Data Privacy Protection and the Conduct of Applied Research

Ruobin Gong, V. Joseph Hotz, and Ian M. Schmutte, editors

Introduction to Volume

Rapid improvements in computational power and coincident increases in the amount of data on individuals and firms that is available from both public and private sector sources have been a boon to empirical research in the social sciences. However, these developments have also created new challenges for protecting the privacy of information on individuals and businesses collected from surveys or censuses released to the public. By combining such data with that from external sources, such has commercial databases, and deploying computational tools that efficiently cross-tabulate the combined data, it is increasingly possible to breach the anonymity of individuals and businesses and their characteristics. Such breaches can violate the strong privacy protections statistical agencies and other data providers are required or pledged to uphold, compromising their mission of providing high quality data on the nation's people and economy essential for government decision-making and the production of valuable research.

Statistical agencies and other data providers have been implementing new procedures for protecting the privacy of the data they collect and release to the public and various research user communities. Two approaches to disclosure limitation have received the most attention: data constructed to meet formal privacy criteria, notably the differential privacy criterion (Dwork et al., 2006), and synthetic data methods (Rubin, 1993, Raghunathan, Reiter & Rubin, 2003). While the scientific literature has established some understanding about the theoretical and empirical properties of these methods for protecting individuals' privacy and on the accuracy of tabulations they produce, less well understood and resolved are the implications of using these data for conducting applied research. It remains unclear the extent to which a new privacy protection method applied to a certain data product would impact the estimation of and the inferences about what we shall refer to as "parameters of interest."

Similarly, while recognizing that tradeoffs exist between data privacy and data usability, less attention has been devoted to how to quantify and assess these tradeoffs. While research has established the conceptual foundations of these new approaches for protecting data privacy, important gaps exist in our understanding of disclosure risks in actual implementations. It is important to quantify the tradeoff between privacy risks and the degradation of data usability.

Finally, there has been more limited attention to defining a coherent and theoretically rigorous foundation for what standards data providers should adopt for protecting privacy and what tradeoffs individuals, firms and society are willing to make between any loss of their privacy and wider access to high quality data to support research and advance our understanding of a range of substantive and social issues.

This volume is a collection of new research findings and ideas that shed light on the current movement to advance data privacy protection methods and the impact these innovations have on the conduct of empirical analysis in economics, computer science and statistics using privacy-protected data. The authors participated in one of two related NBER conferences focused on Data Privacy Protection and the Conduct of Applied Research in 2023 and 2024. Authors were invited to write shorter articles to provide an introduction and overview of some of the key issues related to privacy protection of data used in social science research. In some cases, these chapters describe key findings of the research the authors presented at these conferences or on authors' other related work. Authors were asked to explain why their underlying research question is important and then describe the key findings, whether analytical, the result of empirical analysis, or the result of Monte Carlo analysis. Finally, authors were encouraged to discuss new strategies for protection of data, including those not yet in wide usage for social science data, and report on initial studies that assess alternative strategies for privacy protection and its impacts on applied research.

The volume is organized into five different sections. The chapters in Section I describe new approaches and standards for data privacy protection. Chapters 1 and 2 describe innovations in providing privacy-preserving access to administrative tax data. O'Hara et al. discuss the secure query system (SQS), and Bowen et al. consider the use of privacy enhancing technologies as part of a cross-institutional collaboration called the Safe Data Technologies Project. In chapter 3, Bailie et al. reassess swapping, a classic disclosure avoidance method employed in the 1990, 2000 and 2010 U.S. Decennial Censuses, and consider the extent to which it satisfies the requirements of differential privacy. In chapter 4, Vilhuber reviews the history and lessons from the Cornell Synthetic Data Server (SDS) that supported validation of statistical analyses with synthetic data. Vilhuber discusses the potential use of containers, a standalone and executable software package, as a replacement to validation server to provide superior proxy access to confidential data.

Section II focuses on the modernization of privacy protection for survey data. In chapter 5, Drechsler and Bailie discuss a few prominent challenges in adopting differential privacy for survey data, including the complex nature of data collection and processing, and notably the unique challenges associated with survey weights. Regarding the latter issue, Seeman and colleagues explore in chapter 6 differentially private survey-weighted estimation. Chapters 7 and 8 take a deep dive into the Survey of Income & Program Participation (SIPP). Stanley and Totty discuss the SIPP Synthetic Beta (SSB) in chapter 7, a synthetic data product with validation that provides users with tiered access to microdata while satisfying privacy and confidentiality requirements. Chapter 8 summarizes the findings of a recent report by a panel of the National Academy of Sciences, Engineering and Medicine. In this chapter, Hotz and Raghunathan review the challenges that the SIPP and other longitudinal surveys face in balancing confidentiality protection of the respondents with the usability of the data products to support research and policymaking and to inform the public.

Section III contains descriptions of methodological innovations in data analysis and statistical inference in a privacy-preserving manner. In chapter 9, Mukherjee et al. examine the feasibility

and efficacy of differentially private statistical inference methods for randomized control trials (RCTs). In chapter 10, Sakong and Zentefis describe the use of the method of simulated moment (MSM) estimation of a non-linear model of consumer flows to different bank branches with the privacy-protected data generated by a differential privacy algorithm. And in chapter 11, Lin and Kolaczyk discuss the use differentially private algorithms for linear regression for use in analyzing upstream linked datasets.

Section IV takes a broader stance and considers the implication of privacy protection in modern data governance. Ramon Sarmiento discusses in chapter 12 the regulatory challenges associated with "dark patterns," i.e., user interfaces and website practices designed to deceive or trick users into actions or release of personal identifying information (PII) that they would otherwise not release. Chapter 13 presents a study of one aspect of the EU's General Data Protection Regulation (GDPR) in which firms are required to disclose how they use information obtained from consumers and that the description of this information be "readable." The authors, Gambato, Ganglmair and Krämer, investigate the differential compliance of firms with the disclosure and readability requirements. Finally in chapter 14, Desrochers and Rancourt explore the evolving privacy protection consideration of Statistics Canada in developing guidelines for its use of financial data from private sector credit monitoring companies.

Section V tackles the difficult question about striking the balance between privacy protection and data usability and accuracy. As noted by Totty and Watson in chapter 15, statistical agencies like the U.S. Census Bureau face fundamental tension between its mandate to produce accurate and usable data while safeguarding the privacy and confidentiality of data subjects. To balance these two objectives requires measuring accuracy and privacy. The chapter describes some of the challenges for measuring the latter. Chapter 16 presents an analysis that tries to quantify the importance of confidential microdata for economic research by examining various publication and citation statistics for studies that use confidential data versus those that use publicly available data. The authors Stipanicic and Tranchero find that while studies using confidential data are more likely to be published in top economics journals and have higher citation counts, they also tend to be disproportionately authored by senior researchers at more elite institutions. Finally, in chapter 17 Dekel et al. describe a new way of measuring the association with greater degrees of privacy protection and the responsiveness of economic variables and behaviors: the "privacy elasticity of behavior." The authors summarize their earlier work on the rationale for this measure and summarize their use of this measure to assess privacy of public-good contributions in a lab experiment.