Examining Deidentified Data Quality using NIST Datasets and Tools

Saswat Das[1]*, Razane Tajeddine[2], Ferdinando Fioretto[1]

Abstract
Over the course of several months, the authors have utilized the breadth of NIST CRC tools and datasets, in particular the SDNIST data report tool and the Diverse Communities Data Excerpts, as a central element of their research on comparing traditional statistical disclosure control (SDC) methods to differential privacy (DP) primitives and against DP data generation algorithms like SmartNoise AIM. These results provided a holistic look at the nature of the deidentified data releases, and thus allowed the authors to assess the quality of, privacy of (in certain respects), disparities in, and similarity to the target data of the deidentified data releases, among other things. In this short paper, the authors seek to demonstrate how the use of tools provided by NIST enabled such evaluations, and thus motivate their use in future research involving deidentified data.

Keywords
SDNIST, Diverse Communities Data Excerpts, Privacy, Fairness

[1] *Department of Computer Science, University of Virginia, USA*
[2] *Department of Computer Science, University of Helsinki, Finland*
***Corresponding author**: duh6ae@virginia.edu

1. Introduction

Over the course of the past several months, the authors have investigated data privacy in the context of demographic data release that is used for key decisions, such as federal fund allocation, electoral district redrawing, etc. The released data contain highly sensitive microdata that must be deidentified prior to its release. Traditional statistical disclosure control (SDC) techniques are many times used for data deidentification, despite criticism of such methods and their demonstrated inability to defend against some privacy attacks. These techniques have a long history, dating back to the 1930 decennial release by the US Census Bureau, which leveraged traditional SDC techniques such as suppressing certain tables based on the number of people or households in a given area and swapping data in records with similar characteristics [1, 2]. Examples of common SDC methods include swapping [3] and cell suppression [4].

On the other hand, differential privacy (DP) has gained prominence as the de facto gold standard of privacy, and has in recent years been adopted by the US census bureau to produce privacy-preserving data releases.

However, a lot of statistical agencies still make use of SDC methods to deidentify data, citing concerns about the noise addition when using DP mechanisms, as that can and does impact the utility and fairness of the output data. In our works, we studied this problem and showed that using DP for data deidentification gives comparatively better results than the SDC methods we explored, despite the mentioned issues of DP in terms of utility and fairness. To that end, we recognized that no methods to certify the privacy provided by SDC methods with rigorous mathematical guarantees existed, and thus our work provided differentially private approximations of these methods to enable a fair comparison in terms of differential privacy.

Points of Discussion

In this short paper, we will discuss how we utilized some resources provided by NIST in order to evaluate the fairness and quality of deidentified data releases.

In particular, we report the use of the following tools/data.

- The Diverse Communities Data Excerpts [5]
- The SDNIST Report Tool [6]

2. Methods

2.1 Diverse Communities Data Excerpts

The Diverse Communities Data Excerpts are a set of carefully crafted data excerpts from the IPUMS data, involving several public use microdata areas, or PUMAs, from across the United States, and contiguous sets of PUMAs from Massachusetts and Texas. These form what shall be referred to hereon by the authors as the National, Massachusetts (MA), and Texas (TX) datasets, respectively.

Unless otherwise specified, any histograms created using this data use the features PUMA, sex, race, annual income decile, and house ownership status.

2.2 Differential Privacy.

Differential privacy (DP) [7] is a privacy notion which quantifies and bounds the privacy loss of an individual participation to a computation. The action of changing a record from a dataset D, resulting in a new dataset D', defines the notion of *adjacency*, denoted $D \sim D'$.

Definition 1. *A mechanism $\mathscr{M} : \mathscr{D} \to \mathscr{R}$ with domain $\mathscr{D}$ and range $\mathscr{R}$ is (ε, δ)-differentially private if, for any two neighboring datasets $D \sim D' \in \mathscr{D}$, and any subset of output responses $R \subseteq \mathscr{R}$:*

$$\Pr[\mathscr{M}(D) \in R] \leq e^{\varepsilon} \Pr[\mathscr{M}(D') \in R] + \delta.$$

When $\delta = 0$, the mechanism is a pure ε-DP mechanism. The parameter $\varepsilon > 0$ is the privacy budget and it describes the algorithm's *privacy loss*, and parameter $\delta \in [0, 1)$ is a relaxation term that captures the probability of the event that the mechanism fails to satisfy ε-DP.

The *Laplace mechanism* adds Laplace noise to data entries before release. This mechanism satisfies ε-DP.

