CHANCE

ISSN: 0933-2480 (Print) 1867-2280 (Online) Journal homepage: https://www.tandfonline.com/loi/ucha20

SPECIAL ISSUE: A New Generation of Statisticians Tackles Data Privacy

Fang Liu, Saki Kinney & Aleksandra (Seša) Slavkovic´

To cite this article: Fang Liu, Saki Kinney & Aleksandra (Seša) Slavkovic´ (2020) SPECIAL ISSUE: A New Generation of Statisticians Tackles Data Privacy, CHANCE, 33:4, 4-5, DOI: 10.1080/09332480.2020.1847945

To link to this article: https://doi.org/10.1080/09332480.2020.1847945

SPECIAL ISSUE: A New Generation of Statisticians Tackles Data Privacy

Fang Liu, Saki Kinney, and Aleksandra (Seša) Slavković

"What Should Privacy Mean to you?" was the heading of the last special issue of *CHANCE* magazine focused on data privacy and confidentiality, published 16 years ago (Volume 17, Issue 3). In that issue, the authors touched on disclosure limitation methods for contingency tables, synthetic data and remote servers, distributed analyses via secure computation, risk-utility tradeoffs, and the importance of providing confidentiality protection to assure user participation in censuses and surveys. Some ventured to foretell the future of privacy research, but nobody anticipated the emergence of the privacy framework known as differential privacy.

The new formalism, since its first appearance in 2006, has taken computer science and machine learning communities by storm and claims to offer an answer to the question of what privacy should mean to you. It has been implemented by Google, Apple, and the like, and has made its way into the official statistics world and confidentiality protection—the U.S. Census is leading the way by working to release the 2020 Census data using a differentially private algorithm. Many young statisticians (and we need more of them, and us) have joined the statistical data privacy field, realizing the necessity for transparency and sound statistical thinking as the statistical confidentiality protection field, which originated in the 1960s, has begun merging with formal privacy. We feature some of these statisticians and their work and ideas in this special issue of *CHANCE* on statistical data privacy.

In "How Statisticians Should Grapple with Privacy in a Changing Data Landscape," **Joshua Snoke** and **Claire Bowen** discuss the need for statisticians to become better-acquainted with disclosure avoidance methodologies, both traditional approaches and new. In recent years, much attention in statistical disclosure research has been on differential privacy, which

differs from traditional methods in that it provides a framework for quantifying disclosure risk based on methods from theoretical computer science. Differential privacy offers a great deal of promise, but many challenges must be overcome for it to meet the needs of researchers and data producers. Snoke and Bowen highlight these gaps and say that addressing them should not be left to a small group of specialists, but that protecting privacy should be standard practice in producing high-quality research.

If sharing data while protecting confidentiality weren't hard enough, **Michelle Pistner Nixon**, **Steven Nixon**, and **Roberto Molinari** describe the challenges faced by the U.S. Department of Defense (DoD) in sharing data when the data are considered especially sensitive in "Data of the Defense and the Defense of Data." While accessing classified data requires a security clearance, DoD also holds a large amount of unclassified data that is still quite difficult for them to share, even with their own contractors. This article highlights existing disclosure avoidance methodologies that DoD can use to facilitate sharing—differential privacy in particular—as well as unusual challenges faced by DoD, and suggests some steps that DoD can take.

In "Privacy-Preserving Algorithms: the Gain and the Loss," **Zhe Zhang** and **Linjun Zhang** provide a high-level overview of differentially private (DP) algorithms, clarifying the definition of the DP privacy guarantee with a focus on GWAS data. They also discuss the cost to statistical efficiency and utility, and link the positive properties of the privacy-preserving algorithms to notions of regularizations and adaptive data analysis, which are poised to be interesting future research directions.

Synthetic data methodology has received much attention in the statistical disclosure limitation community as a way of sharing public-use microdata files that will minimize disclosure of sensitive information but still offer a great deal of statistical utility. Many questions of how best to capture both the risk and utility are open. **Jingchen (Monika) Hu**, **Terrance Savitsky**, and **Matthew Williams** discuss a new proposal for "Risk-Weighted Data Synthesizers for Microdata Dissemination," as the title implies. Using a sample from the Consumer Expenditure Surveys, published by the U.S. Bureau of Labor Statistics, they give an example of how to fine-tune the utility-risk trade-off of synthetic data that would give a superior risk reduction outcome in comparison to an unweighted synthesizer. They conclude by pointing readers to some interesting new directions tied to formal privacy guarantees.

Marco Avella-Medina argues that private data analysis provides fertile ground for revisiting robust statistics ideas and lessons, and discusses how differential privacy is naturally connected to notions of robustness. He provides examples of how to make robust estimators meet privacy constraints and discusses the efficiency of private estimators and robustness of private estimators. In summary, robust statistics could provide the right set of tools to tackle some of the data privacy problems and help to bridge the gap between the modern work on privacy and statistics.

Jordan Awan, **Matthew Reimherr**, and **Aleksandra (Seša) Slavković** discuss privacy-preserving nonparametric statistics. How to optimize the accuracy of the privatized estimator while still offering formal privacy guarantees is a challenging problem for nonparametric statistics, since many nonparametric methods are very sensitive to changes in individual information. They explore privacy-preserving density estimation and nonparametric regression, and highlight the challenges of satisfying differential privacy while maintaining statistical utility. They also present and discuss privacy challenges for some exciting new problems in nonparametric statistics.

We hope that these six articles will give you some theoretical, methodological, and applied flavors of the past, present, and future research into statistical data privacy. We hope that you and your colleagues will join us in thinking about and contributing to these and many related problems. The sound data privacy methodology built on solid statistical principles is needed to support sound policy and decision-making.

We are thankful to our young colleagues for sharing their work with us and you, to the ASA Privacy and Confidentiality Committee for supporting this effort, and to *CHANCE* for giving us the opportunity to feature these works.


Liu


Kinney


Slavković