent manner. Their research also recommended that data quality be investigated in terms of information systems output, not the information system itself. That is, the data quality aspects focus on information as a by-product of information systems (Shankaranarayanan & Cai, 2006).

PRIVACY ISSUES

There are four primary approaches for protecting privacy: comprehensive laws, sectoral laws, self-regulation, and individual privacy-enhancing technologies (PETs) (Marcella & Stucki, 2003). Most countries use a combination of the above. The United States typically follows a combination of sectoral laws, self-regulation, and individual PETs. In contrast, the European Union has comprehensive laws that directly address the collection, use, and dissemination of individually identifiable information.

The Health Information Portability and Accountability Act (HIPAA) of 1996 protects individually identifiable health information. HIPAA establishes standards by which the privacy and security of medical health information is to be protected. The Act also requires security mechanisms to be used in the electronic exchange of individually identifiable health information. Its security requirements include technical, administrative, and physical controls. The administrative controls cover personnel and hiring practices. The technical controls cover health information that is collected, processed, stored, and shared via computer systems. Finally, physical controls cover necessary protection against physical access to computer systems or records. (Moskop, et al., 2005)

Title V of the Gramm-Leach-Bliley Act of 1999 (GLB), also known as the Financial Services Modernization Act, has privacy-related requirements for financial services sector. According to Senator Phil Gramm, the way financial services have operated has not changed much since the depression era. The GLB Act brings together securities, banking, and insurance industry. Key privacy-related provisions in GLB include the disclosure of privacy policies regarding information sharing with business affiliates and third parties. Additionally, the Act requires financial services organizations to allow consumers to opt-out of sharing individually identifiable information with nonaffiliated third parties. Financial institutions must also disclose privacy policies at the time that a customer relationship is established. A financial institution must then provide a copy of its privacy policy at least once annually.

The *Personal Information Protection and Electronic Documents Act* (PIPEDA) is Canada's primary legislation that addresses privacy issues. Introduced in 2001, the legislation is designed to protect personal information that is collected, disclosed, or used electronically. It establishes principles to govern collection, disclosure and usage of personal information. These principles include accountability, purpose of collection, consent, disclosure, retention, accuracy, security, individual access to data, and the right for consumers to challenge organizational compliance with the principles. Swartz (2006) investigated the extent to which organizations were complying with this legislation and found that even though 94 percent of the organizations surveyed had privacy policies, many of them failed to fulfill some of the basic requirements as set forth in the act. Swartz also found that about 93 percent of online retailers have used personal information for their own marketing initiatives (Swartz, 2006).

Organizations share data for three primary reasons including conducting business transactions such as electronic data interchange (EDI), operational purposes (e.g. for supply chain management and optimization of business processes), and business intelligence or analysis. Sarathy and Muralidhar (2006) introduced a framework that uses operations research and management science models (statistical and mathematical programming models) for record linkage and data protection for organizations that share data. Their approach focuses on the impact that shared data has on organizational decisions. Sharing data can have significant impact on organizations when the data has high utility (i.e. usefulness in obtaining good results). The richness of data is also useful as it pertains to the quantity and quality of relationships embedded within the shared data. Finally, data sharing has impact when the data has either economic or social benefits that emerge from analyses. Economic benefits may include cost savings and increased market share. For government, sharing data may also have social benefits, which could include reduced crime based on analytical results (Sarathy & Muralidhar, 2006).

The relationship between privacy and trust is important. This is because a consumer's level of trust in an organization may depend on how that organization handles the consumers personal information. For example, in business-to-consumer (B-to-C) electronic commerce (e-Commerce), privacy and trust are important factors in building strong relationships (Eastlick, Lotz, & Warrington, 2006). One of the reasons why consumers may feel that their privacy is at risk is because some marketers may merge data sources, which contain a significant amount of individual information about consumers. Eastlick et al. (2006) found that there was a negative relationship between privacy concerns and trust. Additionally, the study found a negative relationship between privacy concerns and intent to purchase online. In other words, these results suggest that a consumer may decide not to purchase online if he or she believes that personal information may be misused. Eastlick et al. also suggested that organizations develop strategies to address these issues. The use of privacy seals from independent third parties have been shown to be effective in establishing increased trust between consumer and businesses in e-Commerce.

