Counting Kindergartners: De-Identification's Impact on Student Representation

Sarah Radway and Miranda Christ

June 18, 2022

1 Introduction

An important application of the U.S. Census Bureau's Demographic and Housing Characteristics (DHC) data is in school planning. Single-year-of-age data is crucial in helping schools estimate the number of incoming students. Doing so is challenging, since exact counts are often not possible. For example, families might not respond to the survey, or families might move after responding to the survey. While uncertainty comes from a variety of sources, we are interested in particular in uncertainty added by disclosure avoidance methods. The methods we consider are *data swapping*, used by the Census Bureau from 1990 to 2010, and *differential privacy*, adopted by the Census Bureau in 2020. We work toward answering the question of which disclosure avoidance method is best suited to support school planning while preserving privacy as is statutorily mandated.

2 Methods

At a high level, our methods follow those of [CRB22].

Generating Data

Because the true microdata collected by the Census Bureau is not publicly available, we use synthetic microdata that we treat as the ground truth throughout our experiments. Our microdata was generated using a method of Flaxman and Haddock, the associated code of which is available at `https://github.com/aflaxman/ppmf_12.2_reid`. This library uses integer programming to reconstruct microdata matching the tables in Summary File 1 of the Census Bureau's published 2010 Decennial Census Data. The resulting synthetic data thus closely matches the characteristics of the populations they simulate. To simulate school districts of varying sizes, we create census tracts from Alabama and Rhode Island of varying sizes, shown in Table 1. Following Flaxman and Haddock, we refer to our ground truth synthetic data as a reconstructed microdata file (ReMF). We use our ReMF to obtain two sets Privacy-Protecting Microdata Files (PPMFs): one using swapping and one using differential privacy.

Tract	Population
RI-51000	1,508
RI-15800	3,771
RI-20101	8,415
AL-100	12,267
AL-5400	4,068
AL-10001	8,035
AL-10704	19,477
AL-400	23,781

Table 1: Population sizes of our synthetic census tracts

Swapping

Swapping is a disclosure avoidance method that involves exchanging some subset of data from one geographic region with data from another geographic region within the same state. For example, a house in Alameda County may be swapped with a house in Contra Costa County. The Census Bureau has used swapping in its decennial census data since 1990.[1]

The implementations of swapping used by the Census Bureau have not been published due to privacy concerns. We create a swapping implementation based upon the limited published guidance from the Census Bureau. Our implementation is parameterized by the **swap rate**, the proportion of households that are exchanged. We choose our swap rates to run based upon published U.S. Census Bureau guidance [cen21]; which states that block swap rates range from 0-50%. Our graphs focus on the lower values of these swap rates, as within our dataset, these would likely be the appropriate numbers.

In our swapping implementation, first, the number of households to be swapped is determined based on the swap rate. We then compile a list of households to be swapped, prioritizing households whose members' characteristics (race, sex, age) make that household unique within its tract. For each household on this list, we choose a household from the state population with which to exchange its data. Our implementation tries to find a household matching the number of minors and adults in the household; if there are multiple, a household is chosen uniformly at random from these options. If no such household exists, this constraint is gradually relaxed until such a household is found.

Once all selected houses have been swapped, we compute a histogram of single-year-of-age counts over the swapped data.

Differential privacy

In our differential privacy (DP) implementations, we use the geometric mechanism, which is parameterized by ϵ, from IBM's Differential Privacy Library.[2] Our two DP implementations, *age only* and *sex* $\times$ *age*, differ in how exactly this noise is added. In *age only*, we first compute a single-year-of-age histogram over the ReMF and add independent noise according to our geometric mechanism to each of these age counts. In *sex* $\times$ *age*, we compute a histogram over each combination of binary sex and single-year-of-age and add independent noise according to our geometric mechanism to each of these product counts. We then obtain a de-identified single-year-of-age histogram by adding the two noisy counts for each age. The *sex* $\times$ *age* adds more noise and thus more inaccuracy compared to *age only* for the same value of ϵ.

We found that the trends for *age only* and *sex* $\times$ *age* DP were similar, and we focus on *age only* for the remainder of these report.

