Statistica Sinica **35** (2025), 533-537

Foreword

DATA PRIVACY: OVERVIEW

Data privacy is a key concern of our increasingly digital world, where vast amounts of personal information are collected, stored, and analyzed daily. As organizations and researchers seek to extract valuable insights from data, the need to protect individual privacy has never been more pressing. Differential privacy (DP), a powerful mathematical framework for quantifying and limiting privacy loss, has emerged as a leading approach to addressing these challenges.

The concept of differential privacy, introduced in Dwork et al. (2006a,b), provides a rigorous foundation for privacy-preserving data analysis. It offers strong privacy guarantees by adding carefully calibrated noise to data or query results, ensuring that the presence or absence of any individual's data has a negligible impact on the statistical output. This approach allows for meaningful statistical analysis while protecting individual privacy, striking a balance between data utility and confidentiality. Since its inception, DP has revolutionized the field of privacy-preserving data analysis and more broadly the area of statistical data privacy (Slavković and Seeman, 2023), opening up new avenues for research and practical applications. It has been adopted by major technology companies (Ding, Kulkarni and Yekhanin, 2017; Kenthapadi and Tran, 2018; Gadotti et al., 2022; Cormode et al., 2018) and government agencies, most visibly by the U.S. Census Bureau for the 2020 census (Abowd et al., 2022), demonstrating its real-world impact and importance.

This special issue celebrates the significant advancements in DP and broader data privacy research over the past two decades. The eight featured articles from authors around the world represent the rich spectrum of current work in data privacy, showcasing both theoretical developments and practical applications. They cover privacy-preserving synthetic data generation, statistical estimation, inference, and statistical disclosure risk assessment based on privacy-preserving synthetic data, privacy-preserving confidence interval construction and hypothesis testing, and differentially private optimization.

Synthetic data is important in protecting individual privacy while allowing for useful research and policy analysis, however, Awan and Cai find that existing methods, particularly the parametric bootstrap approach, lead to inconsistent synthetic data with inefficient estimators. To address this problem, they propose a new method called "one-step synthetic data", which adds an extra step

to the parametric bootstrap. This approach is designed to be widely applicable to various parametric models, easily implemented, and computationally efficient. It allows for both partially synthetic datasets (preserving summary statistics without formal privacy methods) and fully synthetic data satisfying differential privacy. The authors demonstrate that their method preserves efficient estimators with asymptotically negligible error and allows for distribution convergence even with slight parameter differences. This new approach aims to overcome the limitations of previous synthetic data generation techniques by offering a more versatile, easily implemented, and computationally efficient solution while maintaining a balance between data utility and privacy protection.

Hu, Williams, and Savitsky introduce a new method that embeds any Bayesian model used for synthetic data generation into a DP mechanism. The authors propose a censored likelihood approach that induces upper and lower bounds based on the desired level of ε-DP guarantee, and show that this innovative approach is superior to traditional methods, such as the perturbed histogram mechanism, in balancing data utility and privacy protection. The method incorporates a vector-weighted pseudo posterior mechanism within the censoring mechanism to minimize distortion in the posterior distribution. This combined approach allows for the generation of synthetic data with either a weaker asymptotic DP guarantee and higher utility, or a stronger, non-asymptotic DP guarantee with slightly reduced utility.

Nombo and Charest tackle the problem of how to properly perform statistical inference with differentially private synthetic (DIPS) datasets by investigating the applicability of combining rules, originally designed for standard synthetic datasets, to DIPS datasets. They propose to empirically test whether these combining rules can provide valid inference for various differentially private synthesizers, including those based on statistical ideas (Bayesian networks, copulas) and deep learning models (e.g., GANs). They show empirically that this approach can offer accurate inference under certain conditions, such as when a method produces unbiased or minimally biased point estimates and the between-variance sufficiently captures variability due to DP. The authors note that this methodology works well for DPGAN and COPULA-SHIRLEY methods, and sometimes for PATE-GAN method, making it applicable to a wider range of models than previously thought.

On a theoretical side, Györfi and Kroll address the challenge of estimating regression functions from synthetic data within the context of local DP. The authors present a new partitioning estimate for regression functions, and provide a thorough theoretical analysis, including the derivation of a convergence rate for the excess prediction risk over Hölder classes and a matching lower bound. A key contribution of this work is that it eliminates the need for the strong density

assumption on the design distribution, which has been a requirement in previous research on this topic.

Kazan and Reiter introduce Bayesian methods for assessing statistical disclosure risks in differentially private data with a hierarchical structure, under zero-concentrated DP. These methods compute posterior probabilities of disclosure based on released counts and assumptions about adversaries' knowledge. The authors apply their approach to differentially private data releases from the 2020 U.S. decennial census and perform empirical studies using public, individual-level data from the 1940 U.S. decennial census. They explore how disclosure risks vary with privacy parameters and released counts, aiming to provide insights into potential privacy risks.