PRIVACY-PRESERVING DATA MINING ISSUES

Agrawal and Srikant (2000) explained that increases in digital data have raised concerns about information privacy on a global basis. This particular research paper is considered the seminal work in PPDM research. Their research laid the foundation for future research that addresses privacy issues within a data mining context. They explain that the Internet has made data collection and data storage much easier, but the potential for misuse has also risen significantly. Data mining results can show models of aggregate data, but the model's accuracy depends on the quality of data. The authors raise the concern that any changes to data affect the accuracy and output of data mining models. Their approach to this problem allows the consumer to provide a perturbed value for sensitive attributes. This allows consumers to participate in the process and hopefully gives the consumer a sense of control over his or her own information. A major drawback of this approach is that output accuracy is lost during data mining activities. However, the authors maintain that small drops in accuracy are an acceptable trade-off for privacy.

A drop in the accuracy of data mining output may not be acceptable for applications where accuracy of results is significantly important. In PPDM research, there are tradeoffs. For example, an increase in privacy preservation will result in lower accuracy in the data mining model. PPDM research attempts to control these drops in accuracy while still preserving

individual privacy at the aggregate level. Early PPDM research suggests that privacy and data accuracy cannot coexist in data mining activities.

Iyengar (2002) demonstrated that data can be transformed in such a way as to protect individual identity. The author's argument that there is a tradeoff between privacy and information loss is generally agreed upon by PPDM researchers. He suggests that random data can replace any individually identifiable information. Another approach is to exclude any sensitive attributes from a data set. In data mining activities, one must transform data before it is useful for data mining. It is during this step that sensitive attributes may be dropped or excluded from the data set before data mining is used. Random data may also replace existing sensitive attributes during this step. Metrics can also be used to quantify the loss of content within a dataset (Iyengar, 2002). Iyengar's approach to solving the information loss problem is by using a genetic algorithm against the popular *adult* data set available from the University of California Irvine's (UCI) machine learning library, available at http://www.ics.uci.edu/~mlearn/MLRepository.html. As noted in Iyengar's research, this benchmark data set has been used in other machine learning experiments.

Thuraisingham (2002) first suggested that privacy issues occur in data mining and that this is a generalization of the inference problem. The inference problem refers to an issue when a user can infer new knowledge by executing successive queries against a database. Since data mining techniques are designed to help the user discover new knowledge, the results of data mining can raise the likelihood of an inference to occur. Thuraisingham also noted that this may cause ethical issues based on how the information is going to be used (Thuraisingham, 2002).

Vaidya and Clifton (2004) stated that even though there are privacy implications in data mining, organizations should continue to use data mining because there are numerous benefits. They also put forth the idea that there are tradeoffs between the utility of data mining results and maintaining informational privacy of consumers. In other words, when privacy data is present, the utility, or usefulness, of the data mining results will decrease. Therefore, one of the goals of PPDM research is to protect the results of a data mining operation from inference (Vaidya & Clifton, 2004).

The PPDM research area has produced solutions based on different approaches. For example, approaches such as data distribution, data modification, data mining algorithms, data or rule hiding, and privacy preservation techniques have been proposed (Verykios et al., 2004). Privacy preservation techniques are an important solution approach because different techniques can be used to modify data, such as heuristics, cryptography, and randomization. Verykios et al. (2004) also noted that selectively modifying data for privacy preservation is a complex problem. Other methods for modifying data include data sampling, swapping records, aggregation, perturbed values or noise addition, and blocking.

There is a strong relationship between security and PPDM. Several solution approaches to PPDM use concepts from the information security field. The first approach is through the use of secure multiparty computation (SMC). SMC is the general problem of secure computation of a function with multiple distributed inputs (Vaidya, et al., 2006). SMC-based solutions use cryptography as a major portion of the algorithm. For example, two-part secure computation (Yao, 1986) and multiparty secure computation (Goldreich, Micali, & Wigderson, 1987) have been used in a variety of cryptographic and privacy-preserving data mining studies.