2.3 Histogram Deidentification

We next discuss a predominant SDC system which, in contrast to differential privacy, does not provide formal bounds on privacy leakage, as well as a simple DP primitive for histogram release, hereon referred to as the Laplace histogram, named thus for its use of the Laplace mechanism to obscure the actual counts of a histogram.

Cell Suppression

The cell suppression technique [1], frequently employed by statistical agencies (e.g., [8]), aims at concealing the low-frequency counts in histograms before data dissemination.

Definition 2. *Given a histogram $\boldsymbol{x}$ and a threshold value k, cell suppression returns a private histogram $\tilde{\boldsymbol{x}}$ with entries $\tilde{x}_i = x_i$, for all $x_i = 0$ or $x_i \geq k$, and $k/2$ otherwise.*

A significant limitation of this approach is that it only protects groups with a non-zero and low number of records while revealing the counts and information of other groups in full.

Laplace Histogram

This method can be seen as a DP primitive in the histogram data release setting; as the name suggests, this method involves perturbing the counts of a histogram using the Laplace mechanism. The Laplace mechanism for histogram data release is defined by $\mathscr{M}_{\text{Lap}}(\boldsymbol{x}) = \boldsymbol{x} + \text{Lap}(2/\varepsilon)$, where $\text{Lap}(2/\varepsilon)$ is the Laplace distribution centered at 0 and with scaling factor $2/\varepsilon$.

This method is easily seen, by virtue of the Laplace mechanism's DP guarantees to provide ε-DP.

DP Cell Suppression

Here, we introduce a randomized version of cell suppression, *DP cell suppression* and denote it by $\mathscr{M}_{\text{CS}}$. DP cell suppression releases a private count $\hat{x}_i$ for every $i \in [n]$ as follows:

$$\mathscr{M}_{\text{CS}}(D)_i = \hat{x}_i = \begin{cases} x_i & \text{if } x_i \geq k \text{ and } x_i + \eta_i \geq k \\ \lfloor k/2 \rfloor & \text{otherwise} \end{cases}, \quad (1)$$

where $\eta_i \sim \text{Lap}(\lambda)$ is a noise variable sampled from a 0-centered Laplace distribution with factor $\lambda = 2/\varepsilon$ and k is the suppression threshold. The DP cell suppression mechanism $\mathscr{M}_{\text{CS}}$ serves merely as a differentially private approximation of the non-private cell suppression mechanism while being able to quantify the worst-case privacy loss in terms of DP.

2.4 DP Data Generation

In a dataset data release setting, there exist a number of algorithms for DP synthetic data generation, such as CTGAN [9], MST [10], and the state-of-the-art Adaptive Iterative Mechanism (AIM) [11]. In this work, we use the SmartNoise implementation of AIM and compare its data release quality with the quality when using cell suppression and the Laplace mechanism.

2.5 Utility/Fairness Metrics

The notions of utility and fairness central to the analysis rely on the concept of (statistical) *bias*. For any entry $i \in [n]$, the bias associated with a mechanism $\mathscr{M}$ is

$$\mathscr{B}(\mathscr{M})_i = \mathbb{E}\left[\mathscr{M}(D)_i\right] - x_i(D),$$

where the expectation is over the randomness of the mechanism. Fairness is defined as the maximal difference in biases across the histogram entries.

Definition 3 (α-fairness [12])**.** *A mechanism $\mathscr{M}$ is said to be α-fair if the maximum difference among the biases is bounded by α, i.e.,*

$$\|\mathscr{B}(\mathscr{M})\|_{\rightleftharpoons} = \max_{i \in [n]} \mathscr{B}(\mathscr{M})_i - \min_{i \in [n]} \mathscr{B}(\mathscr{M})_i \leq \alpha,$$

where $\mathscr{B}(\mathscr{M}) = [\mathscr{B}(\mathscr{M})_1 \ \ldots \ \mathscr{B}(\mathscr{M})_n]$.

2.6 DP Approximations of SDC Methods

In this work, we will mainly include results on cell suppression when it comes to SDC methods. For further details about the DP variants of the SDC methods, readers are encouraged to go through our preprint[13].

3. Results and Discussion

In this section, we will describe and discuss how NIST's resources allowed us to evaluate the quality and the disparities present in data releases made using different deidentification methods.