Measuring Accuracy

We measure accuracy using mean absolute percentage error (MAPE). We compute the MAPE over the single-year-of-age histograms of our swapping and DP PPMFs, restricted to the following ranges of interest: total population; population of age under 18; population of ages 4 and 5. After restricting the PPMF histogram and the ReMF histogram to the range in question, for each selected age, we compute the absolute error and divide it by the true count of that age group from the ReMF to obtain the absolute percentage error. If this true count is zero, we omit it. We compute the sum of the non-omitted absolute percentage errors of our selected ages, then divide by the number of selected ages to obtain the average.

For each swap rate an epsilon value, we computed the MAPE over five separate swapping and DP runs. In our plots, we show the average MAPE over these five runs to show the trends more clearly.

[1] `https://www.census.gov/library/visualizations/2019/comm/history-privacy-protection.html`
[2] `https://github.com/IBM/differential-privacy-library`

3 Results

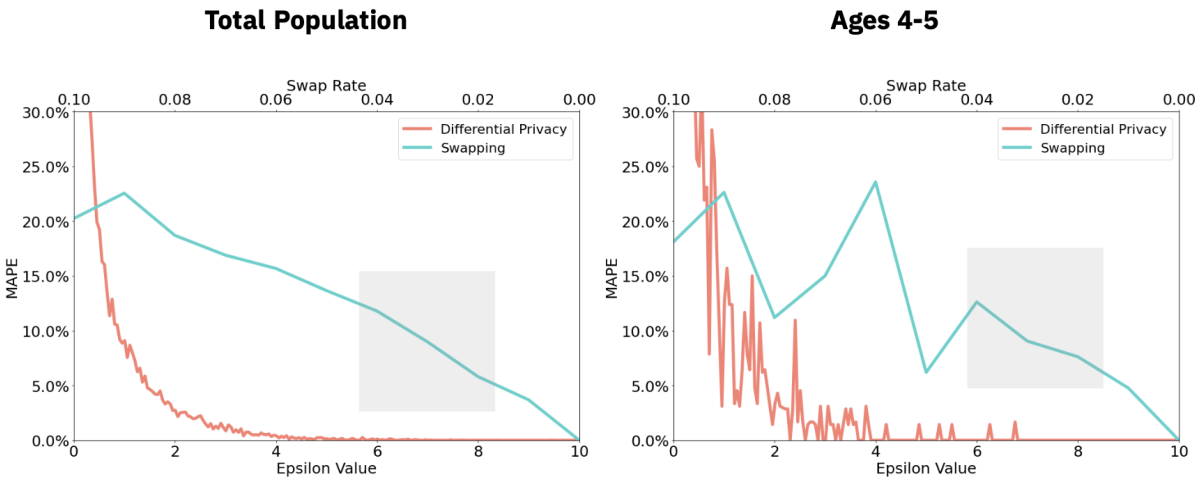

Figure 1: MAPE for RI-51000, a tract of population 1,508, for *age only DP* and swapping. Shown on the left is the MAPE over all ages, and shown on the right is the MAPE over ages 4 and 5 only.

