"HAVING CONFIDENCE IN MY CONFIDENCE INTERVALS": HOW DATA USERS ENGAGE WITH PRIVACY-PROTECTED WIKIPEDIA DATA

Harold Triedman[*1], *Jayshree Sarathy*[†2], *Priyanka Nanayakkara*[‡], *Rachel Cummings*[§], *Gabriel Kaptchuk*[¶], *Sean Kross*[‖], *Elissa M. Redmiles*[**]

[*]Cornell Tech, [†]Northeastern University, [‡]Harvard University, [§]Columbia University, [¶]University of Maryland, [‖]Fred Hutch Cancer Center, [**]Georgetown University

As data has become a critical and valuable resource, organizations have sought to open up access to their data for external researchers and the public. At the same time, privacy researchers have demonstrated potential for attacks that identify individual data subjects using aggregate statistics [5, 12]. To mitigate these threats, organizations have started to incorporate *privacy noise* when generating datasets for publication [4, 9]. Some noising techniques, such as rounding [3], have been used for decades; others like differential privacy (DP) [7] are more modern developments. The goal of these techniques is to protect the privacy of individual data subjects while preserving overall patterns in the data.

While noising techniques are well-studied from technical perspectives, real-world data releases that include privacy noise, such as with the 2020 US Census and Facebook's Social Science One initiative, have generated dissatisfaction from data users (e.g. [18, 15]). The rocky reception to these data releases demonstrates the privacy community's gaps in understanding data users' perceptions of noisy data and best practices for communication. Prior work has focused on communication of privacy guarantees to data subjects (e.g., [1, 19, 2, 11, 16]), guidance for data curators who wish to apply privacy protections when releasing sensitive data (e.g. [6]), and usability interventions for developers and analysts who implement privacy-preserving mechanisms (e.g. [14, 13, 17, 10]). However, the experiences of data users who use *already*-noised datasets are less studied.

In this work, we explore the perspectives of data users contending with noise in data releases using datasets from the Wikimedia Foundation (WMF), which has released sensitive data using two different methods of privacy noise: rounding and DP. WMF is the only organization we are aware of[3] that has released the same dataset using different noising methods, which provides a natural opportunity to understand how data users engage with heuristic versus formal privacy noise techniques in datasets. We ask: How do data users perceive, interact with, and interpret noise injected into data for privacy protection? Does the nature of the noise (rounding vs. DP) impact data users' engagement with the data?

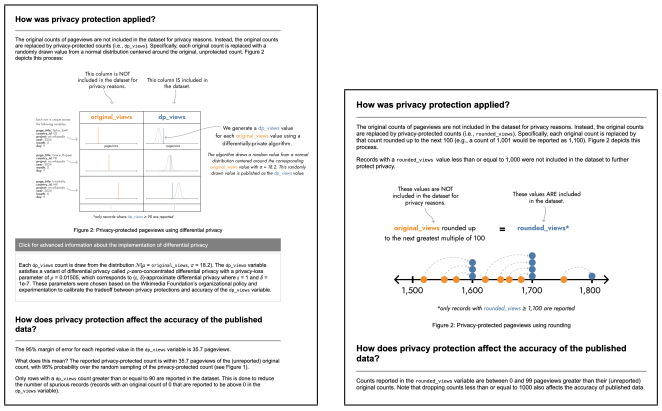

Figure 1. Screenshots of DP and Rounding documentation for Pageviews by Country Dataset.

To explore these questions, we conducted a task-based analysis study with 15 participants experienced in data science. Each study session included a contextual inquiry using data analysis tasks, as well as a semi-structured interview. We provide participants with dataset documentation, which is critical for enabling data users to responsibly work with data [8]. We design documentation for each of the datasets, incorporating feedback from five experts in privacy communication, and provide these documents to participants during the study (see Fig. 1).

We find that participants were well-versed in existing sources of error in Wikipedia datasets, which allowed them to think through the impact of additional noise for privacy purposes. They were easily able to understand the process of rounding, but felt better equipped to devise simulations and develop empirical confidence intervals using the DP data. For both rounding and DP, participants struggled to understand how uncertainty would scale across multiple perturbed data points. Participants had mixed and surprising perceptions about the privacy protections

[1]Email: hjt36@cornell.edu.

[2]Email: j.sarathy@northeastern.edu.

