Data Mining, National Security, Privacy and Civil Liberties

Bhavani Thuraisingham
The National Science Foundation
Arlington, VA
(On Leave from the MITRE Corporation,
Bedford, MA)
bthurais@nsf.gov

ABSTRACT

In this paper, we describe the threats to privacy that can occur through data mining and then view the privacy problem as a variation of the inference problem in databases.

Keywords

Data Mining, National Security, Counter-terrorism, Inference Problem, Security Constraints, Privacy, Privacy Problem, Privacy Constraints, Privacy Sensitive Data Mining, World Wide Web, Civil Liberties

1. INTRODUCTION

There has been much interest recently on using data mining for counter-terrorism applications. For example, data mining can be used to detect unusual patterns, terrorist activities and fraudulent behavior. While all of these applications of data mining can benefit humans and save lives, there is also a negative side to this technology, since it could be a threat to the privacy of individuals. This is because data mining tools are available on the web or otherwise and even naïve users can apply these tools to extract information from the data stored in various databases and files and consequently violate the privacy of the individuals. As we have stressed in [THUR03], to carry out effective data mining and extract useful information for counter-terrorism and national security, we need to gather all kinds of information about individuals. However, this information could be a threat to the individuals' privacy and civil liberties.

Privacy is getting more attention partly because of counter-terrorism and national security. We started giving talks on privacy due to data mining back in 1996 [THUR96] and wrote about data mining and privacy in [THUR98]. While this work received some attention, the topic did not get the widespread attention it is receiving today. Recently we have heard a lot about national security vs. privacy in newspapers, magazines and television talk shows. This is mainly due to the fact that people are now realizing that to handle terrorism, the government may need to collect information about individuals. This is causing a major concern with various civil liberties unions.

We are beginning to realize that many of the techniques that were developed for the past two decades or so on the inference problem can now be used to handle privacy. One of the challenges to securing databases is the inference problem. Inference is the process of users posing queries and deducing unauthorized information from the legitimate responses that they receive. This problem has been discussed quite a lot over the past two decades. However, data mining makes this problem worse. Users now have sophisticated tools that they can use to get data and deduce patterns that could be sensitive. Without these data mining tools, users would have to be fairly sophisticated in their reasoning to be able to deduce information from posing queries to the databases. That is, data mining tools make the inference problem quite dangerous. While the inference problem mainly deals with secrecy and confidentiality we are beginning to see many parallels between the inference problem and what we now call the privacy problem.

This paper discusses both the inference problems through data mining as well as privacy issues. In Section 2, we first provide an overview of the inference problem to give the reader some background. In Section 3, we discuss approaches to handling the inference problem that arises through data mining. In Section 4, we discuss privacy issues. These privacy issues also depend on the various policies and procedures enforced. That is, technical, political, as well as social issues play a role here. Then in Section 5 we revisit the inference problem with respect to privacy. Privacy enhanced/sensitive data mining will be discussed in Section 6. Finally in section 7, we provide an overview of civil liberties vs. national security. The paper is summarized in section 8.

2 BACKGROUND ON THE INFERENCE PROBLEM

Inference is the process of posing queries and deducing unauthorized information from the legitimate responses received. For example, the names and salaries of individuals may be unclassified individually, but while taken together they are classified. This means that one could retrieve names and employee numbers, and then later retrieve the salaries and employee numbers, and make the associations between names and salaries. The problem that occurs through this inference is called the inference problem.

In the early 1970s, much of the work on the inference problem was on statistical databases. Organizations such as the census bureau were interested in this problem. However, in the mid 1970s and then in the 1980s, the United States Department of Defense started an active research program on multilevel secure databases, and research on the inference problem (see, for example, [AFSB83]) was conducted as part of this effort. The pioneers included Morgenstern [MORG87], Thuraisingham [THUR87], and Hinke [HINK88].

We have conducted extensive research on this subject and worked on various aspects. In particular, it was shown that the general inference problem was unsolvable by Thuraisingham [THUR90], and then approaches were developed to handle various types of inferences. These approaches included those based on security constraints as well as those based on conceptual structures (see for example [THUR91a], and [THUR93]). These approaches handled the inference problem during database design, query, and update

operations. Furthermore, logic-based approaches were also developed to handle the inference problem (see, for example, [THUR91b]).

