Differential Privacy for Government Agencies – Are We There Yet?

Keywords: *Confidentiality, Privacy, NSI*

1 Abstract

The concept of differential privacy gained substantial attention in recent years being the only standing approach offering formal privacy guarantees, which hold irrespective of the sensitivity of the data and of the assumptions regarding the background knowledge of a potential attacker. Since the seminal paper by Dwork et al. [1] was published in 2006, thousands of papers mostly from the computer science community addressed the topic from various perspectives. Given its roots in theoretical computer science, it is perhaps not surprising that most of these works approached the problem from a theoretical perspective. While new algorithms that satisfy the differential privacy requirements for specific analysis tasks are proposed almost every other day, their performance is typically only evaluated by looking at measures such as the maximum expected error under asymptotic regimes (i.e., assuming $n \to \infty$). Evaluations based on real data are sparse, although it is well understood that the relative advantages of different algorithms crucially depend on the size and structure of the available data (see [2] for a nice illustration of this phenomenon). Despite the limited experience in practice, the concept of differential privacy has been embraced by the industry in recent years. Many companies, especially in the tech industry, such as Google [3], Apple [4], Microsoft [5], Facebook [6] or Uber [7] have deployed the concept for some of their products or are currently conducting research with the aim of implementing the approach in the future.

Despite the excitement in academia and industry, the enthusiasm at government agencies and national statistical institutes (NSIs) has been limited so far. While some agencies explored the feasibility of the approach in limited settings [8, 9], the only large-scale deployment of the approach for many years was OntheMap, a graphical interface offered by the U.S. Census Bureau showing commuting patterns in the United States. The underlying data are protected using an algorithm which satisfies $\epsilon - \delta$-probabilistic differential privacy [10]. This changed recently, when the U.S. Census Bureau announced that it will adopt differential privacy for the decennial Census 2020 [11]. Compared to most other data products gathered at NSIs which are based on surveys with limited sample sizes and hundreds of variables, protecting the Census seems to be a straightforward task: it contains hundred millions of records and only asks seven questions. Still, the fact that a research team of computer scientists and statisticians has been working on this problem for several years now and the severe concerns regarding the accuracy of the results that were raised after results from a test run of the algorithm using 2010 Census data were released [12, 13, 14, 15] illustrate the difficulties when trying to implement the ideas in practice. Many problems arise, since the requirements when implementing differential privacy approaches at government agencies are fundamentally different from the requirements in industry: The data should be available for many years, results should be reproducible, users of the data are typically interested in making inferences regarding a specific target population, agencies are not the final users of the data, incentives for sharing the data are virtually non-existent, etc. All these aspects need to be taken into account

when considering whether the concept might be a viable approach for solving the ever existing dilemma between confidentiality protection and broad access to the data. This talk is not meant to provide a road map how to implement differential privacy at government agencies. Instead, it will highlight some important aspects that need to be on everybody's radar and open questions that still need to be addressed when thinking about if and how the concept could be applied in the government context. The talk will address the following aspects:

- **Data availability and access:** I will discuss the challenges arising for the typical data products at government agencies with limited sample sizes but very detailed information. I will also illustrate why neither the query response system, for which differential privacy was originally developed nor restricted access for accredited researchers will be an option when adopting differential privacy for government agencies.

- **Understanding the privacy guarantees and impacts on accuracy**: I will highlight the difficulties in anticipating the impacts that guaranteeing differential privacy will have on the accuracy of the results obtainable from the protected data. I will also show that guaranteeing differential privacy alone is not sufficient to protect respondents from harm and that understanding the level of protection provided through differential privacy is not trivial.

- **Differential privacy in the survey context:** I will illustrate why implementing differential privacy might have negative effects on the willingness to participate in a survey. I will also discuss the difficulties in understanding the interaction between differential privacy and common data processing steps such as weighting and imputation and the challenges when trying to account for the data protection procedures when making inference to the underlying population.

- **Setting the value of ε:** Deciding which value to choose for the privacy parameter ε is always difficult. However, statistical agencies face additional challenges as they cannot fully anticipate the future uses of the data, they have to take into account that low accuracy may also have negative consequences for the respondents and that the data release might also affect units that did not have a chance to decide whether they want to be part of the database.

References

[1] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284, 2006.

[2] Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam Smith, and Salil Vadhan. Differentially private simple linear regression. *arXiv preprint arXiv:2007.05157*, 2020.

[3] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[4] Apple's Differential Privacy Team. Learning with privacy at scale. *Apple Machine Learning Journal*, 1(8), 2017.

[5] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pages 3571–3580, 2017.

[6] Solomon Messing, Christina DeGregorio, Bennett Hillenbrand, Gary King, Saurav Mahanti, Zagreb Mukerjee, Chaya Nayak, Nate Persily, Bogdan State, and Arjun Wilkins. Facebook Privacy-Protected Full URLs Data Set, 2020. URL https://doi.org/10.7910/DVN/TDOAPG.

[7] Uber Security. Uber releases open source project for differential privacy, 2017. URL https://medium.com/uber-security-privacy/differential-privacy-open-source-7892c82c42b6.

[8] Jordi Soria-Cormas and Jörg Drechsler. Evaluating the potential of differential privacy mechanisms for census data. In *UNECE work session on statistical data confidentiality*, 2013.

[9] James Bailie and Chien-Hung Chien. Abs perturbation methodology through the lens of differential privacy, 2019.

[10] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *2008 IEEE 24th international conference on data engineering*, pages 277–286. IEEE, 2008.

[11] John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018.

[12] Steven Ruggles, Catherine Fitch, Diana Magnuson, and Jonathan Schroeder. Differential privacy and census data: Implications for social and economic research. In *AEA papers and proceedings*, volume 109, pages 403–08, 2019.

[13] X Wezerek and D Van Ripper. Changes to the census could make small towns disappear, 2020. URL https://www.nytimes.com/interactive/2020/02/06/opinion/census-algorithm-privacy.html.

[14] David Van Riper, Tracy Kugler, and Steven Ruggles. Disclosure avoidance in the census bureau's 2010 demonstration data product. In *International Conference on Privacy in Statistical Databases*, pages 353–368. Springer, 2020.

[15] Maine state economist letter to census on differential privacy. URL availableat:https://sdcclearinghouse.com/2020/02/27/maine-state-economist-letter-to-census-on-differential-privacy/.