CAN SWAPPING BE DIFFERENTIALLY PRIVATE? A REFRESHMENT STIRRED, NOT SHAKEN

James Bailie[†], *Ruobin Gong*[‡], *Xiao-Li Meng*[†]

[†]Statistics Department, Harvard University; [‡]Statistics Department, Rutgers University

To directly address the title's query, an answer must necessarily presuppose a precise specification of differential privacy (DP) [8]. Indeed, as there are many different formulations of DP [6] – which range, both qualitatively and quantitatively, from being practically and theoretically vacuous to providing gold-standard privacy protection [7] – a "yes/no" answer to the question, "is X differentially private?" is not particularly informative and is likely to lead to confusion or even dispute, especially when the presupposed DP specification is not clearly spelt out and fully comprehended.

A true answer to the title's query must therefore be predicated upon an understanding of how formulations of DP differ, which is best explored, as is often the case, by starting with their unifying commonality. DP specifications are, in essence, Lipschitz conditions on the data-release mechanism. The core philosophy of DP is thus to manage relative privacy loss by limiting the rate of change of the variations in the noise-injected output statistics when the confidential input data are (counterfactually) altered. Hence, DP conceives of privacy protection specifically as control over the Lipschitz constant – i.e. over this rate of change; and different DP specifications correspond to different choices of how to measure input alterations and output variations, in addition to the choice of how much to control this rate of variations-to-alterations. Following this line of thinking through existing literature [9, 3, 11, 5, 4, 17, 10, 12] leads to five necessary building blocks for a DP specification. They are, in order of mathematical prerequisite, the protection domain, the scope of protection, the protection unit, the standard of protection, and the intensity of protection. In simple terms, these are respectively the "who", "where", "what", "how", and "how much" questions of DP.

Under this framework, we consider DP's applicability in scenarios like the US Census where the disclosure of certain aggregates is mandated by the US Constitution [1]. We design and analyze a data swapping method, called the Permutation Swapping Algorithm (PSA), which is reminiscent of the statistical disclosure control (SDC) procedures employed in several US Decennial Censuses before 2020 [13]. For comparative purposes, we are also interested in the principal SDC method of the 2020 Census, the TopDown algorithm (TDA) [1], which melds the DP specification of [3] (zero-concentrated DP) with Census policy and constitutional mandates.

We analyze the DP properties of both data swapping and the TDA. Both [3]'s specification and the original ϵ-DP specification of [8] demand that no data summary is disclosed without noise – which is impossible for swapping methods as they inherently preserve, and hence disclose, some margins; and is also impossible for the TDA since it too keeps some counts invariant. Therefore, for the same reasons that the TDA cannot satisfy the DP specification of [3], data swapping cannot satisfy the original ϵ-DP specification of [8]. On the other hand, we establish that the PSA is ϵ-DP, subject to the invariants it necessarily induces and we show how the privacy-loss budget ϵ is determined by the swapping rate and the maximal size of the swapping strata. We also prove a DP specification for the TDA, by subjecting [3]'s specification to the TDA's invariants. Drawing a parallel, we assess the privacy budget for the PSA in the hypothetical situation where it was adopted for the 2020 Census.

Our overarching ambition is three-fold: firstly, to leverage the merits of DP, including its mathematical assurances and algorithmic transparency, without sidelining the advantages of classical SDC [18]; secondly, to unveil the nuances and potential pitfalls in employing DP as a theoretical yardstick for SDC procedures; and thirdly, to build connections between social and technical conceptualizations of privacy [16, 14, 19, 15, 2] by outlining real-world considerations behind the five building blocks. By spotlighting data swapping, we aspire to stimulate rigorous evaluations of other SDC techniques, and demonstrate that the privacy-loss budget ϵ is merely one of five components of DP.

References

[1] John Abowd et al. "The 2020 Census Disclosure Avoidance System TopDown Algorithm". In: *Harvard Data Science Review* Special Issue 2 (June 2022). DOI: `10.1162/99608f92.529e3cb9`.

[2] James Bailie and Ruobin Gong. "The Five Safes as a Privacy Context". In: *The 5th Annual Symposium on Applications of Contextual Integrity.* Toronto, Canada, 2023.

