Some Requirements for Adoption of Privacy-Preserving Data Mining

PORTIA Project White Paper, April 2005

Joan Feigenbaum, Yale University, feigenbaum@cs.yale.edu
Benny Pinkas, HP Labs, Benny.Pinkas@hp.com
Raphael Ryger, Yale University, ryger@cs.yale.edu
Felipe Saint-Jean, Yale University, felipe.saint-jean@yale.edu

Abstract

We attempt to explain why our privacy-preserving salary-survey package was not adopted by our first group of intended users. Based on this evaluation, we formulate productive next steps toward the PORTIA goal of implementing and deploying practical, usable, privacy-preserving data-mining systems.

Introduction

Last year, we implemented a user-friendly, privacy-preserving software package for Taulbee-Survey computation [Taulbee]; the design and implementation of our system is described in [FPRS04]. The proximate motivation for our work was the decision by the Computing Research Association (CRA) to begin asking each Taulbee-participating department to supply a complete, anonymized list of faculty salaries, rather than just the max, min, median, and mean within each faculty rank, as it had done in previous iterations of the survey. Although we offered our open-source software to the CRA free of charge, the CRA chose not to use it but rather just to ask the member departments to trust it with the complete salary lists. Our goal in this brief white paper is to attempt to explain the non-adoption of our system and to formulate productive next steps toward the PORTIA goal of implementing and deploying practical, usable, privacy-preserving data-mining schemes.

Basics

First and foremost, it is important to note that the CRA did not ask for our help. As privacy researchers, we predicted (correctly, as it turns out and as we explain below) that some of the Taulbee-participating departments would object on privacy grounds to providing complete lists of salaries; we thought that these departments might view a privacy-preserving survey package (built on top of the FairPlay system [MNPS04]) as a solution that would allow them both to provide the requested data and to maintain privacy; and we saw this as an opportunity to pursue the PORTIA research-agenda item of building and experimenting with usable, privacy-preserving data-mining schemes. CRA Surveys-Committee Co-Chairs Stu Zweben and Bill Aspray and CRA Executive Director Andrew Bernat were very generous with time and information; they met with us to discuss their requirements before the 2004 survey and responded to our request for compliance statistics after the 2004 survey. That they chose not to use our software instead of the data-collection and data-analysis procedures that they already had in place and had been using for some time is not surprising; in fact, it is what usually happens when a CS-research group attempts to "transfer its technology" to an operations group that did not ask for that technology. We undertook the work reported in [FPRS04] fully aware of this fact, and our primary goal (which we achieved) was to conduct experimental research on privacy-preserving data mining.

CRA's Explanation

Essentially, CRA people offered two reasons that they did not use our system. The first is that they were not sure that it was necessary, i.e., that they were not sure that privacy concerns would actually deter many survey-participating departments from providing complete, anonymized lists of salaries. The second is that they themselves needed access to all of the cleartext data. Currently, they do have it, because access to cleartext data is an essential feature of the "trusted-party" model of survey computation. They would not have access to the cleartext data if our system were used, because it is an essential feature of the "secure, multiparty-computation" model on which FairPlay (and any package built on top of it, including ours) is based that no single party ever obtains another's cleartext data unless those data are logically inferable from that single party's inputs and the protocol's outputs.

After the survey was complete, Professor Zweben informed us [Zweben] that, out of the 182 departments that provided salary data of any kind, 147 provided the complete, anonymized lists, and 35 did not. These numbers include both US and Canadian departments. Zweben regards this 80.77% compliance rate as "excellent." Furthermore, those departments that gave reasons for not providing complete lists said that "institutional barriers" tied their hands, i.e., that their universities had policies in place that prohibited the disclosure of a complete (albeit anonymized) list of salaries for a department, not that they doubted that the CRA was indeed a trustworthy third party with respect to salary data. Whether an 80.77% compliance rate is acceptable is clearly a matter of opinion, and so we will not comment further on this aspect of things. We discuss "institutional barriers" in the next section.

