The Journal of Privacy and Confidentiality (2009) **1**, Number 1, pp. 1–6

First Issue Editorial

John M. Abowd*, Kobbi Nissim† and Chris Skinner‡

1 Privacy and Confidentiality

When the founders of this *Journal*—Cynthia Dwork, Stephen Fienberg and Alan Karr—made its initial call for papers, they and we identified many constituencies that participate in the scientific analysis of privacy and confidentiality. Statisticians, particularly those working within national statistical offices, have developed the field of statistical disclosure limitation. Computer scientists contribute work in privacy-preserving data-mining and cryptographic analyses of privacy. Lawyers and social scientists study the role of government and regulation in the creation and protection of individual and business privacy. Health researchers struggle with the trade-off between a patient's privacy and the contribution to science that access to integrated medical records might allow. Survey designers in all fields of human endeavor wrestle with methods of enticing survey cooperation under a variety of ethical and privacy guarantees. Gargantuan online services gather petabytes of data on search queries, online purchases, e-mail exchanges, and other social network interactions while pushing their computer scientists to exploit the corporate asset these data represent without damaging the companies' ability to do future business by breaching the confidence of their client/users. And many, many data users from all of the fields listed above perform analyses that are conditioned on the privacy and confidentiality protections imposed on their work without all the means to assess the consequences of those measures on the inferences they have made.

We are certainly not the first journal to venture into this domain. But we are the first journal to solicit actively contributions from the entire community that are aimed at multiple constituencies within that community. We think that a brief illustration of how the research questions share a common theme would provide a useful introduction to this first volume.

Government agencies around the world are mandated to provide general statistical information on population, economic activity, health, education, crime, labor markets, trade, and numerous other activities. The provision of general-purpose statistical data was placed in the public sector as a solution to the public goods problem. There is no rivalry in the consumption of statistical information: one person's use of a price index does not reduce the amount of "price index" available for another user. There is also no feasible exclusion from the use of statistical information: once an agency or business has published a price index, it can no longer reasonably control who uses that index and for what purpose. Thus, statistical data are a classical public good. And they will be underproduced and underutilized when they are provided by profit-oriented businesses (10; 2).

Recognizing the central role that publication of the public good plays in their mission, most government agencies have adopted a "trusted custodian" model for collecting and disseminating statistical data. Data entities (usually households or businesses) report identifiable

*School of Industrial and Labor Relations, Cornell University, Ithaca, NY, mailto:john.abowd@cornell.edu

†Department of Computer Science, Ben-Gurion University, Israel, mailto:kobbi@cs.bgu.ac.il

‡School of Social Sciences, University of Southampton, Southampton, UK mailto:C.J.Skinner@soton.ac.uk

information to the statistical agency, which undertakes to release summary versions of these data in "non-identifiable" form. The trusted custodian must, therefore, develop publication standards that define "non-identifiable." Statistical disclosure limitation (SDL) emerged to address this question (9; 3).

One of the critical SDL concepts is the notion of a "sensitive item" whose confidentiality must be protected. When a "trusted custodian" releases statistical data, the publication usually consists of collections of counts or magnitudes stratified by the characteristics of the entities to which these measures apply. The SDL literature designates one of these counts or magnitudes as "sensitive" if an external entity can estimate the value of another entity too precisely. "Trusted custodians" of statistical data operationalize their obligation to release non-identifiable data by enforcing rules that prohibit the release of the data in cells that are at "too much" risk to violate these limits, and often prohibit release in enough other cells to prevent reconstruction of the sensitive items from the cells in the released data (8).

While cell suppression remains standard at many agencies, the field of SDL has expanded greatly with many other different approaches—synthetic data, perturbation models, swapping, reporting bounds, reporting marginal tables for contingency tables, and the use of risk-utility trade-off models. Furthermore, in the computer science/data-mining literature we have seen many ideas revisited using new formal methods and new ideas proposed under the banner of privacy-preserving data-mining. In these applications, it makes sense to move to secure computation and differential privacy, which is where many of these ideas are sharpened and refocused.