Narayanan and Shmatikov (2005) demonstrated that data can be encrypted in such a way that users can still use the information contained within it (Narayanan & Shmatikov, 2005). Their study used provably secure obfuscation techniques while permitting certain types of queries to be

generated. A limitation to their study was that they only examined its use on small databases, therefore the approach may not scale well to larger databases. In order for their approach to work, they developed a new query language. Their approach may also be impractical if a user wanted to use widely available databases such as Microsoft SQL Server or Oracle. Pinkas (2002) also discussed cryptographic techniques for preserving privacy within a data mining context. Pinkas begins with the use of oblivious transfer, which is used to construct secure computation functions.

Randomization techniques have been used as one method for preserving privacy during data mining activities (Dinur & Nissim, 2003; Du & Zhan, 2003). Du and Zhan (2003) proposed a randomized response technique to perturb data so that users cannot tell whether the data contains truthful information or false information. They used a decision-tree classifier along with randomization methods to perturb the data so that aggregate results still show some degree of accuracy, while at the same time maintain individual privacy. Once this deliberate noise addition occurs within the data warehouse, data mining results have reduced accuracy, but may be within a predefined tolerance level. One drawback with this approach is that it only focused on Boolean data types to test their technique. Future research could focus on other frequently used data types, especially plain text and multimedia objects. Randomization techniques ought to define what the tolerance level is for loss of accuracy in a data mining result. Du and Zhan also neglected to define exactly what tolerances are acceptable during data mining with privacy-preservation. However, Du and Zhan, along with other researchers have noted that small drops in accuracy are acceptable (Dinur & Nissim, 2003; Du & Zhan, 2003; Dutta et al., 2003; Evfimievski et al., 2003).

How does missing or perturbed data affect data mining? Brown and Kros (2003) concur with others that the accuracy of data mining models is based on the quality of underlying data. It is important or data mining users to understand how underlying data affects the model. Data mining users must address inaccurate and missing data issues before applying data mining algorithms to a dataset (Brown & Kros, 2003). There are different kinds of missing data, such as data missing at random. One could also decide to treat outliers as missing data, or simply use complete records only. Missing and perturbed data impact data mining approaches including clustering, association rule mining, decision trees, neural networks, and k-nearest neighbor (kNN) (Brown & Kros, 2003). In fact, any missing or perturbed data will compromise the utility of any data mining output.

When an algorithm is developed to protect privacy within data mining, it must be developed for a specific data mining task. That is, each privacy-preservation algorithm is specific to its associated data mining task. For example, an association rule mining algorithm developed by Liu et al. (2006) combines hashing and cryptography to produce association rules that also have privacy preservation properties. The association rules produced with this algorithm therefore provide no new information about a specific individual, and also ensures that individual privacy is maintained by reducing the possibility of any inferences from occurring (Liu, Piao, & Huang, 2006).

Yang et al. (2005) proposed a cryptographic approach for preserving privacy during the data collection process. This approach focuses largely on the actions that are taken prior to executing specific data mining tasks. Yang et al. suggest that cryptographic approaches to privacy preserving data mining are superior to random perturbation and randomized response techniques because the former maintain data accuracy and provide adequate data privacy (Yang, et al., 2005). Their approach suggests that data miners collect data anonymously. That is, data is

collected so that the data miner has no knowledge of who provided a particular piece of data. If the data miner has no way of knowing who provided which data, he or she need not worry about protecting individual privacy. Design of a specific cryptographic approach to protecting individual privacy must be done in the context of a specific data mining task. Their approach to this problem developed protocols for protecting anonymity, including one that can handle malicious data mining and malicious respondents.

There have been a variety of solution approaches to PPDM. This is because each approach focuses on a different meaning of privacy, what results are desired, and how data is distributed. In response to this, Clifton et al. (2002) recommended that a toolkit of components be developed that draws different components together for solving privacy-preserving data mining problems. Their work extends PPDM by focusing on the distributed aspects of data mining because data often exists in multiple locations. In distributed data mining, two or more sites can share global mining results without learning anything about data at an individual site (Clifton, et al., 2002).

SPECIFIC RECOMMENDATIONS

After careful examination and analysis of the above literature, there are a variety of issues to address prior to widespread adoption.