CRA's reasons for wanting access to the cleartext data are quite compelling. First, they regularly go through an interactive process of "data cleaning" with survey participants; that is, they receive a completed survey form, read it and decide that something in it "does not look right," ask the survey participant about the suspect data, and receive a corrected form from the participant as a result. The secure, multiparty-computation model does not include a data-cleaning phase; rather, it regards the inputs supplied by the participants as "correct" by definition. Second, CRA representatives said that they wanted to retain complete sets of cleartext survey responses so that data collected in one year could be reused in a later year if, during that later year, new statistics were added to the Taulbee report. For example, Universities are currently grouped by "tier," and various aggregate statistics are reported for each tier; how these values change over time for a particular tier can be seen from the outputs of the survey. If, at some point in the future, Universities are also grouped by geographic region, CRA will be able retrospectively to track how aggregate statistics for a region have changed over time if it has the raw data but not if it only has the outputs of the survey. Note that, although this particular example is harmless, and it is indeed hard to imagine an aggregate Taulbee-survey statistic that would not be harmless, the same statement cannot be made about surveys in general. The prospect that survey data could be used for a purpose other than the one for which they were collected is exactly what causes some people to refuse to participate in surveys; the fact that the secure, multiparty-computation model precludes such use might help allay these people's fears and convince them to participate.

A Survey Participant's Explanation

We spoke with the Assistant Provost for Science and Technology at one of the Universities at which the Computer Science department did not provide a complete salary list because of institutional barriers. He confirmed that it was indeed longstanding University policy that departments are not allowed to disclose complete lists of salaries, even anonymized ones. Interestingly, the University has a detailed policy that describes which salary statistics a department is allowed to disclose, and this policy does not permit the department to answer the old Taulbee-survey salary questions of max, min, median, and mean either; specifically, it does not permit the disclosure of exact max or exact min. We described the secure, multiparty-computation model to the Assistant Provost and asked whether departments would be allowed to use a software package that conforms to this model in order to make their complete salary lists available for privacy-preserving statistical computations. Understandably, he was unable simply to answer yes or no (and suggested that we ask the University's General Counsel if we needed a definitive answer right away, but of course we had no such need, because the CRA had not adopted our package). He gave the following interesting reasons that the answer might be no, at least in the short run: (1) If an institution is bound by law, contract, or policy not to disclose certain data, then it is unclear whether that institution is in compliance with that law, policy, or contract if it makes those data available for processing through a software package built in the secure, multiparty-computation model; (2) although we claimed that the fact that our source code is available should allay worries that the software itself could leak confidential data, he pointed out that it may be difficult or expensive to verify correctness and security claims about this type of software package; and (3) the University's policy about confidentiality of salary lists is consistent with the policies of other comparable Universities, and no single member of this peer group would change its policy unilaterally – whether or not the use of privacy-preserving software packages is allowed is something that would have to be decided by the peer group as a whole.

Conclusions

In light of this experience, we conclude that the following three agenda items are high priorities for the PORTIA work on privacy-preserving data mining:

1. Work with a user community that has explicitly expressed interest in using privacy-preserving data-mining software. Fortunately, PORTIA colleague Alejandro Schaffer of the National Institutes of Health has informed us that genetics researchers form such a community [Scha04].
2. Address the problem of privacy-preserving data cleaning, and, more generally, bridge the gap between the idealized computational model of the literature on secure, multiparty computation and the way in which data mining is actually done.
3. Devise methods to determine whether privacy-preserving data-mining protocols, as formalized in the CS-research community, comply with laws and organizational policies. PORTIA colleagues at the Yale Law School suggest that an appropriate starting point for this part of the investigation is the "American pragmatism" school of legal theory, exemplified by the work of Mortin Horwitz [Horw92].

References

[FPRS04] J. Feigenbaum, B. Pinkas, R. Ryger, and F. Saint-Jean, "Secure Computation of Surveys," EU Workshop on Secure Multiparty Computation (SMP), Amsterdam, The Netherlands, 2004. http://www.zurich.ibm.com/~cca/smp2004/

[Horw92] M. Horwitz, **The Transformation of American Law, 1870-1960: The Crisis of Legal Orthodoxy**, Oxford University Press, 1992.

[MPNS04] D. Malkhi, N. Nisan, B. Pinkas, Y. Sella, “FairPlay – A Secure Two-Party Computation System,” Proceedings of the 13th Symposium on Security, Usenix, 2004, pp. 287-302.

[Scha04] A. Schaffer, “Two Problems with Multiple Genetic Studies of the Same Disease,” PORTIA Project White Paper, November 2004. http://crypto.stanford.edu/portia/pubs/articles/S169701188.html

[Taulbee] Computing Research Association, Taulbee Survey, http://www.cra.org/statistics.

[Zweben] Stu Zweben, private correspondence, February 23, 2005 and April 19, 2005.