3.1 Linear Regression + Heatplots

Figure 1 includes heatplots showing errors on 2-way marginals on income decile (`PINCP_DECILE`) and educational status (`EDU`) and linear regression between these variables for target

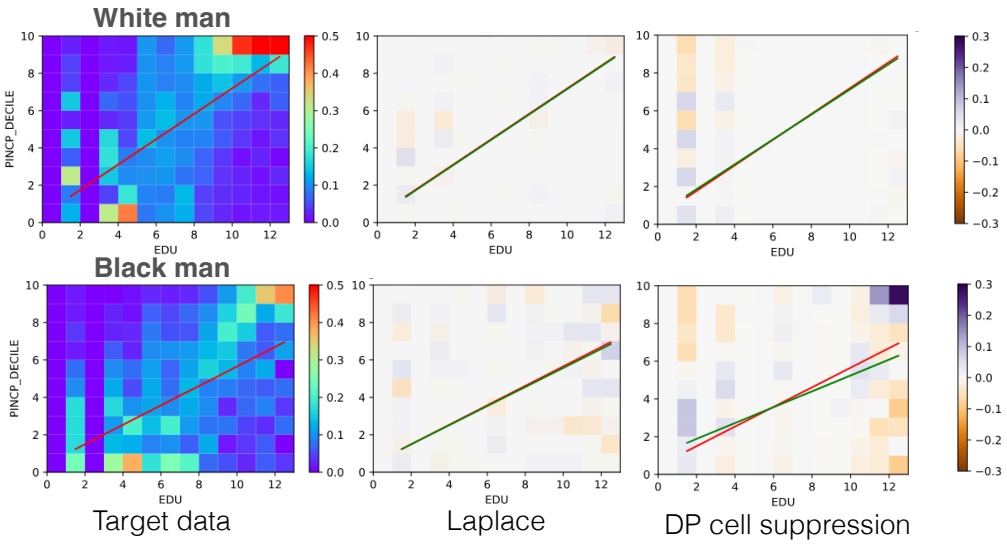

Figure 1. National ACS Dataset: Heatplots for two-way marginals on *income decile* and *education* for the Laplace mechanism (left) and DP Cell Suppression (right) comparing White vs Black Men subgroups. These results were obtained using $\varepsilon = 2$ and threshold $k = 10$. The red line refers to the target data and the green line refers to the deidentified data.

and deidentified data. These heatplots and linear regression plots were generated for several (race, sex) pairs using the SDNIST tool, and here we present plots for white males vs. black males on the National dataset.

We found these heatplots interesting for several reasons. For one, they provide a way to compare the relative levels of utility offered by the different deidentification methods, by comparing these plots across methods. In addition, the production of plots for different demographics provided an insight into the disparities incurred by each method and helped us see that minority groups suffer more loss in utility than majority groups. Moreover, we realized that some methods (viz. Laplace histogram) seem to perform better than others (viz. (DP) cell suppression) when it comes to fairness.

3.2 Propensity Plots

The SDNIST tool provided us with propensity plots that give an idea of the quality of output deidentified data (as in how much it resembles the original target data) and the number of records that were successfully detected as being fake records. The plots are shown in Figure 2. The blue curve represents the original data's propensity curve, and the red represents the deidentified propensity curve; these propensities are derived from a classifier trained to distinguish between members and non-members of the target data. The proximity of these curves shows how close the deidentified data is to the target data. The spikes beyond the center of these curves to the right show the members of the deidentified data that are classified as being fake, increasing in confidence the further right they are. Therefore, the spike seen prominently for cell suppression on the far right (around the 100 mark), and to a much smaller extent for Laplace, provides information about the number of records confidently classified as being fake.

This helped provide a neat illustration of the quality of deidentified data output using each of these methods. The data is subset by the features mentioned in subsection 2.1. AIM provides very high quality as well as realistic deidentified data, while Laplace histogram slightly less so, and cell suppression seems to offer significantly worse utility with clearly fake records (perhaps due to the imputation of suppressed counts with $\lfloor k/2 \rfloor$).

3.3 Map Plots

In addition, we briefly present choropleth plots in Figure 3 for all the 31 PUMAs present in the Diverse community excerpts dataset showing the differences in ℓ_1 error values between Laplace histogram and DP cell suppression for the same value of ε on histograms produced using the aforementioned features on each PUMA separately. This allowed us to explore a wide breadth of geographical regions across the United States to observe the behaviour of these mechanisms on them and disparities in utilities across PUMAs. This provided a helpful illustration of the advantages of using even simple DP mechanisms/primitives over cell suppression; Laplace histogram outperforms suppression with respect to the utility and fairness of the output data release.