A parallel approach to SDL emerged in the cryptographic literature, which seeks rigorous foundations for private data analysis. In this work data holders (custodians and proprietary database owners) can control precisely the information leakage, and its consequences, regarding any individual in the presence of attackers who are armed with a complete description of the database protection mechanism and with auxiliary knowledge about the data (*e.g.*, knowledge of a subset of the actual contents of the database, all answers to previously asked queries, *etc.*). The notion of privacy that has emerged is "differential privacy"(7; 5), where, informally, the behavior of the protection mechanism (and hence the view of the attacker) remain almost unchanged when an individual record is modified. The theoretical work on differential privacy has yielded solutions for function approximation, statistical analysis, data-mining, and sanitized databases.[1] It remains to see how these theoretical results might influence the privacy and confidentiality preserving practices of government agencies and private enterprises.

The distinction between a government agency acting as a "trusted custodian" and a non-governmental information source allowing some public queries of proprietary data is artificial. A "trusted custodian" can ensure trustworthiness by releasing nothing, which makes the government agency equivalent to the proprietary database owner except that there is no solution to the public good problem and the agency then withers away for lack of a mandate. Similarly, a proprietary database has no asset value if it is not used for some economic purpose that eventually produces a revenue stream. This economic activity necessarily involves the release of some information from the database during the transactions or queries that monetize its economic value. Upon this release of information the proprietary data are subject to leakage even if no other explicit queries are allowed.

The users of statistical data have a much more central role to play than either the "trusted custodian" or "privacy-protected" data-mining literatures recognize. The inferences and de-

[1] See (6) for a review of recent results.

cisions made by users are affected by the SDL and privacy-protection schemes adopted by providers. The conventions in SDL make the assessment of the effects of these protections on the inferences very difficult because critical parameters of the SDL process are secret. The privacy protection literature, which does provide tools for answering inference quality questions, insists that formal analysis can only be conducted when the protection parameters are public. Data users are caught in the middle. They use data that were protected by SDL or cryptographic methods, and they make assessments about the quality of their inferences in the face of such protections. But they rarely contribute to the design or assessment of the protection schemes, as noted by the National Academy of Science's Duncan et al. (4). Consequently, a rather large divide has emerged.

One of the central purposes of this *Journal* is to make the case for a consistent and rigorous discussion among these diverse research communities. Our goal is to publish cutting-edge work, including the contributions of all our target communities, that are accessible and useful to several distinct privacy and confidentiality constituencies. We summarize below the contributions of the articles in our first issue.

2 The First Issue

Our first issue contains four articles that are aimed at readers from the user, SDL, and computer science communities.

In "Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues," Margo Anderson and William Seltzer consider the history of the use of confidential micro data collected from business in the production of statistical products by U.S. government agencies. This article revisits the controversy they first ignited in their paper on breaches of confidentiality in U.S. census data (1). They review both the legal and the statistical underpinnings of these uses. More importantly, they consider the interaction of the legal environment and the practice of the major statistical agencies. There are important lessons from this study stemming from the ways in which potential breaches of confidentiality are managed by statistical agencies. Former Census Bureau Director, C. Louis Kincannon, comments on the issues raised by Anderson and Seltzer, and the authors rejoin. This interaction highlights the tensions between the duties of a statistical agency to meet its missions and the scientific evaluation of the successes, and failures, of those efforts.

