Crossroads
The ACM Student Magazine

Multilevel Security: Privacy by Design

by ***Stephany Filimon***

Introduction

Although data mining and database security are common topics of conversation in the computer and information sciences, they do not often inspire furious articles from New York Times columnists and letters of outrage from readers. Yet, in the fall of 2002, high-profile newspapers began to run stories on the potential threats to privacy and civil liberties posed by Total Information Awareness (TIA). Total Information Awareness was a program created within the Information Awareness Office (IAO) of the United States Defense Advanced Research Projects Agency (DARPA) in February 2002. Total Information Awareness sought to develop and use technologies such as data mining and biometrics in order to monitor private communications and commercial transactions, process large amounts of data from different sources, and to look for patterns of behavior that might reveal terrorist activity [**1**].

The events of the September 11, 2001 terrorist attack on the US and reports of unheeded security threats led many Americans to encourage the US Department of Defense (DoD) to adopt high-tech approaches to homeland security. News about high-tech methods, however, came in the context of increased public attention to other threats to privacy, namely the USA PATRIOT Act and TIA program. This article introduces the social, technological, and ethical considerations of the need for both

national security and privacy protection while reviewing the database concepts most closely related to this discussion. This paper also introduces the Genisys project, which may enable both information access and privacy protection. It is important to note, since the time of writing, Congress has voted to eliminate funding for Total Information Awareness and closed the Information Awareness Office. However, this particular action by Congress does not signal the end of this highly contested issue or the related technologies.

Background

Total Information Awareness sought to create mechanisms for collecting and mining vast amounts of data. The name was coined by the DoD and intended to signify "the consideration of every information source available worldwide to uncover terrorists" [**9**]. The DoD may have had good intentions when it coined the name for the TIA program, but it failed to include complementary privacy protections and policies that could protect innocent citizens from unwarranted suspicion and prevent abuse [**10**].

While better information awareness and processing may be necessary to detect and prevent acts of terrorism, the technology to do so is never neutral. The potential for abuse is enormous, and the legal system designed to protect privacy has yet to catch up with so much new technology [**1**]. At the heart of debates over the Patriot Act are larger and more general ethical questions about national security and privacy. What are the alternatives to privacy violation if we are to combat terrorism effectively [**9**]? What is more important: protecting nations from terrorist attacks or protecting the privacy of individuals? Must we choose between the two [**9**]? Can we find, through combined efforts in the creation of technologies, policies, and legal parameters, a happy medium? These questions and more have crossed disciplinary lines, from ethics, public policy, social science, and law to computer security and information science.

General Terminology

An overview of the following concepts and techniques may be helpful when reading about the database technologies and projects discussed in this piece.

Multilevel Analysis or Multilevel Modeling

Multilevel analysis is a familiar method of research in disciplines like sociology and

health sciences. When analyzing statistical data, the relationships between people that sit in the same office or classroom, or live in the same apartment building or neighborhood are often ignored. **Multilevel modeling** is a statistical technique for taking these dependencies into account, arranged in hierarchies (this technique is also known as hierarchical linear modeling) [2]. Correlations are sought between students within classrooms, between classrooms within schools, between schools within school districts, and so forth [2]. Multilevel modeling attempts to model the hierarchical relationships that are found in the real world. Multilevel modeling is certainly not the same thing as multilevel relational data models, as discussed later, but familiarity with the concept of multilevel modeling can be helpful.

Classical Relational Database Model

Relational databases (RDBs) have existed since the 1970s and are the most commercially available systems in the world [8]. A **relational database** stores data in one or more tables, which can be joined in various ways to allow efficient information access. Most government database systems are based on the classical relational model, a model that cannot provide the information analysis the government seeks or the privacy protections many desire. To its credit, government reliance on the classical relational model is one of the things the Genisys project seeks to change [3].

Multilevel Secure Relational Model

A **multilevel secure relational database** represents information in a multilevel state that is representative of the "real," exterior world [5], meaning (in this case) a government employee's access to and knowledge of information is directly related to his/her government-imposed security level. Most commercial database management systems (DBMSs) provide some form of data security by controlling modes of user access privileges to data, but these access controls are discretionary and do not provide adequate mechanisms for preventing the unauthorized disclosure of information [4]. Multilevel systems, on the other hand, require additional mechanisms for enforcing mandatory (nondiscretionary) access controls [4].

