NISS

Homeland Insecurity: Datamining, Terrorism Detection, and Confidentiality

Stephen E. Fienberg

Technical Report Number 148
December 2004

Homeland Insecurity: Datamining, Terrorism Detection, and Confidentiality

Stephen E. Fienberg
Carnegie Mellon University
Department of Statistics
Pittsburgh, PA 15213-3890, U.S.A.
fienberg@stat.cmu.edu

Following the terrorist attacks of September 11, 2001, there was heightened attention in the United States on the use of multiple government and private databases for the identification of possible perpetrators of future attacks, part of an unprecedented expansion of federal government datamining activities, many involving databases containing personal information This article reviews some proposals that have surfaced for the search of multiple databases without compromising possible pledges of confidentiality to the individuals whose data are included and their link to the related literature on privacy-preserving datamining. In particular, we focus on the concept of *selective revelation* and its confidentiality implications.

HOMELAND SECURITY AND THE SEARCH FOR TERRORISTS

A recently issued report from the U.S. General Accounting Office [8] notes that at least 52 agencies are using or are using or planning to use data mining, "factual data analysis," or "predictive analytics," in some 199 different efforts. Of these, at least 29 projects involve analyzing intelligence and detecting terrorist activities, or detecting criminal activities or patterns.[1]

Perhaps the most visible of these efforts has been the Total Information Awareness (TIA) program initiated by the Defense Advanced Research Program (DARPA) in DARPA's Information Awareness Office (IAO), which was established in January 2002, in the aftermath of the September 11 terrorist attacks. The TIA research and development program was aimed at integrating information technologies into a prototype to provide tools to better detect, classify, and identify potential foreign terrorists. When it came under public scrutiny in 2003, TIA morphed into the Terrorist Information Program (still TIA) with essentially the same objectives.

Another closely related example is the Multistate Anti-terrorism Information Exchange System (MATRIX), in use in five states, intended to provide "the capability to store, analyze, and exchange sensitive terrorism-related information in MATRIX databases among agencies, within a state, among states, and between state and federal agencies." MATRIX databases all involve personally identifiable information in what is not otherwise generally accessible form.

In both TIA and MATRIX, the dataminer can issue queries to the multiple linked databases and receive responses that combine data on individuals across the data bases. The goal is the identification of terrorists or criminals in a way that would not be possible from the individual databases. We distinguish between two aspects of this goal: (1) identification of known terrorists which is a form of retro- or postdiction, and (2) identification of potential future terrorists and profiling, which involves prediction. Prediction cannot be separated from uncertainty, postdiction might conceivably be. Most of the public outcry regarding TIA and MATRIX has focused on concerns regarding what has been described as "dataveillance" [2] and terrorist profiling–concerns both about the use of data for purposes other than those for

---

[1] Notable among the nonresponders to the GAO inquiry were agencies like the Central Intelligence Agency and the National Security Agency.

which they were collected without the consent of the individual, and about the quality and accuracy of the mined data and the likelihood that they may help falsely identify individuals as terrorists. Here we explore the related issues of the implications of the use of "linked" databases for the privacy of the individuals whose confidential information in contained in them.

PRIVACY-PRESERVING DATAMINING

Among the methods advocated to carry out such datamining exercises are those that are described as privacy-preserving datamining (PPDM). PPDM typically refers to datamining computations performed on the combined data sets of multiple parties without revealing each party's data to the other parties. The data consist of possibly overlapping sets of variables contained in the separate data bases of the parties and overlapping sets of individuals. When the the parties have data for the same variables but different individuals the data are said to be horizontally partitioned whereas when the individuals are the same but the variables are different the data are said to be vertically partitioned. Here we are concerned with the more complex case involving both overlapping variables *and* overlapping sets of individuals. PPDM research comes in two varieties. In the first, sometimes referred to as the construction of "privacy-preserving statistical databases," the data are altered prior to delivery for datamining, e.g., through the addition of random noise or some other form of perturbation. While these approaches share much in common with the methods in the literature on statistical disclosure limitation, they are of little use when it comes to the identification of terrorists. In the second variety, the problem is solved using what is known as "multi-party secure computation," where no party knows anything except its own input and the results. The literature typically presumes that data are included without error and thus could be matched perfectly if only there were no privacy concerns. The methods also focus largely on situations where the results are of some computation, such as a dot product or the description of an association rule. See the related discussion in Fienberg and Slavkovic [5].

A major problem with the PPDM literature is that the so-called proofs of security are designed not to protect the individuals in the database but rather the database owners, as in the case of two companies sharing information but not wanting to reveal information about their customers to one another beyond that contained in the shared computation. Once the results of the datamining consist of linked extracts of the data themselves, however, the real question is whether one of the parties can use the extra information to infer something about the individuals in the other party's data that would otherwise not be available.

Secure computation is a technique for carrying our computations across multiple databases without revealing any information about data elements found only in one database. The technique consists of a protocol for exchanging messages. We assume the parties to be *semi-honest*— i.e., they correctly follow the protocol specification, yet attempt to learn additional information by analyzing the messages that are passed. For example, Agrawal, Evfimievski, and Srikant [1] illustrate the secure computation notion via an approach to the matching problem for parties A and B. They introduce a pair of encryption functions E (known only to A) and E' (known only to B) such that for all x, $E(E'(x)) = E'(E(x))$. A's database consists of a list $\mathbf{A}$ and B's consists of a list $\mathbf{B}$. A sends B the message $E(\mathbf{A})$; B computes $E'(E(\mathbf{A}))$ and then sends to A the two messages $E'(E(\mathbf{A}))$ and $E'(\mathbf{B})$. A then applies E to $E'(\mathbf{B})$, yielding $E'(E(A))$ and $E'(E(B))$. A computes $E'(E(\mathbf{A})) \cap E'(E(\mathbf{B}))$. Since A knows the order of items in $\mathbf{A}$, A also knows the order of items in $E'E(A))$ and can quickly determine $\mathbf{A} \cap \mathbf{B}$. The main problems with this approach are (1) it is asymmetric, i.e., B must trust A to send $\mathbf{A} \cap \mathbf{B}$ back, and (2) it presumes semi-honest behavior.