The second paper is Yehuda Lindell and Benny Pinkas' invited overview "Secure Multiparty Computation for Privacy-preserving Data-mining." Secure multiparty computation allows several distrusting parties to compute a function of their joint private inputs while not leaking any information beyond the intended computation. The paper reviews secure multiparty computation starting from its definitions and their rationale, going through the basic building blocks to generic and problem specific constructions. The generality of the results (secure computation protocols exists for all feasible computable functions) suggests a very wide spectrum of potential applications, and indeed secure computation is a growing interest also of non-cryptographers. It is important, however, to understand the limitations and to clear up some of the misconceptions regarding secure computation—specific objectives that are addressed by this mini-tutorial. First, the generic constructions of secure computation protocols lead in many cases to practically inefficient protocols; hence, protocols for specific functions are often designed using less general techniques. The design of secure protocols is often quite intricate and delicate. Some common pitfalls are highlighted. Finally, even when efficiency is taken

aside, one should remember that although secure multiparty computation is an important tool for privacy, it is not a panacea. It should usually be used in conjunction with other techniques, such as those developed for private data analysis and SDL.

Partially synthetic data protect confidentiality by releasing samples from a properly constructed predictive distribution in lieu of releasing parts of the actual confidential database. Jerome Reiter and Robin Mitra tackle the problem of how to quantify the trade-off between confidentiality protection and data quality for these methods in "Estimating Risks of Identification Disclosure in Partially Synthetic Data." Inferential quality depends upon releasing multiple samples from the predictive distribution, but these reduce the confidentiality protection. This practical trade-off is studied in detail. As synthetic data techniques proliferate, users should study this trade-off carefully.

"Global Measures of Data Utility for Microdata Masked for Disclosure Limitation" by Mi-Ja Woo, Jerome Reiter, Anna Oganian, and Alan Karr contributes to the growing literature designed to help both data custodians and users assess the consequences of data protection systems on the usability of the release data. Global measures summarize data utility in manner that is applicable to many different analyses–for example, all inferences that depend upon the first and second moments or all conclusions influenced by the shape of the cumulative distribution. By studying the effects of confidentiality protections on such global measures, these authors provide very general tools for judging the quality of released statistical data.

3 Call for More Submissions

We hope that the process of finding and editing papers that have broad appeal in the privacy and confidentiality research communities benefits from your reading of our first issue. There are many papers produced by statisticians, computer scientists, social scientists, and others that have a direct message for the much broader scholarly community that studies privacy and confidentiality. Since the *Journal* seeks privacy research from statisticians, computer scientists and social scientists, we are particularly interested in collaborations among authors reflecting these multiple perspectives. Please help us locate and disseminate such contributions for future issues and volumes of our *Journal*.

References

[1] Anderson, M. and Seltzer, W. (2007). "Challenges to the Confidentiality of U.S. Federal Statistics, 1910-1965." *Journal of Official Statistics*, 23(1): 1–34.
URL http://www.jos.nu/Articles/abstract.asp?article=231001 3

[2] Bergstrom, T., Blume, L., and Varian, H. (1986). "On the Private Provision of Public Goods." *Journal of Public Economics*, 29: 25–49. 1

[3] Dalenius, T. (1977). "Towards a Methodology for Statistical Disclosure Control." *Statistik Tidskrift*, (5): 35–64. 2

[4] Duncan, G. T., Jabine, T. B., and de Wolf, V. A. (eds.) (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. National Academy of Sciences–National Research Council and Social Science Research Council. 3

[5] Dwork, C. (2006). "Differential Privacy." *33rd International Colloquium on Automata, Languages and Programming–ICALP*, II: 1–12. 2

[6] — (2009). "The Differential Privacy Frontier (Extended Abstract)." *Theory of Cryptography Conference*, 6: 496–502. 2

[7] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). "Calibrating Noise to Sensitivity in Private Data Analysis." *Theory of Cryptography Conference*, 3: 265–284. 2

[8] Federal Committee on Statistical Methodology (2005). "Statistical Policy Working Paper 22 (Second version): Report on Statistical Disclosure Limitation Methodology." NTIS PB94-165305.
URL http://www.fcsm.gov/working-papers/spwp22.html 2

[9] Fellegi, I. P. (1972). "On the Question of Statistical Confidentiality." *Journal of the American Statistical Association*, 67: 7–18. 2

[10] Stigler, G. J. (1980). "An Introduction to Privacy in Economics and Politics." *Journal of Legal Studies*, 9: 623–44. 1