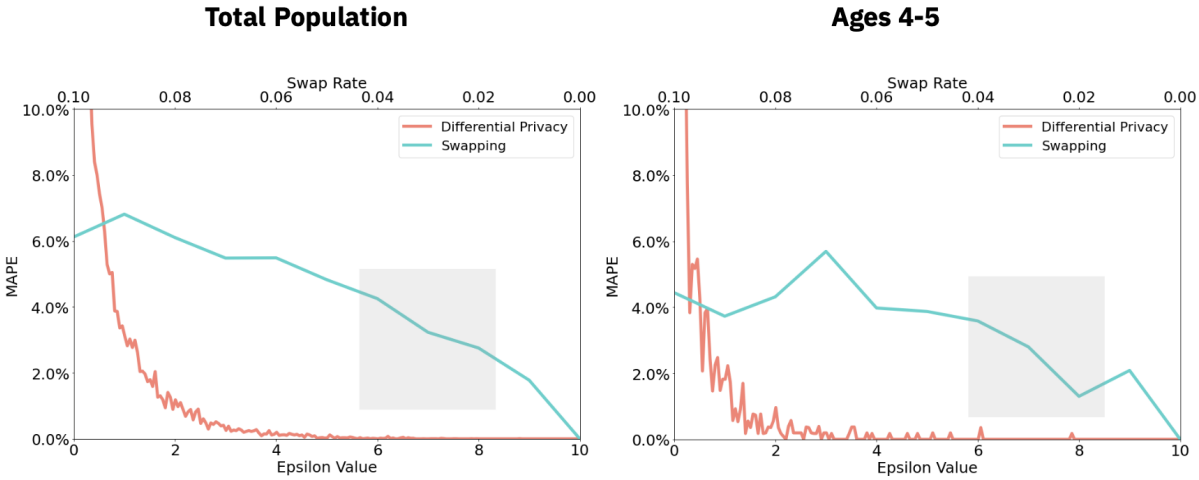

Figure 2: MAPE for RI-15800, a tract of population 3,771, for *age only DP* and swapping. Shown on the left is the MAPE over all ages, and shown on the right is the MAPE over ages 4 and 5 only.

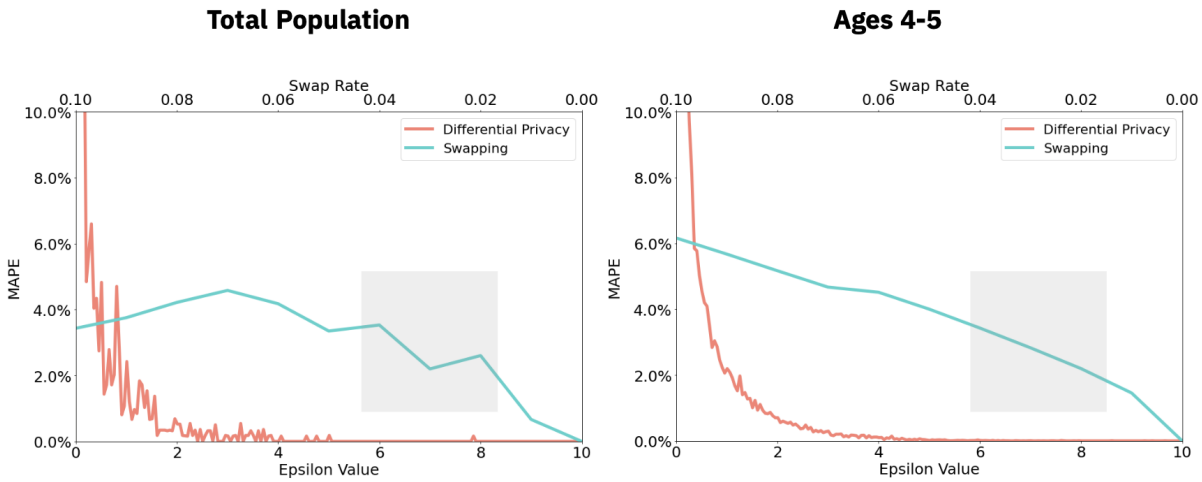

Figure 3: MAPE for AL-5400, a tract of population 4,068, for *age only DP* and swapping. Shown on the left is the MAPE over all ages, and shown on the right is the MAPE over ages 4 and 5 only.