3.4 The Effect of Sparsity

One other dimension to this problem is the effect of the level of sparsity on ℓ_1 errors of the released data. We use here a simple definition of sparsity: the proportion of the histogram populated by zero bins to the total number of bins. In some of our other work, we studied the theoretical results for this notion of sparsity for cell suppression and for Laplace histogram. After which, we were able to provide empirical results on marginal queries made on the Massachusetts dataset, while varying the number of zero bins within the vector of counts and recording their ℓ_1 errors. The results are presented in Figure 4 along with a brief discussion in the caption. This helped us understand the effect of sparsity on the different private

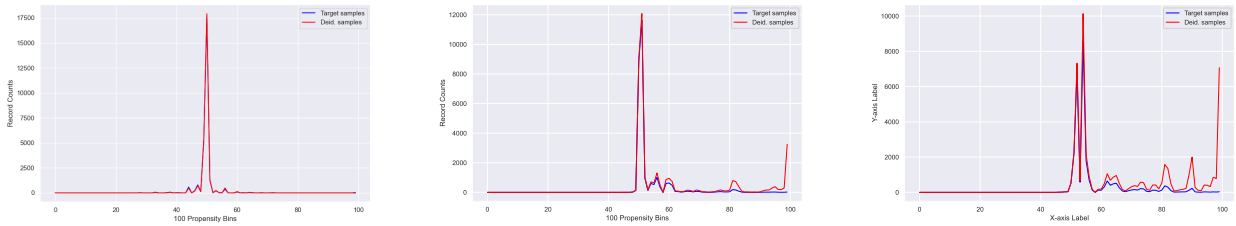

Figure 2. Propensity plots for SmartNoise AIM (left), Laplace histogram (center), and DP cell suppression (right) for $\varepsilon = 4$.

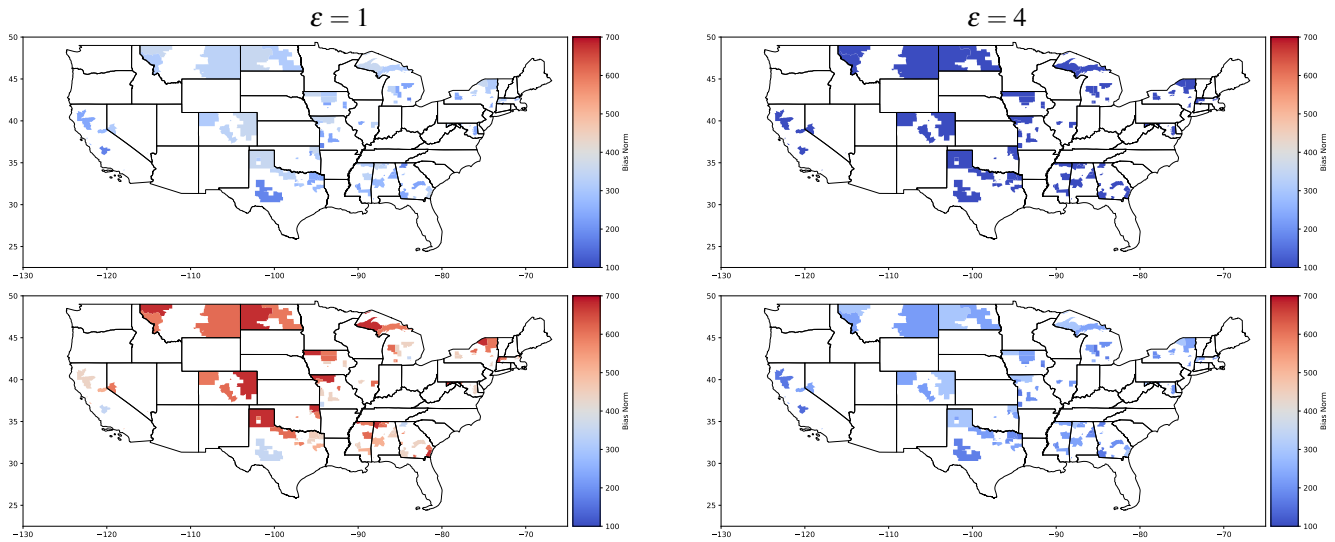

Figure 3. Choropleth Plots: Plotting the bias observed over each PUMA (in the union of the MA, TX, and National DCDE datasets) for Laplace histogram (top) and DP cell suppression (DPCS) (bottom) for $\varepsilon = 1, 4$ (left and right, respectively). Regions colored with white are not included in the DCDE datasets. Plotting is done on a color scale of dark blue (low bias) to dark red (high bias).

