Is Privacy Still an Issue for Data Mining?

(Extended Abstract)

Chris Clifton Wei Jiang Mummoorthy Muruguesan M. Ercan Nergiz
Department of Computer Science
Purdue University
305 North University Street
West Lafayette, IN 47907-2107
{clifton, wjiang, mmuruges, mnergiz}@cs.purdue.edu

Abstract

Privacy-preserving data mining has been the subject of substantial research. This paper summarizes accomplishments, the privacy debate, and outlines areas where privacy issues still impact data mining research and practice.

1. Introduction

Five years ago, the National Science Foundation held a workshop on “Next Generation Data Mining”. At that time, privacy was a relatively new issue to the data mining community; there had been half a dozen research papers on privacy-preserving data mining techniques [3, 19, 2, 22, 15, 31, 26] and even a couple of articles in the popular press[9, 21]. The ensuing years have seen substantial research in privacy-preserving data mining techniques, several workshops on the subject. At the same time, “data mining” has been vilified as a threat to privacy and civil liberties – witness the 2003 letter from the USACM suggesting that data mining technology could contribute to the growth of privacy-compromising databases[29] (and the ensuing response from SIGKDD[17]), and perhaps more critically, the continuing efforts to restrict “data mining” in the U.S. Senate (ranging from the proposed Data Mining Moratorium Act of 2003 [10], which would have banned data-mining by the Department of Defense, to this year’s “Data Mining Reporting Act of 2007” that would require a report to Congress from any Federal Government department or agency engaged in or developing data mining activities [11].)

What has been the impact of privacy-preserving data mining research over the last five years? In commercial terms, the answer is little or none – privacy-preserving data mining technology is still in the research paper, or at best research prototype, stage. However, the research may have had an impact on the “privacy vs. data mining” debate; researchers have pointed out the privacy implications of data mining technology, and the debate had become more reasoned. For example, the Moratorium Act of 2003 banned “data-mining”, with exceptions for “computer searches of public information” or “computer searches that are based on a particularized suspicion of an individual”. By the 2007 act, the term “data mining” had been limited to pattern-based “queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals”. This is much more specific than the research community’s view of data mining, and shows recognition that data mining technology is not inherently bad, but is (perhaps unusually) subject to misuse.

Where does this leave privacy-preserving data mining research? While it could be argued that the direction of the debate makes such research irrelevant, an alternative view is that the debate has lead to a better framework for research in the next five years. In particular, the following have emerged from the debate, and can serve as guidance for privacy research in the data mining community:

- Misuse of data doesn’t require data mining;
- Misunderstanding data mining technology can lead to misuse; and
- Privacy is about individually identifiable data.

The following sections will elaborate on these points, with suggestions on how research can address them as well as pointers to successes.

2 Misuse of data doesn’t require data mining

Most high-profile cases of misuse of private data appear to have nothing to do with data mining. Instead, it is problems with security of the database that lead to security breach and misuse. The USACM letter questioning the Total Information Awareness program recognized this [29]; it is unfortunate that the term “data mining” was featured so prominently, as the security risks described by the letter were based on potential for misuse of the immense databases proposed by the program, rather than the technology to analyze them. Identity theft is an aggravating and expensive problem, but results from direct disclosure of information about individuals (the underlying database) rather than analysis of that data. Most high profile privacy breaches are similar; it is poor security of the database (contained on laptops, backup tapes, or through electronic break-in) that leads to the breach.

Does this mean data mining can be exonerated? Unfortunately, the answer is no. One of the highest profile cases was the 2005 theft of credit card information from CardSystems (a credit card transaction processing company)[25]. A breach of such magnitude should not have been able to happen; CardSystems was only supposed to use the data to process the transaction, not store it. However, CardSystems stored data on some transactions “for research purposes in order to determine why these transactions did not successfully complete”[25]. Without data mining technology, meaningful analysis of such a large amount of data (at least 263,000 records were stolen) would have been difficult or impossible. Without data mining, there would have been no reason to keep the data, and thus nothing to steal.

