Journal of Privacy and Confidentiality
Vol. 14 (3) 2024

Noisy measurement workshop 2022

Submitted July 2024
Published August 2024

FOREWORD FOR THE COLLECTION OF PAPERS FROM THE WORKSHOP ON THE ANALYSIS OF CENSUS NOISY MEASUREMENT FILES AND DIFFERENTIAL PRIVACY

JÖRG DRECHSLER[1], RUOBIN GONG[2], WEIJIE SU[3], AND LINJUN ZHANG[2]

[1]Institute for Employment Research, 90478 Nuremberg, Germany, LMU Munich, 80539 Munich, and University of Maryland, College Park, MD 20742
e-mail address: joerg.drechsler@iab.de

[2]Rutgers University, Piscataway, NJ 08854
e-mail address: {ruobin.gong; linjun.zhang}@rutgers.edu

[3]University of Pennsylvania, Philadelphia, PA 19104
e-mail address: suw@wharton.upenn.edu

Abstract. The Noisy Measurement Files (NMFs) are a set of differentially private intermediate statistical queries that underlie the official 2020 U.S. Decennial Census data products. While they do not satisfy certain constraints that the official data products must obey, they enjoy crucial theoretical qualities, including unbiasedness and transparency, that support principled statistical analysis and uncertainty quantification. This issue of the *Journal of Privacy and Confidentiality* collects three papers that result from the *Workshop on the Analysis of Census Noisy Measurement Files and Differential Privacy*, held at the DIMACS Center of Rutgers University in April 2022, where experts from diverse disciplines gathered to discuss key challenges in the use of NMFs to support social research and policy decisions. The U.S. Census Bureau released the NMFs in 2023. Since their release, the NMFs have shed light on methodological questions relating to the design of the Census Disclosure Avoidance System (DAS).

The 2020 Decennial Census of the United States employs a novel disclosure avoidance system that uses differential privacy to conduct privacy-loss accounting (Abowd et al., 2022). The TopDown Algorithm, which supports the Redistricting Data Summary File (P.L. 94-171) and the Demographic and Housing Characteristic File, works by first generating a full set of statistical queries (called the Noisy Measurements File, or NMF, via additive noise infusion that satisfy zero-concentrated differential privacy (Bun and Steinke, 2016), implemented via the discrete Gaussian Mechanism (Canonne et al., 2020). In a second step, it post-processes the NMFs to enforce invariants, edit constraints, and maintain structural zeros to create the end data product for public release. Invariants describe a set of aggregate statistics such as total housing units at the Census block level, which have to be reported unperturbed. Edit constraints are constraints enforced on the noisy data to ensure consistency, for example, by ensuring that counts at the state level add up to the counts at the country level. Constraints

Key words and phrases: unbiasedness; uncertainty quantification; social and policy research.

JPrivacy Confidentiality

www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.v0.i0.934

also ensure that the released tables do not contain any negative counts. Finally, structural zeros are implausible combinations of attributes such as age of natural child being larger than the age of its mother. The post-processing step ensures that these structural zeros are maintained in the released tables.

Despite being the intermediate output of the TopDown, the NMFs possess crucial advantages over the post-processed end data products. Post-processing can deliver consistency and enhanced usability in many use cases, but induces bias which may undermine data quality and the soundness of subsequent policy decisions.[1] Moreover, due to the data-dependent nature of the post-processing optimization procedure, bias due to post-processing may be difficult to quantify (Zhu et al., 2021). On the other hand, even though the NMFs may not look facially valid due to the presence of negative counts and logical inconsistency, they are generated via a highly transparent process and are theoretically unbiased. These qualities make the NMFs the most straightforward solution to the issues that arise due to post-processing, allowing researchers to quantify post-processing biases and to conduct principled downstream statistical analysis with correct uncertainty quantification. Moreover, since the NMFs are differentially private themselves, releasing them to the public in conjunction with the post-processed data files may not further degrade the privacy guarantee afforded to the latter.

On April 28-29, 2022, the *Workshop on the Analysis of Census Noisy Measurement Files and Differential Privacy* was held at the DIMACS Center of Rutgers University. The Workshop brought together experts from demography, public policy, social sciences, statistics, and computer science to address key challenges in the use of the differentially private Census NMFs to support social research and policy decisions. The workshop participants deliberated the following questions: A) What are the important use cases for which the analysis of noisy measurement files can support reliable research conclusions and trustworthy policy decisions? B) What are the statistical and computational methodologies needed to support the analyses of the noisy measurement files, and what are some challenges to their design and practical implementation? C) What are some inference methodologies that can support the analysis of minimally post-processed data files, or a combination of noisy measurements and post-processed data sources? D) Looking ahead, what are some good design principles for privacy-protected data files for public release, such that they are conducive to principled analysis?

This issue of the *Journal of Privacy and Confidentiality* curates three papers that resulted from the work presented at the Workshop. Cumings-Menon et al. (2024) discusses the technical design of the *geographic spine*, the crucial structural element that determines the hierarchical organization of the 2020 Census data products. Seeman (2024) calls for better alignment between academic and practical aspects of DP, highlighting political and communication issues in current approaches. They argue for integrating both theoretical and empirical perspectives to make DP theory more applicable to real-world data practices. Snoke et al. (2024) describes applying differential privacy to a complex dataset for statistical

[1] For example, Asian Americans Advancing Justice-Mexican American Legal Defense and Educational Fund (2021) observed that the post-processed counts in an early version of the demonstration data released during testing exhibited an apparent "transfer" of population from urban to rural areas, an issue that may prevent the accurate redistricting. Gao et al. (2022) reported that the TopDown algorithm associated larger counts with negative errors and smaller counts with positive errors, when the total count is held as invariant. In general, systematic bias and under-counting in the Census public data products may disproportionally impact marginalized groups (Kirkendall et al., 2020).

analysis, finding fundamental incompatibilities with current practices. The authors suggest that resolving these issues will require significant changes to either statistical data analysis or differential privacy approaches, or both.