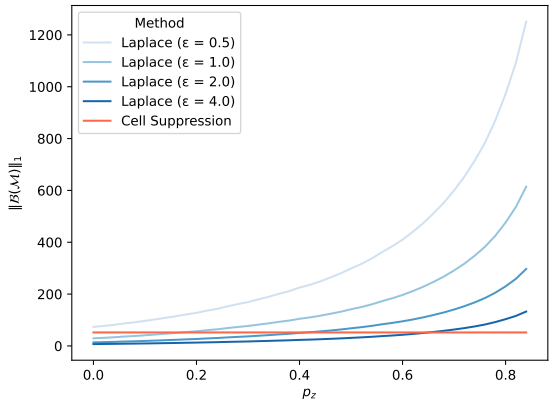

Figure 4. Plot showing the ℓ_1-norm of the bias versus the sparsity factor (*i.e.*, the proportion of zero bins). We can see that as the factor grows, the bias from using Laplace histogram grows while the bias from cell suppression is constant.

histogram release methods. Recall that while Laplace noise may be drawn from a distribution centered at 0, projection of negative counts post perturbation to 0 coupled with the fact that zero-count bins are perturbed by Laplace histogram and not by suppression contribute to what is observed.

Takeaways

These results helped us observe several facets of disclosure avoidance mechanisms, showing that even simple DP mechanisms can outperform SDC methods, *e.g.*, suppression, in terms of fairness and utility, while providing much stronger privacy guarantees (also proven theoretically in [13]).

4. Conclusion

The tools and datasets provided by NIST have enabled the authors to extensively perform research on the effects of the use of data disclosure avoidance mechanisms on high-quality demographic datasets that are close to Census data releases, therefore allowing us to perform a holistic evaluation so as to inform responsible use of these methods and further encourage the use of differential privacy in this domain. These results form but a fraction of the results obtained using NIST's resources, including those on privacy attacks and for different levels of sparsity, which are out of the scope of this short paper. The authors firmly believe that the provision of quality resources by NIST has been a key component for their work in this domain, and will continue to aid evaluations of this nature on quality datasets and using helpful and informative evaluation methods and metrics.

Acknowledgments

The works mentioned in this paper were done by the authors in collaboration with Keyu Zhu, Christine Task, and Pascal van Hentenryck, whose contributions and advice we gratefully acknowledge. The work of R. Tajeddine was supported by the Research Council of Finland (grant 343555).

References

[1] James P Kelly, Bruce L Golden, and Arjang A Assad. Cell suppression: Disclosure protection for sensitive tabular data. *Networks*, 22(4):397–417, 1992.

[2] Tore Dalenius and Steven P Reiss. Data-swapping: A technique for disclosure control. *Journal of statistical planning and inference*, 6(1):73–85, 1982.

[3] Steven P Reiss. Practical data-swapping: The first steps. *ACM Transactions on Database Systems (TODS)*, 9(1): 20–37, 1984.

[4] Matthias Templ. Statistical disclosure control for microdata. *Cham: Springer*, 2017.

[5] Christine Task, Karan Bhagat, Streat Damon, and Gary Howarth. NIST Diverse Community Excerpts Data, December 2022. URL `https://data.nist.gov/od/id/mds2-2895`.

[6] C. Task, K. Bhagat, and G.S. Howarth. Sdnist v2: Deidentified data report tool. National Institute of Standards and Technology, 2023.

[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[8] Tatauranga Aotearoa. Microdata output guide. `https://www.stats.govt.nz/assets/Methods/Microdata-Output-Guide-2020-v5-Sept22update.pdf`, 2020.

[9] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan, 2019.

[10] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the nist contest: A scalable and general approach to differentially private synthetic data, 2021.

[11] Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. Aim: An adaptive and iterative mechanism for differentially private synthetic data, 2022.

[12] Keyu Zhu, Ferdinando Fioretto, and Pascal Van Hentenryck. Post-processing of differentially private data: A fairness perspective. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4029–4035. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/559. URL `https://doi.org/10.24963/ijcai.2022/559`. Main Track.

[13] Keyu Zhu, Ferdinando Fioretto, Pascal Van Hentenryck, Saswat Das, and Christine Task. Privacy and bias analysis of disclosure avoidance systems, 2023.