Li et al. [6] describe a variety of scenarios in which the AgES protocol can easily be exploited by one party to obtain a great deal of information about the other's database and they explain drawbacks of some other secure computation methods including the use of one-way hash-based schemes. As Dwork and Nissim [4] note: "There is also a very large literature in secure multiparty computation. In secure multi-party computation, functionality is paramount, and privacy is only preserved to the extent that the function outcome itself does not reveal information about the individual inputs. In privacy-preserving statistical databases, privacy is paramount." The problem with privacy-preserving data-mining methods for terrorist detection is that they seek the protection of the latter while revealing individual records using the functionality of the former.

SELECTIVE REVELATION AND THE RISK-UTILITY TRADEOFF

To get around the problems associated with privacy-preserving datamining approaches such as those referred to above, those involved in the development of the TIA and MATRIX systems have advocated the use of what has come to be called "selected revelation," involving something like the risk-utility tradeoff in statistical disclosure limitation [3].

One specific approach has been the work on *privacy appliances* by Lunt [7]: "a mix of software and hardware to allow data scanning and 'selective revelation' of personally identifiable information." [9]. The privacy appliance is intended to be a stand-alone device that would sit between the analyst and the private data source so that private data stays in authorized hands. These privacy controls would also be independently operated to keep them isolated from the government. According to Lunt [7] the device would provide:

- **Inference control** to prevent unauthorized individuals from completing queries that would allow identification of ordinary citizens.

- **Access control** to return sensitive identifying data only to authorized users.

- **Immutable audit trail** for accountability.

Such claims for selective revelation and privacy appliances sound much like the impossible combination secure multi-party computation combined with an ensemble of privacy-preserving data-bases. To date there are no publicly-available prototypes of the privacy appliance. has there been any technical demonstration that data of the sort sought after for terrorist detect can be made available without seriously compromising the integrity of confidential databases containing personal information on individuals.

While the U.S. Congress stopped funding for DARPA's TIA program in 2003, and Lunt's research in particular, claims for similar systems that can aid in homeland security without compromising confidentiality abound. Statisticians in particular remain skeptical.

References

[1] Agrawal, R., Evfimievski, A., and Srikant, R. (2003). "Information sharing across private databases." In *Proceedings of the 2003 ACM SIGMOD Intl Conf. on Management of Data*, San Diego, CA.

[2] Clarke, Roger (1988). "Information technology and dataveillance." *Communications of the ACM, 31, (5), 498-512.*

[3] Duncan, George T., Keller-McNulty, Sallie A., and Stokes, S. Lynne (2004). "Database security and confidentiality: Examining disclosure risk vs. data utility through the R-U confidentiality map." Technical Report Number 142, National Institute of Statistical Sciences, March, 2004.

[4] Dwork, Cynthia and Nissim, Kobbi (2004). Privacy-preserving data mining in vertically partitioned databases." *Proc. CRYPTO 2004, 24th International Conference on Cryptology*, University of California, Santa Barbara.

[5] Fienberg, Stephen E. and Slavkovic, Aleksandra B. (2004). "Preserving the confidentiality of categorical statistical data bases when releasing information for association rules." Submitted for publication.

[6] Li, Yaping, Tygar, J.D., and Hellerstein, Joseph M. (2004). "Private Matching." IRB-TR-04-005, University of California, Berkeley, February, 2004.

[7] Lunt, Teresa (2003). "Protecting privacy in terrorist tracking applications." Presentation to the Department of Defense Technology and Privacy Advisory Committee, September 29, 2003. `www.sainc.com/tapac/library/Sept29/LuntPresentation.pdf`

[8] U.S. General Accounting Office (2004). *Data Mining: Federal Efforts Cover A Wide Range of Uses.* GAO-04-548, a report to the Ranking Minority Member, Subcommittee on Financial Management, the Budget, and International Security, Committee on Governmental Affairs, U.S. Senate, Washington, DC.

[9] Walker, Leslie (2003). "Balancing data needs and privacy." *Washington Post*, May 8, 2003, page E1.

ACKNOWLEDGMENTS

RÉSUMÉ

A la suite des attaques terroristes du 11 Septembre 2001, s'est produit aux Etats-Unis un intérêt accru pour l'utilisation des bases de données privées et celles de diverses administrations afin, dans l'éventualité d'un attentat, de pouvoir à l'avance en identifier les auteurs. Cela représente une partie d'un développement sans précédent des activités de "datamining" par le gouvernement fédéral. Parmi toutes ces bases exploitées actuellement, nombreuses sont celles qui contiennent des informations personnelles. Cet article passe en revue quelques propositions ayant émergé pour rechercher des bases de données multiples, sans compromettre les éventuelles promesses de confidentialité faites aux individus dont les données sont incluses. Il étudie également les liens entre ces propositions et les textes existants sur la préservation de la vie privée dans le "datamining". Nous nous concentrons, en particulier, sur le concept de "selective revelation" et ses implications dans le domaine de la confidentialité.