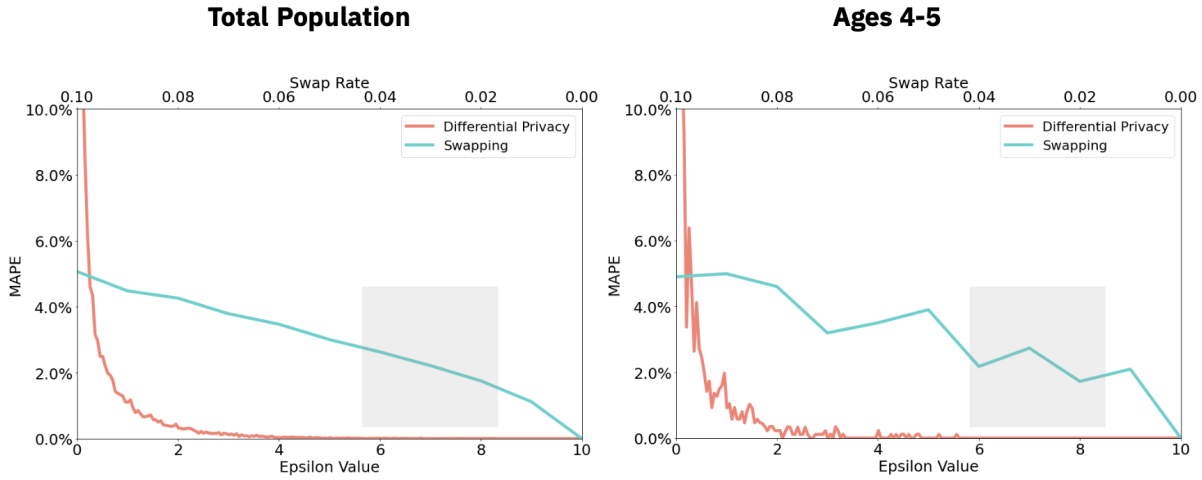

Figure 4: MAPE for AL-10001, a tract of population 8,035, for *age only DP* and swapping. Shown on the left is the MAPE over all ages, and shown on the right is the MAPE over ages 4 and 5 only.

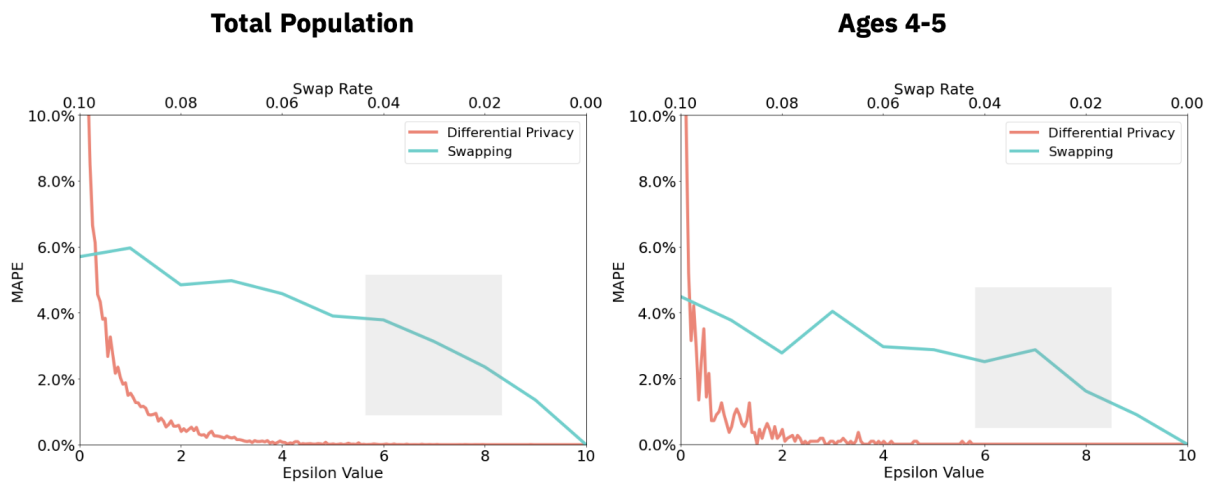

Figure 5: MAPE for RI-20101, a tract of population 8,415, for *age only DP* and swapping. Shown on the left is the MAPE over all ages, and shown on the right is the MAPE over ages 4 and 5 only.