Most privacy-preserving data mining research to date can be used to address this problem. Much of the work falls in two basic categories, exemplified by two papers titled “Privacy Preserving Data Mining” that appeared in 2000 [3, 19]. In [3], data was distorted before placing it in the database, obscuring actual values. Privacy-preserving data mining techniques on such data recover the correct data mining *results*, based on the data and knowledge of the process by which it was distorted, but recovery of actual data values (even knowing the distortion process) is (presumably) impossible. If CardSystems had used such a technique to save *distorted* transactions, theft of the data would have had no (privacy) impact.

The second approach, exemplified by [19], is to mine data from distributed sources without requiring the sources to disclose the data (even to each other.) Such approaches could alleviate the need for the immense databases that raised concerns with the total information awareness program. Techniques have been developed that replicate the results of several data mining algorithms, while allowing data about an individual to be split among several sites (starting with [31]); gathering enough information about an individual to result in a serious invasion of privacy would require compromising several databases.

There have been techniques developed to support many types of data mining in these two approaches, for a more detailed discussion and citations to much of the research see [32]. However, work is not done: additive random noise must be used carefully, as in some cases (e.g., attempting to mask correlated data items) signal processing techniques can be used to recover original values with relatively high accuracy [16]. Multiplicative randomization techniques can help this problem [20], but further investigation is needed.

The Secure Multiparty Computation approach also has weaknesses. Much of what has been published has only been proven secure in the semi-honest model; the assumption that parties will not “cheat” to try to obtain private information is not sufficient for many practical applications. Methods proven secure under the malicious model have not yet shown the efficiency needed for practical applications. Intermediate approaches such as *accountable computing*[14] and *non-cooperative computation*[28] have been proposed; development of protocols under these models is needed.

The key to acceptances of this technology as a viable alternative to the monolithic (and vulnerable) data warehouse is to develop tools that make business sense. Two key possibilities are:

Enhancing user trust to get better data: Studies suggest that reputation and ability to protect privacy result in a greater willingness to provide accurate personal data[18]. Businesses that use privacy-preserving data mining technologies, and can convince individuals of the efficacy of that technology, stand to improve the value of their data.

Corporate collaboration using sensitive business data: Privacy-preserving data mining technology can be used to protect sensitive data that is not concerned with individual privacy. Companies may need to keep data secret from collaborators, but still wish to use shared data for common analysis purposes. This is commonly done using a trusted broker to manage information, but is such trust necessary (or cost-effective)? One area where this has been investigated (and found corporate interest) is in supply chain management [5], other areas surely exist where a business case can be made for development and use of privacy-preserving data analysis.

The goal of research on privacy-preserving data mining techniques needs to go beyond developing the basic techniques. The future lies with developing technology that ties to a business model where privacy is a demonstrable asset.

3 Misunderstanding data mining technology can lead to misuse

A key component of the Data Mining Reporting Act of 2007 is a requirement that agencies give "An assessment of the efficacy or likely efficacy of the data mining activity in providing accurate information consistent with and valuable to the stated goals and plans for the use or development of the data mining activity." Predictive data mining techniques typically give at best a probability that the given prediction is correct. While this is of some value, it is not sufficient. For example, "X is a member of a terrorist organization with 50% probability" is much less actionable than saying "X is a member of a terrorist organization with 50% probability, and a dangerous crackpot with 50% probability" vs. "X is a member of a terrorist organization with 50% probability, and a Senator with 50% probability".

The first step is understanding what can be expected of data mining technology. Classical work on limits of learning and sample complexity generally talks of the expected accuracy of a prediction AND the confidence that the expected accuracy will be met. Work such as that of [33] can form a much better basis for justifying both the value and privacy implications of data mining in privacy-sensitive situations. Methodologies to apply such theory in the early stages of a data mining project are needed; only when we can justify the value and quantify the privacy risk should we begin the process of collecting data.

While most concerns focus on the potential inaccuracy of data mining, the converse is also important to privacy. Highly accurate predictive modeling can be a threat to privacy, if the prediction is with respect to sensitive information. The parameters in this case are somewhat different, it isn't average accuracy, but the expected accuracy of a PARTICULAR prediction that matters. It is little comfort to say that on average sensitive data will have low accuracy if you are one of the individuals for whom sensitive data can be predicted. While some work has been done that can be applied to this problem [7], more is needed. A key component is understanding when privacy is at risk, leaded to our next topic.

