Journal of Privacy and Confidentiality (2012) **4**, Number 1, 1–3

Special Issue on Statistical and Learning-Theoretic Challenges in Data Privacy

Aleksandra Slavkovic* and Adam Smith†, *Guest Editors*

This special issue presents papers based on talks from a workshop on "Statistical and Learning-Theoretic Challenges in Data Privacy" held at UCLA's Institute for Pure and Applied Mathematics (IPAM), February 22-26, 2010.

The workshop brought together researchers from a variety of areas in statistics, machine learning, cryptography, and data mining to discuss the theoretical foundations for research on data privacy. Two recurring themes were how the conflicting goals of privacy and utility can or should be formulated mathematically; and how the constraints of privacy—in their various incarnations—affect the accuracy of statistical inference and machine learning.

In all, there were 23 speakers on the program, four of whom gave tutorials. There was also a poster session, as well as a "rump session" consisting of very short (five-minute) talks. Participants at the workshop were invited to submit papers to this special issue. Eight papers were accepted, most of which directly reflect talks presented at the workshop.

In This Special Issue

Daniel Kifer and Bing-Rong Lin's contribution, "An Axiomatic View of Statistical Privacy and Utility," takes direct aim at the mathematical formulation of "privacy" and "utility" for statistical data. They formulate a class of privacy definitions and identify some basic axioms—closure under post-processing and convexity, for example—that they feel should be satisfied by reasonable definitions and explore the relationship of these axioms to existing definitions. They also look at measures of utility that satisfy similarly basic axioms (for example, not increasing under post-processing), paving the way for a more general theory of statistical privacy.

Motivated by the second theme, there are several papers on the meaning and potential of *differentially* private algorithms.

Larry Wasserman in "Minimaxity, Statistical Thinking and Differential Privacy" takes some initial steps at bridging the linguistic and conceptual gaps between the computer sciences and statistics communities working on data privacy. He first distinguishes between the query-response privacy model and sanitized database model, offering his perspective on why statisticians would prefer the latter—having access to a whole dataset rather than just summary statistics; it should be noted, however, that the

*Department of Statistics, Penn State University, University Park, PA, mailto:sesa@psu.edu
†Department of Computer Science & Engineering, Penn State University, University Park, PA, mailto:asmith@cse.psu.edu

sanitized data also have analytic limitations compared to the original data. In the second part of this essay, he discusses the role of minimax statistical theory for differential privacy, focusing mostly on nonparametric density estimation.

Benjamin Rubinstein, Peter Bartlett, Ling Huang, and Nina Taft address the design of differentially private algorithms for certain convex optimization problems in "Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning." They look at algorithms for empirical risk minimization (also sometimes called M-estimators) for which the loss function is convex, and give bounds on the distortion in the minimizer that are necessary to satisfy differential privacy. They also show how these ideas can be combined with dimensionality reduction techniques to handle learning in high-dimensional parameter spaces.

Xiaolin Yang, Stephen E. Fienberg, and Alessandro Rinaldo present an evaluation of a statistical utility of a differentially private algorithm for releasing binary contingency tables and propose an extension that is applicable to non-binary multi-way tables in "Differential Privacy for Protecting Multi-dimensional Contingency Table Data: Extensions and Applications." They evaluate the utility of differential privacy of the proposed algorithms on three real-life datasets in terms of "statistical distance" between the original and perturbed data. They conclude that these algorithms impair realistic statistical analysis in the case of large-sparse contingency tables, and that designing algorithms that stem from the principles of smooth sensitivity framework and are data dependent may provide better trade-off between utility and risk.

Several papers also looked at privacy issues tied to current statistical practice.

"Confidentialising Survival Analysis Output in a Remote Data Access System" by Christine M. O'Keefe, Ross Stewart Sparks, Damien McAullay, and Bronwyn Loong, describes the implementation of a remote analysis system enabling survival analysis. The focus here is on creating a usable system that is compatible with current practice yet provides some well-explained protections for the data (aggregation and outlier removal, for example).

Natalie Shlomo and Chris Skinner in "Privacy Protection from Sampling and Perturbation in Survey Microdata" examine a few commonly used procedures in statistical disclosure control of microdata from social surveys in the light of differential privacy and probabilistic differential privacy. The procedures examined are random sampling and perturbation schemes such as random data swapping, PRAM, and recoding. They illustrate that the viability of the two privacy definitions under perturbation schemes of survey data depend on the presence of zeros in a misclassification matrix.

David McClure and Jerome P. Reiter address the important issue of getting users to trust the results of inference from synthetic data in "Towards Providing Automated Feedback on the Quality of Inferences from Synthetic Datasets." They propose ways to release fidelity measures of data utility via remote verification servers for multiple-imputed partially synthetic data such that the information leaked by the quality measures themselves is minimized, and consider strengths and weaknesses of the proposed methods.

Finally, the last two papers consider the design of secure distributed protocols for statistical analysis. In "Achieving Both Valid and Secure Logistic Regression Analysis on Aggregated Data from Different Private Sources," Stephen E. Fienberg, Robert J. Hall, and Yuval Nardi propose a new way of using existing tools from the secure-function evaluation literature to compute parameters of logistic regression to high numerical accuracy when data are held by multiple parties without actually combining the data. They demonstrate the accuracy claims and running time on an extract from a real-life dataset, and provide a detailed discussion on possible information leakage and ways to mitigate it.

Stephen E. Fienberg and Jiashun Jin investigate connections between sparsity constraints in inference problems and confidentiality. Specifically, they show that the "phases" that appear in sparse inference problems (corresponding to qualitatively different types of possible inference, depending on the amount of available data and the dimension) have analogues in a setting where a dataset is shared among several parties who would like to jointly analyze it.

Acknowledgments

We are grateful to the referees for this special issue, who dedicated considerable time and effort to the review process, ensuring the high quality of the papers. We would also like to thank all the participants at the workshop for a stimulating, fun week in Westwood and, in particular, the authors who submitted papers to this special issue.