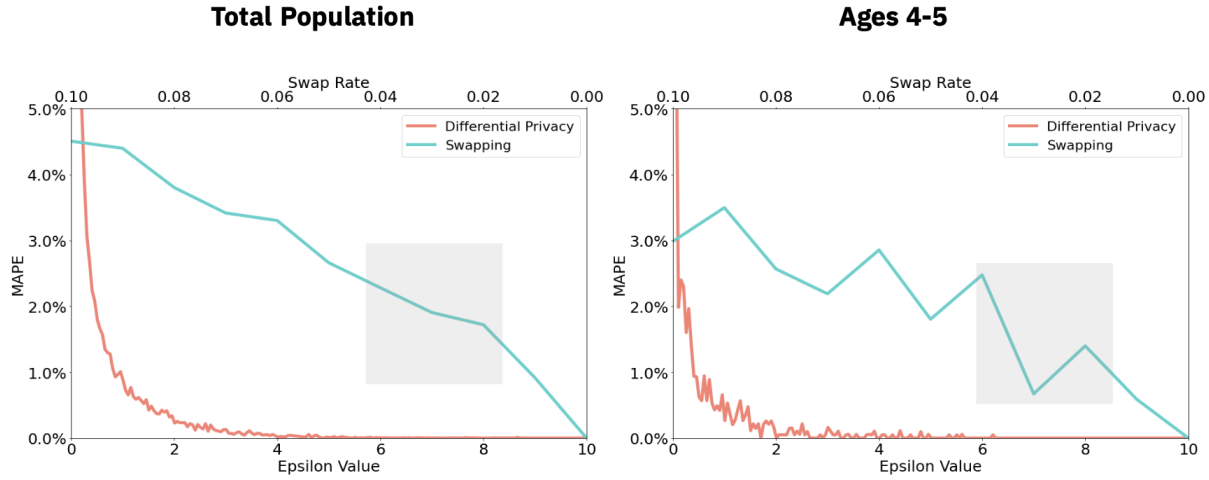

Figure 6: MAPE for AL-100, a tract of population 12,267, for *age only DP* and swapping. Shown on the left is the MAPE over all ages, and shown on the right is the MAPE over ages 4 and 5 only.

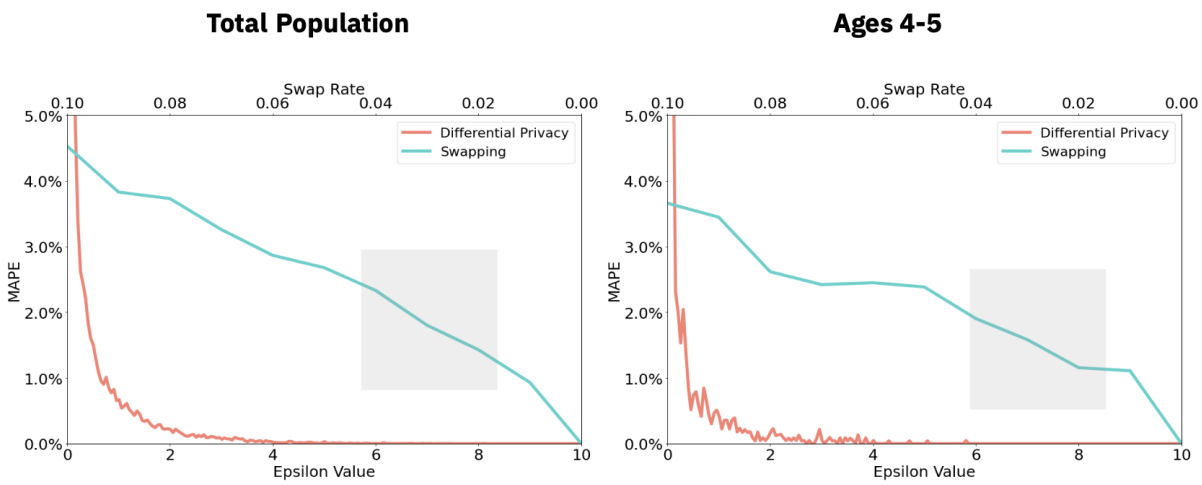

Figure 7: MAPE for AL-10704, a tract of population 19,477, for *age only DP* and swapping. Shown on the left is the MAPE over all ages, and shown on the right is the MAPE over ages 4 and 5 only.