1. Those implementing privacy-preserving data mining solutions must first address internal privacy-related policies by investigating the following.
    a. Is there an existing privacy policy?
    b. Who is responsible for that policy?
    c. To whom does the policy refer?
    d. Does the policy address legal and ethical concerns?
    e. Who is responsible for executing procedures related to the policy?
    f. Managers will need to make policy decisions about whether data should be encrypted and at what level of granularity.

2. Data mining resources and processes must be defined.
    a. There may be a significant barrier to adoption for SMEs due to the lack of data mining and privacy expertise among staff.
    b. What are the costs and benefits of moving forward with a PPDM strategy?
    c. What does the overall PPDM architecture look like?
    d. What process(es) are going to be used to ensure privacy is preserved throughout the entire data mining life cycle (DMLC)?
    e. It should be possible to integrate privacy-preservation processes into the standard data mining process known as CRISP-DM (CRoss Industry Standard Process for Data Mining).

3. PPDM algorithms must be integrated into existing systems prior to adoption.
    a. Add-on software modules for relational databases could be developed and integrated into existing systems.
    b. Relational database vendors could include PPDM algorithms in its core software.
    c. PPDM algorithms could become a core feature of data mining packages.

d. Algorithms must become more generalized. As of this writing, algorithms are designed to solve one specific task. Since there are so many different data mining tasks, it would be beneficial if PPDM techniques were able to be applied to a variety of data mining tasks.

CONCLUSION AND FUTURE WORK

While there are many barriers to implementing privacy-preserving data mining (PPDM) techniques, there are almost too many issues that must be addressed prior to implementation. It also appears that there are significant policy, process, and technological issues that must be addressed. For SMEs, adoption of PPDM technologies may be prohibitively expensive. Furthermore, SMEs usually do not have on-site staff with expertise in privacy, data mining, and privacy-preservation techniques. Large organizations do have the capacity to engage in PPDM techniques, especially large financial institutions and medical institutions where meeting privacy legislation requirements are extremely important. The good news is that some data mining software packages are free and easy to use. Packages such as Weka and Orange are freely available, have online tutorials, and are open source so one can develop their own algorithms if necessary. Orange is especially easy to use since it is written in Python and many companies have an IT staff person who knows Python. If organizations concentrate largely on getting the PPDM process and related policy developed and organized correctly, the technical implementation should not be prohibitively difficult. Weka is available at: http://www.cs.waikato.ac.nz/ml/weka/. Orange is available at: http://orange.biolab.si/. There are also commercial data mining packages, but they generally do not have privacy-preserving tools in them (yet). The best recommendation is to run PPDM pilot studies using low-cost or free data mining tools before investing in large scale PPDM solutions.

Future work in this area will integrate some basic PPDM algorithms into the Orange or Weka data mining toolset and run several experiments using data sets from the UCI machine learning library. Thankfully, there are a large number of data sets available for experimentation in this area. Furthermore, the Technology Acceptance Model (TAM) could be used to evaluate PPDM adoption within organizations.

REFERENCES

Abelson, J. (2007, January 25, 2007). TJX breach snares over 200,000 cards in region, *The Boston Globe*. Retrieved from http://www.boston.com/business/globe/articles/2007/01/25/tjx_breach_snares_over_200000_cards_in_region/

Aggarwal, C., & Parthasarathy, S. (2001). Mining massively incomplete data sets by conceptual reconstruction. *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 227-232.

Aggarwal, C., Pei, J., & Zhang, B. (2006). On Privacy Preservation against Adversarial Data Mining. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 510-516.

Brown, M., & Kros, J. (2003). Data mining and the impact of missing data. *Industrial Management & Data Systems, 103*(8), 611-621.

Children's Online Privacy Protection Act, 15 U.S.C. 6501-6508 § 1301-1308 (1998).

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. (2002). Tools for privacy-preserving distributed data mining. *ACM SIGKDD Explorations Newsletter, 4*(2), 28-34.

Dinur, I., & Nissim, K. (2003). Revealing information while preserving privacy. *Proceedings of the 2003 ACM SIGMOD Symposium on Principles of Database Systems*, 202-210.

Du, W., & Zhan, Z. (2003). Using randomized response techniques for privacy-preserving data mining. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 505-510.

