Journal of Privacy and Confidentiality (2015–2016) **7**, Number 1, 1–2

In This Issue

Stephen E. Fienberg*
Editor-in-Chief, *Journal of Privacy and Confidentiality*

This issue of the *Journal of Privacy and Confidentiality* consists four papers on diverse topics related to privacy and confidentiality.

Record linkage is in some sense the antithesis of privacy protection because its raison d'être is that data from merged or linked files carry more information than is available from the files considered separately, and the linked files pose greater threats to privacy protection. Furthermore, linkage attacks represent the most obvious means for an intruder to gain access to information that should otherwise be kept private or confidential. Yet, record linkage remains at the forefront of approaches to produce high quality information from diverse administrative records without increasing the burden on respondents, and thus it lies at the heart of many of the activities of government statistical agencies such as the U.S. Bureau of the Census.

At the core of most of the current methods for linking the data on individuals from two separate files is a method due to Fellegi and Sunter (1969). One of the major challenges of the Fellegi-Sunter methodology and its variants and extensions is to make algorithms/methods scale as the size of the files to be linked grow. This is because the methodology requires examining all possible pairs of records form the two files. To date the statistical literature has addressed this scaling problem in ad hoc ways and has not taken into account the formal statistical implications of the ad hoc approaches. In a path-breaking paper, "Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering," Jared Murray provides a formal analytical framework for dealing with the most common ad hoc ways of making the Fellegi-Sunter approach scalable to deal with the linkage of substantial datasets. The technical details of the formalization rely on contingency table representations involving quasi-independence.

The primary tool for confidentiality protection of U.S. Census data and its derivatives over the past three decennial censuses has been the method of data swapping, e.g., see Fienberg and McIntyre (2005). Proposed originally by Reiss (1980) and Dalenius and Reiss (1982), the version of data swapping used by the Census Bureau has been extended in a variety of ways. Because the details of these extensions and the rate of swapping are not publicly available, it has been difficult to assess both the extent to which the methodology actually protects confidentiality of individual and household data, and the implications of the swapping for the utility of the resulting data in terms of their use in subsequent analyses.

In "The Effect of Data Swapping on Analyses of American Community Survey Data," Nicolas Kim provides the first systematic analysis of both aspects of data swapping in the context of of the American Community Survey, an ongoing byproduct of the

*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, mailto:fienberg@stat.cmu.edu.

U.S. decennial census.

In "Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multivariate Normal and Multiple Linear Regression Models," Martin Klein and Bimal Sinha develop a systematic approach to a targeted confidentiality problem. They develop a likelihood-based finite sample inference approach based on singly imputed partially synthetic data, when the original data follow either a multivariate normal or a multiple linear regression model and they apply their approach to data from the Census Bureaus Current Population Survey.

The final paper in this issue addresses privacy issues in the commercial sphere where retailers offer personalized advertisements (coupons) to individuals (consumers), but run the risk of strong reactions from consumers who want a customized shopping experience but are concerned about protecting their privacy. In "Designing Incentive Schemes for Privacy-Sensitive Users," Chong Huang and Lalitha Sankar propose a Markov decision process model to capture (i) different consumer privacy sensitivities via a time-varying state; (ii) different coupon types (action set) for the retailer; and (iii) the action-and-state-dependent cost for perceived privacy violations. For the simple case with two states ("Normal" and "Alerted"), two coupons (targeted and untargeted), and consumer behavior statistics known to the retailer, they derive an optimal coupon-offering strategy for a retailer that wishes to minimize its expected discounted cost.

References

Dalenius, T. and Reiss, S. P. (1982). Data swapping: A technique for disclosure control. *Journal of Statistical Planning*, 6: 73–85.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328): 1183–1210.

Fienberg, S. E. and McIntyre, J. (2005). Data swapping: Variations on a theme by Dalenius and Reiss. *Journal of Official Statistics*, 21(2): 309–323.

Reiss, S. P. (1980). Practical data-swapping: The first steps. In *IEEE Symposium on Security and Privacy*, 38–42.