4 Privacy is about individually identifiable data

In order to find the common ground between privacy and data mining, one of the key questions to ask is if any kind of information about a *population* is subject to privacy concerns. Fortunately, the answer to this question seems to be no. The European Community Directive 95/46/EC [8] protects only "personal data" that can tied to (individual) persons. Similarly The United States Healthcare Information Portability and Accountability Act (HIPAA) privacy rules [13] state that the privacy standards apply to "individually identifiable health information"; other information, including *de-identified* personal data, is not subject to privacy regulations. De-identification can protect privacy while allowing use of (personal) data.

However recent research showed that simply removing unique identifiers (SSN, name, $\cdots$) from data is not sufficient; external information can be used to *re-identify* data by linking information to individuals. In US, the combination of zip code, and birth date is unique for 87% of the citizens [30]. Sweeney et al. showed that they could re-identify supposedly anonymous health records via linking them to a publicly available voter registration list. Anonymization techniques such as k-anonymity [27] have been proposed to address this issue by restricting linkage to groups of people. However, such techniques do not provide a statistical way of reasoning about the amount of disclosure inherent in the size of the groups. The intent of HIPAA safe harbor rules is stated as "providing a means to produce some de-identified information that could be used for many purposes with a *very small* risk of privacy violation". But how much small is *very small*? Implying that 99.9% of the patients in a given hospital are diabetics certainly exposes private information even though the group is relatively large.

A better approach is to work with the risk of reidentification. Work in [24] addresses this issue by bounding the probability of a given person being in a private dataset so that risk of identification can be controlled; it is interesting to note that there is no correspondence between this risk-based measure and a choice for k.

Still the risk of identification is not well studied in the literature. It is certain that the risk is very dependent on the prior knowledge of adversaries and may be different for each individual in the data. (E.g., risk of identification is different for a young and old person when we identify the person as being a diabetics with 21% probability. The reason is that the probability of being a diabetic is publicly known to be 9.6% for a young person. This probability increases to 20.9% for an old person.[23]) A second issue is to evaluate the cost of disclosure on an individual basis. The *real* cost (in terms of factors like economics, sociology, $\cdots$) of being identified or in other words, the *risk from identification* should be studied. Finally, the trade off between the risk and benefit (e.g., from any data mining operation on de-identified data) should be taken into consideration.

The problem only gets more difficult when we consider social networks, personalization of services, and other services based on aggregated data. These do not require that private data of individuals be maintained in a central database, but the supposedly anonymized aggregate data is not necessarily free of privacy concerns. The AOL query log release incident[4] is an example of how such aggregate data could affect a user's privacy. By doing simple analysis on the (supposedly) anonymized AOL data, a New York Times reporter was able to identify the owner of a set of queries[6]. Recent works have come up with anonymization techniques for query logs [1]. This can be applied on the query log data before being made available for mining purposes. Similarly, users of location-based services leave traces of information that could be combined over time to create user profiles. Consistent usage of location-based services help creating a pattern of movements, and identify the person associated with it [12]. Data mining techniques such as clustering and frequent pattern mining on these aggregate data produce patterns, which in turn, help in identifying the individuals associated with the pattern. Could we use this to avoid releasing potentially identifiable data?

In addition to the open issue of understanding when data is individually identifiable (and thus subject to privacy laws), recognizing that the goal of privacy is to protect individually identifiable data (and not necessarily anything more) could open the door for more efficient privacy-preserving data mining techniques. For instance, combining k-anonymity and homomorphic encryption could lead to more efficient solutions than currently envisioned by secure multiparty computation-based techniques. One could use k-anonymous data as an index to a small group of similar data items, with the non-anonymous data encrypted. Algebraic operations could be applied to the (homomorphically encrypted) data within the group instead of the whole data set.

5 Conclusions

Privacy-preserving data mining still has room to grow, but needs to become focused to have impact. As the privacy debate is growing more reasoned, the need for privacy is becoming more clear and succinct. Researchers must look to the privacy debate to ensure that privacy-preserving data mining research meets real privacy needs. This is both a challenge and an opportunity; while much has been done, the new problems that are arising are even greater.

References

[1] E. Adar. User 4xxxxx9: Anonymizing query logs. May 2007.

[2] D. Agrawal and C. C. Aggarwal. 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*, pages 247–255, Santa Barbara, California, May 21-23 2001. ACM.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*, pages 439–450, Dallas, TX, May 14-19 2000. ACM.