The U.S. Census Bureau released the NMFs on June 15, 2023 (U.S. Census Bureau, 2023a) and the demographic and housing characteristics NMFs on October 23, 2023 (U.S. Census Bureau, 2023b). Since their release, the NMFs have been used for methodological research, for example to evaluate the bias and the noise of the DAS (Kenny et al., 2024) or to develop procedures for obtaining valid confidence intervals for arbitrary tabulations from the NMFs (Cumings-Menon, 2024). The full potential for the NMFs to support substantive social and policy research is yet to be realized. As a novel data product that has only been released about a year ago, the NMFs are an unfamiliar source of information for researchers, and their accessibility is hindered by both complexity (with trillions of statistics) and the currently limited documentation; see the discussions in McCartan et al. (2023); Abowd (2024) and McCartan et al. (2024). We hope to see more applied research that utilize the NMFs in the days to come.

Acknowledgment

As organizers of the Workshop, RG, WS, and LZ gratefully thank the Alfred P. Sloan Foundation for providing financial support and Rutgers University DIMACS Center for logistic assistance. They also thank all participants of the Workshop for the stimulating, cross-disciplinary conversation that contributed to advancing the frontier of data privacy protection in support of social and policy research.

References

J. M. Abowd. Noisy Measurements Are Important; The Design of Census Products Is Much More Important. *Harvard Data Science Review*, 6(2), apr 30 2024. https://doi.org/10.1162/99608f92.79d4660d.

J. M. Abowd, R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns, D. Kifer, P. Leclerc, A. Machanavajjhala, et al. The 2020 Census disclosure avoidance system TopDown algorithm. *Harvard Data Science Review*, 2, 2022. https://doi.org/10.1162/99608f92.529e3cb9.

Asian Americans Advancing Justice-Mexican American Legal Defense and Educational Fund. Preliminary report: Impact of differential privacy the 2020 census on latinos, asian americans and redistricting. Technical report, April 2021. https://www.advancingjustice-aajc.org/sites/default/files/2021-04/MALDEF%20AAJC%20Differential%20Privacy%20Preliminary%20Report%20FINAL%204.5.2021.pdf.

M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016. https://doi.org/10.1007/978-3-662-53641-4_24.

C. L. Canonne, G. Kamath, and T. Steinke. The discrete Gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33:15676–15688, 2020. https://dl.acm.org/doi/10.5555/3495724.3497039.

R. Cumings-Menon. Full-information estimation for hierarchical data, 2024. https://arxiv.org/abs/2404.13164.

R. Cumings-Menon, J. M. Abowd, R. Ashmead, D. Kifer, P. Leclerc, J. Ocker, M. Ratcliffe, and P. Zhuravlev. Geographic spines in the 2020 census disclosure avoidance system. *Journal of Privacy and Confidentiality*, 2024. https://doi.org/10.29012/jpc.875.

J. Gao, R. Gong, and F.-Y. Yu. Subspace differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3986–3995, 2022. https://aaai.org/papers/03986-subspace-differential-privacy/.

C. T. Kenny, C. McCartan, S. Kuriwaki, T. Simko, and K. Imai. Evaluating bias and noise induced by the US Census Bureau's privacy protection methods. *Science Advances*, 10 (18):eadl2524, 2024. https://doi.org/10.1126/sciadv.adl2524.

N. J. Kirkendall, C. F. Citro, and D. L. Cork. *2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop*. National Academies Press, 2020. https://doi.org/10.17226/25978.

C. McCartan, T. Simko, and K. Imai. Making differential privacy work for Census data users. *Harvard Data Science Review*, 5(4), 2023. https://doi.org/10.1162/99608f92.c3c87223.

C. McCartan, T. Simko, and K. Imai. Rejoinder: We Can Improve the Usability of the Census Noisy Measurements File. *Harvard Data Science Review*, 6(2), may 7 2024. https://doi.org/10.1162/99608f92.f9f4b9a4.

J. Seeman. Bettery privacy theorists for better data stewards. *Journal of Privacy and Confidentiality*, 2024. https://doi.org/10.29012/jpc.865.

J. Snoke, C. M. Bowen, A. R. Williams, and A. F. Barrientos. Incompatibilities between current practices in statistical data analysis and differential privacy. *Journal of Privacy and Confidentiality*, 2024. https://doi.org/10.29012/jpc.872.

U.S. Census Bureau. 2020 census redistricting noisy measurement file (NMF). https://www.census.gov/newsroom/press-releases/2023/2020-redistricting-noisy-measurement-file.html, 2023a.

U.S. Census Bureau. Census Bureau releases 2020 census DHC noisy measurement file. https://www.census.gov/newsroom/press-releases/2023/2020-census-dhc-noisy-measurement-file.html, 2023b.

K. Zhu, P. Van Hentenryck, and F. Fioretto. Bias and variance of post-processing in differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11177–11184, 2021. https://aaai.org/papers/11177-bias-and-variance-of-post-processing-in-differential-privacy/.