Much of the earlier research on the inference problem did not take data mining into consideration. With data mining, users now have tools to make deductions and extract patterns, which could be sensitive. In the next section we address inference problem and data mining. We also include some information on data warehousing and inference.

3. DATA MINING, WAREHOUSING, AND INFERENCE

First let us give a motivating example where data mining tools are applied to cause security problems. Consider a user who has the ability to apply data mining tools. This user can pose various queries and infer sensitive hypotheses. That is, the inference problem occurs via data mining. There are various ways to handle this problem. Given a database and a particular data-mining tool, one can apply the tool to see if sensitive information can be deduced from the unclassified information legitimately obtained. If so, then there is an inference problem. There are some issues with this approach. One is that we are applying only one tool. In reality, the user may have several tools available to him. Furthermore, it is impossible to cover all ways that the inference problem could occur. Some of the security implications are discussed in the paper by Clifton and Marks [CLIF96]. This is one of the first papers to discuss privacy and data mining.

Another solution to the inference problem is to build an inference controller that can detect the motives of the user and prevent the inference problem from occurring. Such an inference controller lies between the data mining tool and the data source or database, possibly managed by a database system. Discussions of this inference controller approach is given in [THUR96] and [THUR98].

Clifton [CLIF00] has also conducted some theoretical studies on handling the inference problems that arise through data mining. Clifton's approach is the following. If it is possible to cause doubts in the mind of the adversary that his data mining tool is not a good one, then he will not have confidence in the results. For example, if the classifier built is not a good one for data mining through classification, then we cannot have sufficient confidence in the rules that are generated. Therefore, the data mining results may not be useful. Now what are the challenges in making this happen? That is, how can we ensure that the adversary will not have enough confidence in the results? One of the ways is to give only samples of the data to the adversary so that one cannot build a good classifier from these samples. The question then is what should the sample be? Clifton has used classification theory to determine the limits of what can be given. There have been some concerns about this approach, as one could give multiple samples to different groups, and the groups can work together in building a good classifier. But the answer to this is that one needs to keep track of what information is to be given out. At the keynote address I gave on data mining and security at the Pacific Asia Data Mining Conference in Melbourne Australia in April 1998, it was suggested that the only way to handle the inference problem is not to give out any samples. But this could mean denial of service. That is, data could be withheld when it is definitely safe to give out the data.

Next, let us consider data warehousing and inference. The inference problem becomes an issue here also. For example, the warehouse may store average salaries. A user may have access to the average salaries in the warehouse. From these average salaries, the user may infer the individual salaries stored in the data sources from which the warehouse may be built. These individual salaries may be sensitive. We pointed this out in [THUR96]. To date, little work has been reported on data warehousing and the inference problem. This is an area that needs much research.

4. PRIVACY ISSUES

At the IFIP (International Federation for Information Processing) working conference in database security in 1997 at Lake Tahoe, the group began discussions on privacy issues and the role of web, data mining, and data warehousing. This discussion continued at the IFIP meeting in Greece in 1998 and it was felt that the IFIP group should monitor the developments made by the security working group of the world wide web consortium. The discussions included those based on technical, social, and political aspects. However it was only at the IFIP Conference in July 2002 at the University of Cambridge, England that there was tremendous interest in privacy. In this section we will examine all aspects.

First of all, with the world wide web, there is now an abundance of information about individuals that one can obtain within seconds. This information could be obtained through mining or just from information retrieval. Therefore, one needs to enforce controls on databases and data mining tools. This is a very difficult problem especially with respect to data mining. In summary, one needs to develop techniques to prevent users from mining and extracting information from the data whether they are on the web or on servers. Now this goes against all that we have said about data mining in our previous papers (see for example, [THUR03]). That is, we have portrayed data mining as a technology that is critical for say analysts and other users so that they can get the right information at the right time. Furthermore, they can also extract patterns previously unknown. This is all true. However, we do not want the information to be used in an incorrect manner. For example, based on information about a person, an insurance company could deny insurance or a loan agency could deny loans. In many cases these denials may not be legitimate. Therefore, information providers have to be very careful in what they release. Also, data mining researchers have to ensure that privacy aspects are addressed.