[4] M. Arrington. Aol proudly releases massive amounts of private data.

[5] M. J. Atallah, H. G. Elmongui, V. Deshpande, and L. B. Schwarz. Secure supply-chain protocols. In *IEEE International Conference on E-Commerce*, pages 293–302, Newport Beach, California, June 24-27 2003.

[6] M. Barbaro and T. Z. Jr. A face is exposed for aol searcher no. 4417749, Aug. 9 2006.

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

[8] 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, Oct. 24 1995.

[9] A. Eisenberg. With false numbers, data crunchers try to mine the truth. *New York Times*, July 18 2002.

[10] M. Feingold, M. Corzine, M. Wyden, and M. Nelson. Data Mining Moratorium Act of 2003. U.S. Senate Bill (proposed), Jan. 16 2003.

[11] M. Feingold, M. Sununu, M. Leahy, M. Akaka, M. Kennedy, M. Cardin, M. Feinstein, and M. Whitehouse. Federal Agency Data Mining Reporting Act of 2007. U.S. Senate Bill (Introduced), Jan. 10 2007.

[12] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of First ACM/USENIX International Conference on Mobile Systems, Applications, and Services (MobiSys)*, May 2003.

[13] Standard for privacy of individually identifiable health information. *Federal Register*, 67(157):53181–53273, Aug. 14 2002.

[14] W. Jiang, C. Clifton, and M. Kantarcioglu. Transforming semi-honest protocols to ensure accountability. *Data and Knowledge Engineering*, To appear.

[15] M. Kantarcıoğlu and C. Clifton. 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)*, pages 24–31, Madison, Wisconsin, June 2 2002.

[16] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. Random data perturbation techniques and privacy preserving data mining. *Knowledge and Information Systems*, 7(4):387–414, May 2004.

[17] W. Kim, R. Agrawal, C. Faloutsos, U. Fayyad, J. Han, G. Piatetsky-Shaprio, D. Pregibon, and R. Uthurasamy. "data mining" is not against civil liberties. Open Letter from the Directors of the Executive Committee of ACM SIGKDD, July 28 2003.

[18] A. Kobsa. Privacy-enhanced personalization. *Communications of the ACM*, 50:24–33, Aug. 2007.

[19] Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Advances in Cryptology – CRYPTO 2000*, pages 36–54. Springer-Verlag, Aug. 20-24 2000.

[20] K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):92–106, Jan. 2006.

[21] S. Lohr. Online industry seizes the initiative on privacy. *The New York Times on the web*, Oct. 11 1999.

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

[23] National Institute of Diabetes and Digestive and Kidney Diseases. National diabetes statistics fact sheet: general information and national estimates on diabetes in the United States. Technical Report NIH Publication No. 06–3892, U.S. Department of Health and Human Services, National Institute of Health, Bethesda, MD, Nov. 2005.

[24] M. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *2007 ACM SIGMOD International Conference on Management of Data*, Beijing, China, June 11-14 2007.

[25] J. M. Perry. Statement of john m. perry, president and ceo, cardsystems solutions, inc. before the united states house of representatives subcommittee on oversight and investigations of the committee on financial services. `http://financialservices.house.gov/media/pdf/072105jmp.pdf`, July 21 2005.

[26] S. J. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In *Proceedings of 28th International Conference on Very Large Data Bases*, pages 682–693, Hong Kong, Aug. 20-23 2002. VLDB.

[27] P. Samarati. Protecting respondent's privacy in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, Nov./Dec. 2001.

[28] Y. Shoham and M. Tennenholtz. Non-cooperative computation: boolean functions with correctness and exclusivity. *Theor. Comput. Sci.*, 343(1-2):97–113, 2005.

[29] B. Simons and E. H. Spafford. Letter from the usacm to the senate committee on armed services, Jan. 23 2003.

[30] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, (5):557–570, 2002.

[31] J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 639–644, Edmonton, Alberta, Canada, July 23-26 2002.

[32] J. Vaidya, C. Clifton, and M. Zhu. *Privacy Preserving Data Mining*, volume 19 of *Advances in Information Security*. Springer, 2006.

[33] V. N. Vapnik. *Estimation of dependences based on empirical data*. Springer-Verlag, New York, 1982.