Covington, He, Honaker, and Kamath discuss a challenge in the application of DP– DP algorithms typically require the user to, without looking at the data, specify a domain to which the data will be clipped. The paper introduces a novel framework: a general-purpose meta-algorithm that converts non-private estimators into DP estimators while maintaining unbiasedness and producing valid confidence intervals. This framework combines the Bag of Little Bootstraps algorithm (Kleiner et al., 2014) and a modified version of the CoinPress private mean estimation algorithm (Biswas et al., 2020) by precision weighting techniques. This approach addresses the difficulty of specifying data bounds without introducing substantial error. This work is positioned as a step towards making DP more practical for applied research, offering a method that allows for conducting statistical inference without introducing bias and is potentially applicable to a wide range of estimators.

Peña and Barrientos propose a novel method combining the subsample and aggregate technique with randomized response to create differentially private versions of existing hypothesis tests. This approach is shown to be conceptually simple, widely applicable, and capable of achieving high privacy levels and low type-I error rates simultaneously. The method is particularly effective for tests with low significance levels, addressing concerns related to the replication crisis in scientific research. The authors demonstrate that their approach, which outputs a binary decision rather than p-values or Bayes factors, can be more practical and potentially more powerful in certain scenarios. Through extensive simulation studies, they illustrate the performance of their method in implementing differentially private versions of various statistical tests, including goodness-of-fit tests, the one-sample Wilcoxon test, and the Kruskal-Wallis test.

Xie, Pietrosanu, Liu, Tu, Jiang, and Kong discuss challenges in privacy-preserving convex optimization, particularly for regularized problems with heavy-tailed data. They propose three novel differentially private algorithms for regularized stochastic convex optimization problems with heavy-tailed responses. The

first algorithm is a vanilla (ε, δ)-DP approach applicable to a wide range of data distributions. The second algorithm utilizes a robust mean estimator to achieve an improved upper bound on the population excess risk under certain assumptions. The third algorithm incorporates a different robust mean estimator, further improving the upper bound with weaker assumptions. These methods are shown to be theoretically and empirically superior to existing approaches, especially in handling non-smooth regularizers and heavy-tailed data distributions. The authors demonstrate that their algorithms address limitations in current literature, which often assumes Lipschitz continuity of the loss function, and provide robust solutions for privacy-preserving regularized convex optimization in various real-world scenarios.

As we continue to grapple with the complexities of protecting individual privacy in the age of big data and machine learning, the work presented in this special issue contributes to our understanding and ability to develop robust, privacy-preserving data analysis methods. These advancements are crucial for maintaining public trust, complying with evolving privacy regulations, and unlocking the full potential of data-driven innovations while respecting fundamental privacy rights.

We hope that this collection of articles will inspire further research, foster interdisciplinary collaboration, and contribute to the ongoing development of privacy-preserving technologies that will shape the future of data analysis and privacy protection.

Jing Lei[1]
Aleksandra Slavković[2]
Linjun Zhang*[3]
(Editors, in alphabetical order, for this special issue)

1. Carnegie Mellon University
2. The Pennsylvania State University
3. Rutgers Uinversity
* Corresponding author. E-mail: lz412@stat.rutgers.edu

References

Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C. et al. (2022). The 2020 census disclosure avoidance system TopDown algorithm. *Harvard Data Science Review* **Special Issue 2**. DOI: 10.1162/99608f92.529e3cb9.

Biswas, S., Dong, Y., Kamath, G. and Ullman, J. (2020). CoinPress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems* **33**, 14475–14485.

Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D. and Wang, T. (2018). Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*, 1655–1658.

Ding, B., Kulkarni, J. and Yekhanin, S. (2017). Collecting telemetry data privately. *Advances in Neural Information Processing Systems* **30**, 3574–3583.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I. and Naor, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006* (Edited by S. Vaudenay), 486–503. Springer, Berlin, Heidelberg.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006b). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography* (Edited by S. Halevi and T. Rabin), 265–284. Springer, Berlin, Heidelberg.

Gadotti, A., Houssiau, F., Annamalai, M. S. M. S. and de Montjoye, Y.-A. (2022). Pool inference attacks on local differential privacy: Quantifying the privacy guarantees of Apple's count mean sketch in practice. In *31st USENIX Security Symposium (USENIX Security 22)*, 501–518.

Kenthapadi, K. and Tran, T. T. (2018). PriPeARL: A framework for privacy-preserving analytics and reporting at linkedin. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2183–2191.

Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **76**, 795–816.

Slavković, A. and Seeman, J. (2023). Statistical data privacy: A song of privacy and utility. *Annual Review of Statistics and Its Application* **10**, 189–218.