Multilevel Secure Database Management System

A large number of databases in the DoD, intelligence community, and civilian government agencies contain data that are classified to have different security levels. All database users are also assigned security clearances [4]. The responsibility of a

multilevel secure database management system (DBMS) is to assure that each user gains access, directly or indirectly, to only the data for which he or she has proper clearance [4]. Private corporations, like health or insurance providers, also use security levels and clearances to ensure secrecy of sensitive information, but their procedures for assigning security tend to be much less formal than those of the federal government.

Predictive Data Mining

Data mining is the search for valuable information in large volumes of data [13]. **Predictive data mining** is a search for very strong patterns in large data sets that can generalize to accurate future decisions. Most data-mining problems are an excellent fit to the classical problem of prediction from prior samples [13]. Data mining uses sophisticated statistical analysis and modeling techniques to uncover patterns and relationships hidden in organizational databases [12]. Prediction is at the heart of data mining.

The Genisys Program

The Genisys Program, created within TIA, is not a software program but a project. The Genisys Program aims to produce technology that enables the creation of "ultra-large, all-source information repositories" [3]. Genisys also seeks to create technology that allows government intelligence analysts to find patterns in data while preventing access to identifying information (i.e., details about individuals themselves) [6].

The DoD believes that in order to predict, track, and preempt terrorist attacks, the US requires a "full-coverage database containing all information relevant to: identifying potential foreign terrorists and their possible supporters; their activities; prospective targets; and, their operational plans" [3]. As Teresa Lunt, a computer scientist at the Palo Alto Research Center (PARC), explains, "the government wants to have predictive models so it can build up evidence of a significant future attack. It doesn't need identifying access to anybody to build up these models" [6].

The DoD understands that current database technology cannot sufficiently address these needs, and many familiar with database technology would agree. Much of the database technology available and in use today is based on a paradigm defined in the 1980s known as the classical relational model [3]. Today, computer processors, storage media, and networks possess many more capabilities than the classical model

allows for; likewise, Genisys aims to reinvent database technology consistent with today's needs and capabilities [3].

According to the Information Awareness Office and in contrast to more classical relational databases, Genisys will:

- Require no "a priori data modeling" and use a simpler query language, more easily and dramatically increasing information coverage.
- Support the automated restructuring and projection of data, making it easier to declassify and share data between government agencies.
- Store data in context of time and space to help resolve the uncertainty that necessarily exists in data, but for which current models do not allow.
- Create privacy filters to protect those not involved with foreign terrorists.
- Develop a vast, distributed system architecture for managing the huge volume of raw data input, analysis results, and feedback [3].

The fourth item, "create privacy filters to protect those not involved with foreign terrorists," is of particular interest here. Although this item provides clear evidence of a stipulation to create privacy filters, many feel that it is insufficient when considering TIA's potential for abuse and the suggestion of criminal behavior. On the other hand, the inclusion of privacy-protection in the Genisys program may lead to the creation of medium that can protect privacy intrusion and abuse from inside and out, while allowing the creation of a database with increased coverage.

Privacy Protection Techniques in Genisys

In early 2002, the Information Systems Advanced Technology (ISAT) panel, a group within the DoD comprised of computer scientists, privacy experts, government officials, and computer industry executives, commissioned a study of technological methods to protect private data in information systems [10]. The focus of the ISAT study was not TIA; the study aimed to investigate and review methods to protect individual data contained in commercial and government databases [10].

The subject of TIA, however, was discussed at ISAT meetings. The panel suggested that, "one way to make sure that private information in a database is kept safe is selective revelation." Put simply, the computer responds in a need-to-know fashion during a database search, providing information only if the user conducting the search

holds the required federal or legal clearances [10].

As an example the ISAT report stated that, "an analyst might issue a query asking whether there is any individual who has recently bought unusual quantities of certain chemicals, and has rented a large truck. The algorithm could respond by saying yes or no, rather than revealing the identity of an individual. The analyst might then take that information to a judge or other appropriate body, seeking permission to learn the individual's name or other information about the individual. By revealing information iteratively, we prevent the disclosure of private information except when a sufficient showing has been made to justify that revelation" [10].