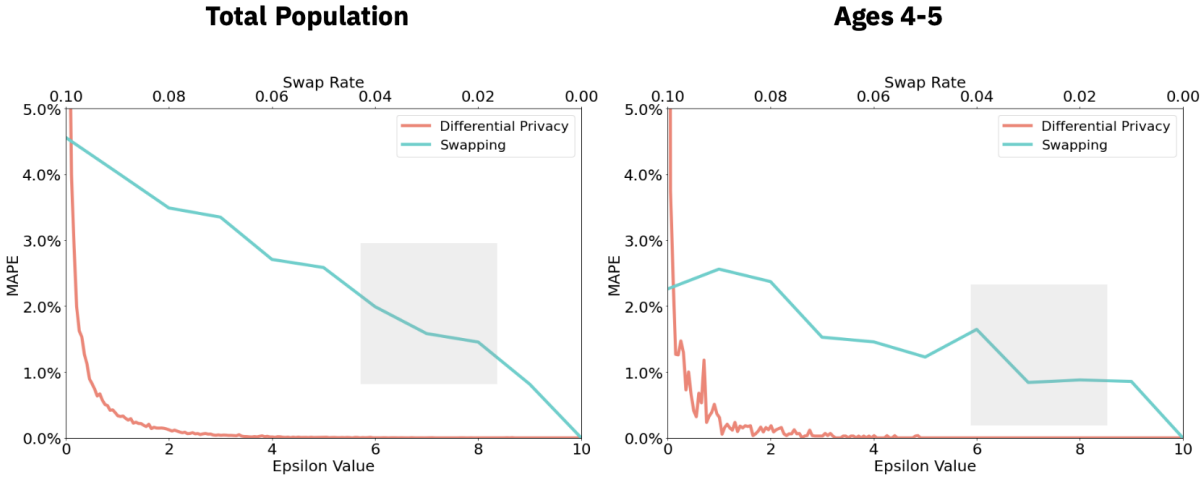

Figure 8: MAPE for AL-400, a tract of population 23,781, for *age only DP* and swapping. Shown on the left is the MAPE over all ages, and shown on the right is the MAPE over ages 4 and 5 only.

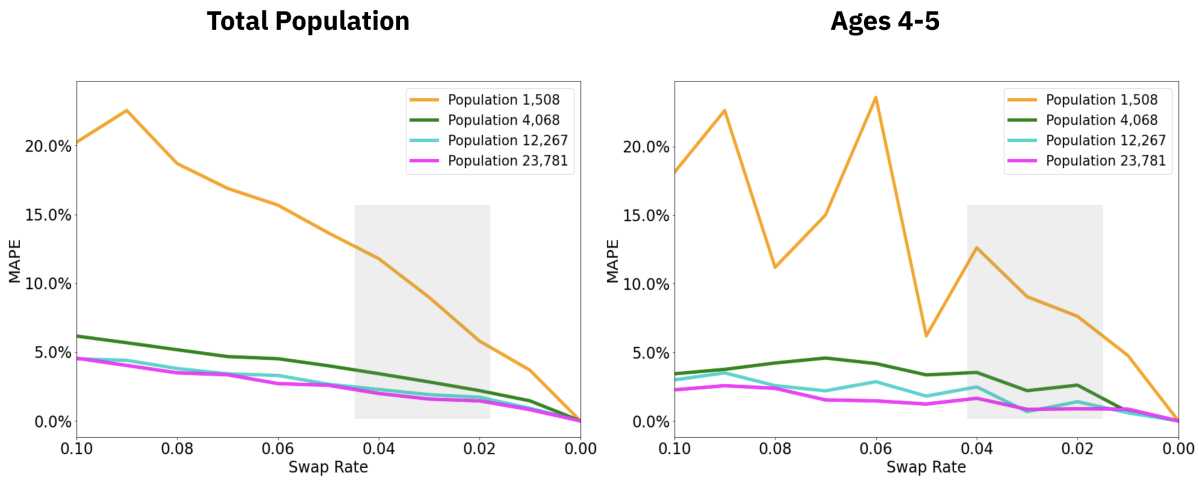

Figure 9: MAPE of swapping for tracts of varying population sizes: RI-51000 (pop. 1,508), AL-5400 (pop. 4,068), AL-100 (pop. 12,267), AL-400 (pop. 23,781)