Next, let us examine the social aspects. In most cultures, privacy of the individuals is important. However, there are certain cultures where it is impossible to ensure privacy. These could be related to political or technological issues or the fact that people have been brought up believing that privacy is not critical. There are places where people divulge their salaries without thinking twice about it, but in many countries, salaries are very private and sensitive. It is not easy to change cultures overnight, and in many cases you do not want to change them, as preserving cultures is important. So what overall effect does this have on data mining and privacy? We do not have an answer to this yet as we are only beginning to look into it. We are however beginning to realize that perhaps we do have many of the technological solutions for handling privacy. That is, many of the technologies we have proposed for information security in general and secrecy and confidentiality in particular could be applied for privacy. However we have to now focus on the social aspects. That is, we need the involvement of

social scientists to work with computer scientists on privacy and data mining.

Next, let us examine the political and legal aspects. We include policies and procedures under this. What sort of secrecy/privacy controls should one enforce for the web? Should these secrecy/privacy polices be mandated or should they be discretionary? What are the consequences of violating the secrecy/privacy polices? Who should be administering these policies as well as managing and implementing them? How is data mining on the web impacted? Can one control how data is mined on the web? Once we have made technological advances on data mining, can we then enforce secrecy/privacy controls on the data mining tools? How is information transferred between countries? Again we have no answers to these questions. We have, however, begun discussions. Note that some of the issues we have discussed are related to privacy and data mining, and some others are related to just privacy in general.

We have raised some interesting questions on privacy issues and data mining as well as privacy in general. As mentioned earlier, data mining is a threat to privacy. The challenge is on protecting the privacy but at the same time not losing all the great benefits of data mining. At the 1998 knowledge discovery in database conference in New York City, there was an interesting panel on the privacy issues for web mining. Much of the focus at that panel was on legal issues. It appears that the data mining as well as the information security communities are now conducting research on privacy. Furthermore, social scientists are also now interested in privacy in the new information technology era (see for example, the work by William Bainbridge [BAIN03]).

5. INFERENCE PROBLEM AND PRIVACY

In an earlier section we discussed the inference problem. In general when we think of the inference problem we have secrecy in mind. However, many of the concepts apply for privacy also. In our previous work (see for example [THUR93]) we defined various types of security constraints and subsequently designed and developed systems to process these security constraints. For example, ships locations and missions taken together are Classified while individually they are Unclassified. Similarly we can define privacy constraints such as names and salaries taken together are Private and individually they are Public. Similarly names and healthcare records taken together are Private while individually they are Public.

When Inference problem is considered to be a privacy problem, then we can use the inference controller approach to address privacy. For example, we can develop privacy controllers similar to our approach to developing inference controllers [THUR93]. Furthermore, we can also have different degrees of privacy. For example, names and age together could be less private while names and salaries together could be more private. Names and healthcare records together could be most private. One can then assign some probability or fuzzy value associated with the privacy of an attribute or a collection of attributes. We need to investigate further as to whether the privacy controllers could process the privacy constants during database design, update and query operations.

Lot of work has been carried out on the inference problem in the past. We need to revisit this research and see whether it is applicable for the privacy problem.

6. PRIVACY ENHANCED / SENSITIVE DATA MINING

As we have mentioned, the challenge is to provide solutions to enhance national security but at the same time ensure privacy. There is now research at various laboratories on privacy enhanced/sensitive data mining (e.g., Agrawal at IBM Almaden, Gehrke at Cornell University and Clifton at Purdue University, see for example [AGRA00], [CLIF02, [GEHR02]). The idea here is to continue with mining but at the same time ensure privacy as much as possible. For example, Clifton has proposed the use of the multiparty security policy approach for carrying out privacy sensitive data mining. While there is some progress we still have a long way to go. Some useful references are provided in [CLIF02] (see also [EVFI02]).

We give some more details on an approach we are proposing. Note that one mines the data and extracts patterns and trends. The privacy constraints determine which patterns are private and to what extent. For example, suppose one could extract the names and healthcare records. If we have a privacy constraint that states that names and healthcare records are private then this information is not released to the general public. If the information is semi-private, then it is released to those who have a need to know. Essentially the inference controller approach we have discussed is one solution to achieving some level of privacy. It could be regarded to be a type of privacy sensitive data mining. In our research we have found many challenges to the inference controller approach we have proposed (see [THUR95]). These challenges will have to be addressed when handling privacy constraints.