[3]The US Census Bureau used two different methods of privacy protection, swapping and DP, for its 2010 decennial census and 2020 decennial census, respectively. These two datasets are similar but not identical.

offered by rounding and DP, illuminating data users' conceptual relationships between dataset utility and perceived privacy. They also had contextually-dependent preferences for which dataset to use when communicating their results to various audiences. Based on these findings, we offer recommendations to help data users work effectively with privacy-noised datasets, such as building tools to automatically compute confidence intervals and track uncertainty in downstream analyses. We point to the need to better understand how different audiences perceive the relationship between privacy and accuracy of noisy data.

References

[1] Brooke Bullek et al. "Towards Understanding Differential Privacy: When Do People Trust Randomized Response Technique?" In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, May 2017, pp. 3833–3837. DOI: 10.1145/3025453.3025698.

[2] Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. ""I need a better description": An Investigation Into User Expectations For Differential Privacy". In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. Nov. 2021, pp. 3037–3052. DOI: 10.1145/3460120.3485252.

[3] Tore Dalenius and Steven P Reiss. "Data-swapping: A technique for disclosure control". In: *Journal of Statistical Planning and Inference* 6.1 (1982), pp. 73–85.

[4] Damien Desfontaines. *A List of Real-World Uses of Differential Privacy*. https://desfontain.es/blog/real-world-differential-privacy.html. 2021.

[5] Irit Dinur and Kobbi Nissim. "Revealing Information While Preserving Privacy". In: *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. ACM, June 2003, pp. 202–210. DOI: 10.1145/773153.773173.

[6] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. "Differential Privacy in Practice: Expose your Epsilons!" In: *Journal of Privacy and Confidentiality* 9.2 (Oct. 2019). DOI: 10.29012/jpc.689.

[7] Cynthia Dwork et al. "Calibrating noise to sensitivity in private data analysis". In: *Theory of Cryptography Conference*. Springer. 2006, pp. 265–284.

[8] Timnit Gebru et al. "Datasheets for Datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[9] Priyanka Nanayakkara, Elena Ghazi, and Salil Vadhan. "Practitioners' Perspectives on a Differential Privacy Deployment Registry". In: *arXiv preprint arXiv:2509.13509* (2025).

[10] Priyanka Nanayakkara et al. "Measure-Observe-Remeasure: An Interactive Paradigm for Differentially-Private Exploratory Analysis". In: *2024 IEEE Symposium on Security and Privacy*. IEEE. 2024, pp. 1047–1064.

[11] Priyanka Nanayakkara et al. "What Are the Chances? Explaining the Epsilon Parameter in Differential Privacy". In: *32nd USENIX Security Symposium*. 2023, pp. 1613–1630.

[12] Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy*. IEEE. 2008, pp. 111–125.

[13] Ivoline C Ngong et al. "Evaluating the Usability of Differential Privacy Tools with Data Practitioners". In: *Twentieth Symposium on Usable Privacy and Security*. 2024, pp. 21–40.

[14] Liudas Panavas et al. "Investigating the Visual Utility of Differentially Private Scatterplots". In: *IEEE Transactions on Visualization and Computer Graphics* 30.8 (2023), pp. 5370–5385.

[15] Craig Silverman. *Funders Are Ready To Pull Out Of Facebook's Academic Data Sharing Project*. BuzzFeed News, Aug. 2019. URL: https://www.buzzfeednews.com/article/craigsilverman/funders-are-ready-to-pull-out-of-facebooks-academic-data.

[16] Mary Anne Smart et al. "Models Matter: Setting Accurate Privacy Expectations for Local and Central Differential Privacy". In: *Proceedings on Privacy Enhancing Technologies* 2025 (4 2025), pp. 653–678. DOI: 10.56553/popets-2025-0150.

[17] Patrick Song et al. "'I inherently just trust that it works': Investigating Mental Models of Open-Source Libraries for Differential Privacy". In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW2 (2024), pp. 1–39. DOI: 10.1145/36870.

[18] Gus Wezerek and David Van Riper. "Changes to the Census Could Make Small Towns Disappear". In: (June 2020). URL: https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html.

[19] Aiping Xiong et al. "Towards Effective Differential Privacy Communication for Users' Data Sharing Decision and Comprehension". In: *2020 IEEE Symposium on Security and Privacy*. IEEE. 2020, pp. 392–410. DOI: 10.1109/SP40000.2020.00088.