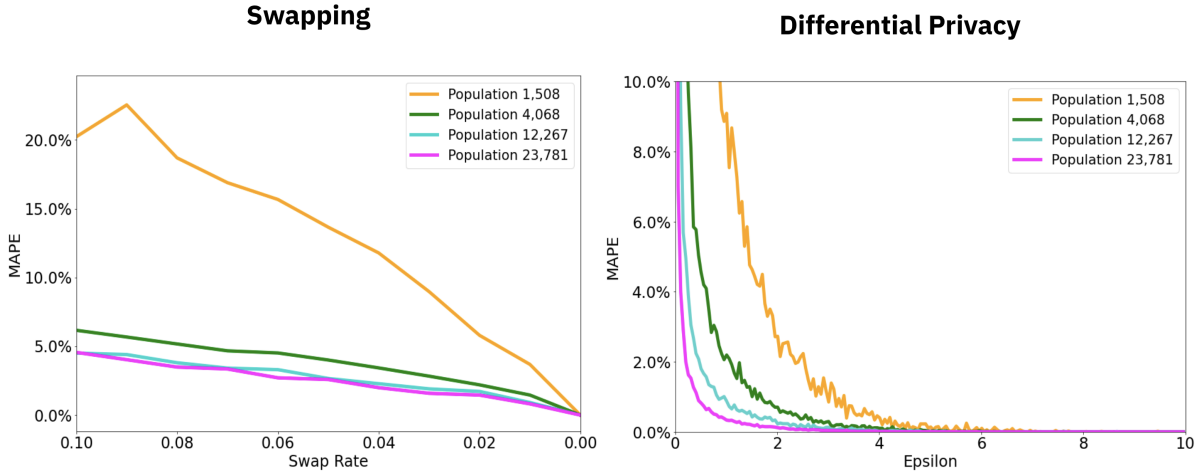

Figure 10: MAPE of swapping and DP for tracts of varying population sizes: RI-51000 (pop. 1,508), AL-5400 (pop. 4,068), AL-100 (pop. 12,267), AL-400 (pop. 23,781)

4 Discussion

We plot the MAPE of DP and swapping for each of our census tracts, shown in Figures 1-7. For each tract, the left graph shows the MAPE over the total population, and the right graph shows the MAPE over only the ages 4 and 5. Swap rates are shown on the top x-axis, and epsilon values are shown on the bottom x-axis, where moving further right represents decreasing privacy.

The gray rectangles highlight the segments of the swapping curves with swap rates between 2% and 4%, which is the speculated estimated national range of percentage of households swapped by the Census Bureau in 2010. While this swap rate can vary from block to block depending on population makeup, it is a reasonable estimate. These epsilon values and swap rates shown are meant to capture a range of reasonable values, and are not meant to be exactly correlated.

Observe that for DP, the MAPE is lower for larger populations. This is because the amount of noise added is parameterized only by epsilon and does not vary across populations. In larger populations, where the number of individuals of each age is greater, the proportion of error to true age count is smaller.

After swapping, for the smallest tract (RI-51000 with a population of 1,508), the de-identified dataset shows very high error. For tracts that are larger, with a population of 4,000 or greater, the MAPE of swapped microdata is more consistent. From these results, it is clear that both mechanisms have higher error for smaller populations than larger populations. However, differential privacy performs more predictably and has more similar curves across populations of different sizes, especially for ages 4-5.

Moreover, we see that for swapped data, accuracy varied greatly between parameter values, especially for small populations. This is represented by the jagged nature of the MAPE: see Figure 5 for the total population, or Figure 4 for Ages 4-5, where the error for the highlighted 2-4% swap rates vary greatly. This was not the case for DP mechanisms, which provide a more predictable relationship between utility and privacy.

Thus, we see that both of these de-identification mechanisms had greater negative impact on data utility when run on smaller populations. This effect was especially pronounced in swapped data. We further observe that differential privacy provided a more consistent relationship between utility and privacy, with swapping producing dramatically varied MAPE results for close swap rates.

Limitations

Our study is limited by both data availability and computational capability.

We were unable to use the TopDown Algorithm, the differentially private disclosure avoidance system used by the U.S. Census Bureau in 2020, as the cost would have been infeasible. Moreover, we were limited by the inability to access both the ground truth data and true swapping implementations used by the U.S. census. We were required to use synthetic data, which could produce different results than true census data, and we were forced to estimate the exact implementation of the census' swapping mechanisms. However, we believe that our design choices are grounded in publicly available Census Bureau guidance, and should be generally representative of the true data and mechanisms.

References

[cen21] Determining the privacy-loss budget: Research into alternatives to differential privacy. `https://www2.census.gov/about/partners/cac/sac/meetings/2021-05/presentation-research-on-alternatives-to-differential-privacy.pdf`, 2021.

[CRB22] Miranda Christ, Sarah Radway, and Steven M Bellovin. Differential privacy and swapping: Examining de-identification's impact on minority representation and privacy preservation in the us census. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1564–1564. IEEE Computer Society, 2022.