Note that not all approaches to privacy enhanced data mining are the same. Researchers are taking different approaches to such data mining. Some have argued that privacy enhanced data mining may be time consuming and may not be scalable. However we need to investigate this area more before we can come up with viable solutions.

7. CIVIL LIBERTIES VS. NATIONAL SECURITY

Civil Liberties are about protecting the rights of the individual whether it is privacy rights, human rights or civil rights. There are various civil liberties unions and laws protecting the rights of individuals (see for example, http://www.aclu.org/]).

There has been much debate recently among the counter-terrorism experts and civil liberties unions and human rights lawyers about the privacy of individuals. That is, gathering information about people, mining information about people, conduction surveillance activities and examining say e-mail messages and phone conversations are all threats to privacy and civil liberties. However, what are the alternatives if we are to combat terrorism effectively? Today we do not have any effective solutions. Do we wait until privacy violations occur and then prosecute or do we wait until national security disasters occur and then gather information? What is more important? Protecting nations from terrorist attacks or protecting the privacy of individuals? This is one of the major challenges faced by technologists, sociologists and lawyers. That is, how can we have privacy but at the same time ensure the safety of nations? What should we be sacrificing and to what extent?

I have served on panels on national security, database technologies and privacy as well given various keynote addresses including at the White House and at the United Nations. I have heard audiences say that if they can be guaranteed that there is national security, then they would not mind sacrificing privacy. However they would not want to sacrifice privacy for a false sense of security. On the other hand I have heard people say that some security is better than nothing. Therefore, even if one cannot guarantee national security, if some security is provided, then sacrificing privacy is not an issue. I have also heard from human rights lawyers about privacy violations by government under the pretext of national security. Some others are very nervous that all the information gathered about individuals may get into the wrong hands one day after we have hopefully eliminated terrorism and then things could be disastrous. Yet some others say that on no account will they sacrifice privacy.

While we have no solutions today, we will certainly hear more about it in coming months and years. The question is, if we assume that there will be no misuse of information, then should we sacrifice privacy for national security? Is it reasonable to make such an assumption? On the other hand should national security be of utmost importance and we prosecute those who have violated privacy on a case-by-case basis? Do we have adequate laws? We have no answers, just questions. However, I have been raising the awareness since my first keynote address on this topic in 1996 at the IFIP Database Security Conference in Como, Italy. It is now that we are hearing much more about this. We still have a lot to do here.

8. SUMMARY

This paper is devoted to the important area of privacy related to web and data mining. While there have been efforts on applying data mining for handling national security and information security problems such as intrusion detection, in this paper we have focused on the negative effects of data mining. In particular, we discussed the inference problem that can result due to mining as well as ways of compromising privacy especially due to web data access.

First, we gave an overview of the inference problem and then discussed approaches to handling this problem that result from mining. Warehousing and inference issues were also discussed. Then we provided an overview of the privacy issues. Next we discussed the inference controller approach for handling privacy and also examined privacy sensitive data mining. Finally we discussed civil liberties vs. national security.

While little work has been reported on privacy issues for web and mining, we are moving in the right direction. There is increased awareness of the problems, and groups such as the IFIP working group in database security are making this a priority. As research initiatives are started in this area, we can expect some progress to be made. Note that there are also social and political aspects to consider. That is, technologists, sociologists, policy experts, counter-terrorism experts, and legal experts have to work together to combat terrorism as well as ensure privacy. Note that the number of web security conferences is increasing including workshops on privacy (see [ACM02]). Recently the NSF sponsored workshop on Next Generation Data Mining focused both on counter-terrorism and privacy (see [NGDM02]). As the web becomes more and more sophisticated, there is also the potential for more and more threats. Therefore we have to be ever vigilant and continue to investigate, design and implement various privacy measures for the web, but at the same time we must ensure national security. This will be our major challenge.

9. ACKNOWLEDGMENTS

I thank NSF and the MITRE Corporation for their support to continue my work on data mining, counter-terrorism, information security and privacy. The views and conclusions expressed in this paper are those of the author and do not reflect the policies or procedures of the National Science Foundation, the MITRE Corporation or of the US Government.

10. REFERENCES

[ACM02] ACM Computer Security Conference Workshop on Privacy, Washington DC, November 2002.

[AFSB83] Air Force Summer Study Report on Multilevel Secure Database Systems, Washington DC, 1983.