The panel had an excellent and reasonable suggestion that may be able to satisfy both the DoD and the American Civil Liberties Union (ACLU). While the solution is easily presented, it is not easily implemented. Before such "selective" or "privacy-sensitive revelation" can take place, a problem of inference must be considered.

The Inference Problem

Inference is the process of posing queries and deducing unauthorized information from the legitimate responses received [9]. An inference problem exists in a multilevel database if knowledge of some objects in the database allows information with a higher security level to be inferred. In other words, an inference problem exists if users are able to infer sensitive information from a sequence of queries that each have a low security classification [7].

For example, the names and salaries of individuals may be unclassified individually but classified when taken together. A person could retrieve names and employee numbers, and then later retrieve the salaries and employee numbers only to make associations on their own between names and salaries [5]. These two queries together constitute an inference channel [7]. If both queries are made, it is possible for the user to infer which person has a certain salary level. Similarly, if there was only one woman taking a college course, asking the average grade of all the women in the course would reveal that woman's grade [6].

The inference problem is a long-standing one. The DoD began to research multilevel secure databases in the mid-1970s [4]. Earlier research did not manage to solve the inference problem, and at that time researchers did not have data mining to take into

account. The inference problem remains open, but researchers have recently made major strides in solving it.

Earlier this year, Teresa Lunt of PARC was the recipient of a $1 million DoD contract [10] out of a $3.5 million project. Lunt will work on this project over the course of three years [6]. The privacy appliance she envisions has been referred to as a kind of "inference firewall" [6] that can perform selective revelation. This inference firewall is designed to sit between data (like phone or credit card records) and government analysts [6] conducting a search at low levels of authorization. The firewall would anticipate queries that seem harmless but might unintentionally reveal details about a particular person [6], and subsequently prevent the revelation of any identifying data. The appliance would block out names, credit card numbers, Social Security numbers, and other identifying information.

Jessica Staddon, one of Lunt's colleagues at PARC, has also made some strides in solving the inference problem. She has found that, when inferences are prevented during query processing, more flexible access control is possible. However, when these inferences are prevented, query processing slows [1]. Staddon has proposed a Dynamic Inference Control, in which access controls can be made sufficiently dynamic to ensure easy access to the information a particular user is entitled and cleared to obtain, while retaining fast query processing [1].

Just as more general questions and concerns about privacy point toward new technologies that allow us to collect and analyze vast amounts of personal and potentially sensitive data, new data analysis and information retrieval techniques point toward larger social questions about privacy and national security.

Social Considerations

Members of Congress, nonprofit organizations, policy groups, intelligence and security analysts, defense specialists, and others have and will continue to debate the ethical and social questions that have been raised. What are the alternatives to privacy violation if we are to combat terrorism effectively? What is more important: protecting nations from terrorist attacks or protecting the privacy of individuals [9]? As is so often the case, there are many more questions than answers. The considerations noted here are not exhaustive, but a sample that covers technological, societal, and political concerns.

- Even if researchers were able to construct a system that protects privacy, it must be noted that there has not been much success with building and maintaining large databases that are secure, let alone the vast information warehouses available online.
- Privacy and civil rights activists should not focus solely on databases containing an abundance of personal data. Missing data that was never collected during an investigative process can also affect database searches. As a simple example, take an individual that may have been convicted of a crime. There is no definite guarantee that all related information, including a successful appeal and overturned decision, is also collected.
- Enforcing controls on databases and data-mining tools is inherently difficult. The ability for controls to work both online and on native databases and servers, as well as the ability to grapple with the use of multiple data-mining tools and users, must be evident. In addition, organizations and corporations must have guidelines in place to help them determine how many individuals should have authorization to access highly personal information. Government agencies have not yet been able to prevent spying and the exchange of information by agents. Access to highly personal information may also play a role in crimes like bribery and blackmail, and involve individuals both within and outside of government offices.
- Similarly, considerations must be made with regards to the database administrator (DBA). Although the database administrator is not responsible for creating or obtaining information, he or she is usually responsible for managing it. Database administrators tend to have access to all information in a given database. The idea of having DBAs that must also obtain clearance levels or be monitored needs to be evaluated.
- If privacy is truly so important, Genisys should not be the only project that bears a privacy stipulation. Privacy protection may be worthy of a program or project of its own [**10**].