[3] Mark Bun and Thomas Steinke. "Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds". In: *Theory of Cryptography.* Ed. by Martin Hirt and Adam Smith. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2016, pp. 635–658. ISBN: 978-3-662-53641-4. DOI: `10.1007/978-3-662-53641-4_24`.

[4] Mark Bun et al. "Controlling Privacy Loss in Sampling Schemes: An Analysis of Stratified and Cluster Sampling". In: *Foundations of Responsible Computing (FORC 2022).* June 2022, p. 24.

[5] Konstantinos Chatzikokolakis et al. "Broadening the Scope of Differential Privacy Using Metrics". In: *Privacy Enhancing Technologies.* Ed. by Emiliano De Cristofaro and Matthew Wright. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 82–102. DOI: `10.1007/978-3-642-39077-7_5`.

[6] Damien Desfontaines and Balázs Pejó. "SoK: Differential privacies". In: vol. 2020. 2. 2020, pp. 288–313.

[7] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. "Differential Privacy in Practice: Expose Your Epsilons!" In: *Journal of Privacy and Confidentiality* 9.2 (Oct. 2019). ISSN: 2575-8527. DOI: `10.29012/jpc.689`.

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

[9] Cynthia Dwork et al. "Our Data, Ourselves: Privacy Via Distributed Noise Generation". In: *Advances in Cryptology - EUROCRYPT 2006.* Ed. by Serge Vaudenay. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 486–503. ISBN: 978-3-540-34547-3. DOI: `10.1007/11761679_29`.

[10] Michael Hay et al. "Accurate Estimation of the Degree Distribution of Private Networks". In: *2009 Ninth IEEE International Conference on Data Mining.* Dec. 2009, pp. 169–178. DOI: `10.1109/ICDM.2009.11`.

[11] Daniel Kifer and Ashwin Machanavajjhala. "No Free Lunch in Data Privacy". In: *Proceedings of the 2011 International Conference on Management of Data - SIGMOD '11.* Athens, Greece: ACM Press, 2011, pp. 193–204. ISBN: 978-1-4503-0661-4. DOI: `10.1145/1989323.1989345`.

[12] Daniel Kifer and Ashwin Machanavajjhala. "Pufferfish: A framework for mathematical privacy definitions". In: *ACM Transactions on Database Systems (TODS)* 39.1 (2014), pp. 1–36.

[13] Laura McKenna. *Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing.* Working Paper. The Research and Methodology Directorate - US Census Bureau, Nov. 2018, p. 39. URL: `https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20for%20the%201970-2010%20Censuses.pdf`.

[14] Helen Fay Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life.* Stanford, Calif: Stanford Law Books, 2010. ISBN: 978-0-8047-5236-7 978-0-8047-5237-4.

[15] Alan Rubel. "The Particularized Judgment Account of Privacy". In: *Res Publica* 17.3 (2011), pp. 275–290. ISSN: 1356-4765. DOI: `10.1007/s11158-011-9160-4`.

[16] Jeremy Seeman and Daniel Susser. "Between Privacy and Utility: On Differential Privacy in Theory and Practice". In: *ACM Journal on Responsible Computing* (Oct. 2023). DOI: `10.1145/3626494`. (Visited on 02/19/2024).

[17] Jeremy Seeman et al. *Privately Answering Queries on Skewed Data via per Record Differential Privacy.* `http://arxiv.org/abs/2310.12827`. Oct. 2023. DOI: `10.48550/arXiv.2310.12827`. arXiv: `2310.12827 [cs]`. (Visited on 02/17/2024).

[18] Aleksandra Slavković and Jeremy Seeman. "Statistical Data Privacy: A Song of Privacy and Utility". In: *Annual Review of Statistics and Its Application* 10.1 (2023), pp. 189–218. DOI: `10.1146/annurev-statistics-033121-112921`. eprint: `2205.03336` (cs, stat). (Visited on 04/23/2023).

[19] Daniel J. Solove. *Understanding Privacy.* Cambridge, MA: Harvard University Press, 2008. ISBN: 978-0-674-02772-5. (Visited on 10/12/2021).