[AGRA00] Agrawal, R, and R. Srikant, "Privacy-preserving Data Mining," Proceedings of the ACM SIGMOD Conference, Dallas, TX, May 2000.

[BAIN03] Bainbridge, W., "Privacy," Encyclopedia of Community, Sage Reference, Thousand Oaks, CA, 2003

[CLIF96] Clifton, C. and D. Marks, "Security and Privacy Implications of Data Mining", Proceedings of the ACM SIGMOD Conference Workshop on Research Issues in Data Mining and Knowledge Discovery, Montreal, June 1996.

[CLIF00] Clifton, C., "Using Sample Size to Limit Exposure to Data Mining," Journal of Computer Security, November 2000.

[CLIF02] Clifton, C., M. Kantarcioglu and J. Vaidya, "Defining Privacy for Data Mining," Purdue University, 2002 (see also Next Generation Data Mining Workshop, Baltimore, MD, November 2002).

[EVFI02] Evfimievski, A., R. Srikant, R. Agrawal, and J. Gehrke, "Privacy Preserving Mining of Association Rules," In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, July 2002.

[GEHR02] Gehrke, J., "Research Problems in Data Stream Processing and Privacy-Preserving Data Mining," Proceedings of the Next Generation Data Mining Workshop, Baltimore, MD, November 2002.

[HINK88] Hinke T., "Inference and Aggregation Detection in Database Management Systems," Proceedings of the Security and Privacy Conference, Oakland, CA, April 1988.

[MORG88] Morgenstern, M., "Security and Inference in Multilevel Database and Knowledge Base Systems," Proceedings of the ACM SIGMOD Conference, San Francisco, CA, June 1987.

[NGDM02] Next Generation Data Mining Workshop, Baltimore, MD, November 2002.

[THUR87] Thuraisingham, B., "Multilevel Security for Relational Database Systems Augmented by an Inference Engine," Computers and Security," December 1987.

[THUR90] Thuraisingham, B., "Recursion Theoretic Properties of the Inference Problem," MITRE Report MTP291, June 1990 (also presented at the 1990 Computer Security Foundations Workshop, Franconia, NH, June 1990).

[THUR91a] Thuraisingham, B., "On the Use of Conceptual Structures to Handle the Inference Problem," Proceedings of the 1991 IFIP Database Security Conference, Shepherdstown, WVA, November 1991.

[THUR91b] Thuraisingham, B., "Nonmonotonic Types Multilevel Logic for Multilevel Secure Data and Knowledge Base Management System," Proceedings of the IEEE Computer Security Foundations Workshop, Franconia, NH, June 1991.

[THUR93] Thuraisingham, B., W. Ford and M. Collins, "Design and Implementation of a Database Infernce Controller," Data and Knowledge Engineering Journal, December 1993.

[THUR95] Thuraisingham B. and W. Ford, "Security Constraint Processing in a Multilevel Distributed Database Management System," IEEE Transactions on Knowledge and Data Engineering, April 1995.

[THUR96] Thuraisingham, B., "Data Warehousing, Data Mining and Security," IFIP Database Security Conference, Como, Italy, July 1996 (paper in book by Chapman and Hall, 1997).

[THUR98] Thuraisingham, B., "Data Mining: Technologies, Techniques, Tools and Trends," CRC Press, FL, 1998.

[THUR03] Thuraisingham, B. "Web Data Mining: Technologies and Their Applications to Counter-terrorism," CRC Press, FL, 2003.

About the author:

Dr. Bhavani Thuraisingham is the Program Director in Data and Applications Security at the National Science Foundation. She has been with the MITRE Corporation since January 1989 and is currently on IPA to NSF. She has worked in secure databases for over seventeen years and is the recipient of IEEE Computer Society's 1997 Technical Achievement Award for "outstanding and innovative contributions to secure distributed data management" and recently IEEE's 2003 Fellow Award for "contributions to secure systems involving database systems, distributed systems and the web". She has published over 400 technical papers and reports including over 50 journal articles in secure data management and information technology. She is the inventor of three patents for MITRE on Database Inference Control. She has written 6 books on data management and data mining for technical managers. Her recent book is on Web Data Management Technologies and Their Applications to Counter-terrorism based on her keynote presentations on the subject at the White House and at the United Nations in 2002.