Eastlick, M., Lotz, S., & Warrington, P. (2006). Understanding online B-to-C relationships: An integrated model of privacy concerns, trust, and commitment. *Journal of Business Research, 59*, 877-886.

Goldreich, O., Micali, S., & Wigderson, A. (1987). *How to play any mental game - a completeness theorem for protocols with honest majority.* Paper presented at the 19th ACM Symposium of the Theory of Computing.

Han, P., & Maclaurin, A. (2002). Do consumers really care about online privacy? *Marketing Management, 11*(1), 35-38.

Iyengar, V. (2002). Transforming data to satisfy privacy constraints. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 279-288.

Lee, S. J., & Siau, K. (2001). A review of data mining techniques. *Industrial Management & Data Systems, 101*(1), 41-46.

Liu, J., Piao, X., & Huang, S. (2006). A privacy-preserving mining algorithm of association rules in distributed databases. *Proceedings of the 1st International Multi-Symposium on Computer and Computational Sciences (IMSCCS)*, 740-750.

Madnick, S., & Zhu, H. (2006). Improving data quality through effective use of data semantics. *Data & Knowledge Engineering, 59*, 460-475.

Marcella, A., & Stucki, C. (2003). *Privacy handbook: Guidelines, exposures, policy implementation, and international issues*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Moskop, J., Marco, C., Larkin, G. L., Geiderman, J., & Derse, A. (2005). From Hippocrates to HIPAA: Privacy and Confidentiality in Emergency Medicine - Part I: Conceptual, Moral, and Legal Foundations. *Annals of Emergency Medicine, 45*(1), 53-59.

Narayanan, A., & Shmatikov, V. (2005). *Obfuscated databases and group privacy*. Paper presented at the Proceedings of the 12th ACM Conference on Computer and Communications Security, Alexandria, VA.

Persson, A., & Hansson, S. (2003). Privacy at work ethical criteria. *Journal of Business Ethics, 42*(1), 59-70.

Peslak, A. (2006). Internet Privacy Policies of the Largest International Companies. *Journal of Electronic Commerce in Organizations, 4*(3), 46-62.

Sarathy, R., & Muralidhar, K. (2006). Secure and useful data sharing. *Decision Support Systems, 42*, 204-220.

Shankaranarayanan, G., & Cai, Y. (2006). Supporting data quality management in decision-making. *Decision Support Systems, 42*, 302-317.

Swartz, N. (2006). No Respect for PIPEDA. *Information Management Journal, 40*(5), 21.

Thuraisingham, B. (2002). Data mining, national security, privacy and civil liberties. *ACM SIGKDD Explorations Newsletter, 4*(2), 1-5.

Torres, E. (2007). Man arrested in theft of 1.8 million Social Security numbers. *The Orange County Register*. Retrieved from http://www.ocregister.com/news/kim-numbers-affairs-1924451-security-social#

Turban, E., Aronson, J., & Liang, T.-P. (2005). *Decision support systems and intelligent systems* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Unknown. (2007). Children's patient info stolen from Edmonton hospital. *CBC News*. Retrieved from http://www.cbc.ca/canada/edmonton/story/2007/11/13/glenrose-breach.html

Vaidya, J., & Clifton, C. (2004). Privacy-preserving data mining: Why, how, and when. *IEEE Security & Privacy, 2*(6), 19-27.

Vaidya, J., Clifton, C., & Zhu, M. (2006). *Privacy Preserving Data Mining*. New York: Springer.

Verykios, V., Bertino, E., Fovino, I., Provenza, L., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record, 33*(1), 50-57.

Winkler, W. (2004). Methods for evaluating and creating data quality. *Information Systems, 29*, 531-550.

Xiong, L., Chitti, S., & Liu, L. (2007). Preserving data privacy in outsourcing data aggregation services. *ACM Transactions on Internet Technology (TOIT), 7*(3).

Yang, Z., Zhong, S., & Wright, R., N. . (2005). *Anonymity-preserving data collection*. Paper presented at the Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, Chicago, Illinois, USA.

Yao, A. C. (1986). How to generate and exchange secrets. *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science*, 162-167.

Zhu, M., & Liu, L. (2004). Optimal randomization for privacy preserving data mining. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 761-766.