In fairness, the DoD has made some recent improvements with regards to privacy protection. The DoD continues to fund research on how to search databases while protecting identities [**6**]. According to Wade Roush of the *Technology Review*, four additions to the DoD's lexicon are:

1. **Anonymization**, which scrambles identifying information like names, addresses, and phone numbers before data are released to investigators

2. **Self-Reporting Data**, wrapped in software or digital watermarks that guard against misuse of private information by tracking who has used the data, and where the data have moved
3. **Immutable Audits**, which would keep tabs on investigators by distributing records of data access to multiple keepers to prevent alteration or tampering
4. **Privacy Appliances**, which could automatically filter out seemingly harmless queries that might allow investigators to infer the identities of individuals

Conclusion

At the time of this writing, Congress had taken a strong, public stand against the programs and projects that were once part of the TIA program, eliminating all funding and closing the Information Awareness Office [**1**]. However, given the powerful law enforcement and national security benefits these technologies may offer, it is unlikely that all related research will simply grind to a halt.

To the contrary, research will continue to exist but will do so in new forms, like the new Novel Intelligence From Massive Data project, for instance. It is also more likely that TIA-related research will move underground into programs beyond the reach of public scrutiny, discourse, and influence. In short, the research that affects us most may proceed without us. If information and control becomes less available to the citizens that will be affected by these new technologies, I doubt we can honestly say that we are better off.

As we have asked ourselves more often during the past two years, what is more important: protecting nations or preserving the privacy of individuals? As a growing body of research suggests, we may not have to choose between the two. Novel approaches to solving the inference problem, privacy-sensitive data-mining capabilities, and other interesting technologies - along with updated legal parameters and public policies - could enable everyone to enjoy the positive aspects of both.

Rather than panic and quickly demand the close of programs like TIA, students of computer security and information science should respond to the challenges these programs pose. The problems reviewed here are vast, complicated, and somewhat messy, but solutions will contribute not only to science but also to privacy protection and national security. The area of privacy-sensitive data mining is ripe for research that crosses many disciplines and fields, across the social sciences, ethics, and

computer and engineering sciences. There are many approaches being made toward privacy-sensitive data mining while revealing useful information, and there is a responsibility to investigate each of these to the fullest extent possible.

References

1

EPIC Total Information Awareness Page. 26 September 2003. *The Electronic Privacy Information Center (EPIC)*, Washington, DC. 28 September 2003.

2

Hox, Joop. Multilevel Analysis: Techniques and Applications. New Jersey: Lawrence Erlbaum Associates, 2002.

3

Information Awareness Office: IAO Mission. DARPA, Washington, DC. 09 September 2003.

4

Jajodia, Sushil and Ravi Sandhu. "Toward a Multilevel Secure Relational Data Model." Proceedings of the 1991 ACM SIGMOD International Conference on Management of Data. 1991: p. 50-59.

5

Qian, Xiaolei and Teresa Lunt. "A Semantic Framework of the Multilevel Secure Relational Model." *IEEE Transactions on Knowledge and Data Engineering*. 1997: p. 292-301.

6

Roush, Wade. "Surveillance with Privacy." *Technology Review*. September 2003: p. 26.

7

Staddon, Jessica. "Dynamic Inference Control." DMKDO3: 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. 2003: p. 94-99.

8

Teorey, Toby J. Database Modeling & Design. San Francisco: Morgan Kaufman Publishers, Inc., 1999.

9

Thuraisingham, Bhavani. "Data Mining, National Security, Privacy and Civil Liberties." *SIGKDD Explorations* 4:2.

10

Total Information Awareness: Down, but not out. Farhad Manjoo. 28 January 2003. *Salon*, San Francisco, CA. 10 September 2003.

11

The USA PATRIOT Act. 20 May 2003. *The Electronic Frontier Foundation (EFF)*, San Francisco, CA. 22 September 2003.

12

Wang, John. Data Mining: Opportunities and Challenges. Hershey, PA: Idea Group, Inc., 2003.

13

Weiss, Sholom M. and Nitin Indurkhya. Predictive Data Mining: A Practical Guide. San Francisco: Morgan Kaufman Publishers, Inc., 1998.

Biography

Stephany Filimon (**stephany@imagetext.net**) is PhD student at the Illinois Institute of Technology in Chicago, IL. She researches the role of geographic information systems in interaction design and information search and retrieval, and enjoys writing about technology.