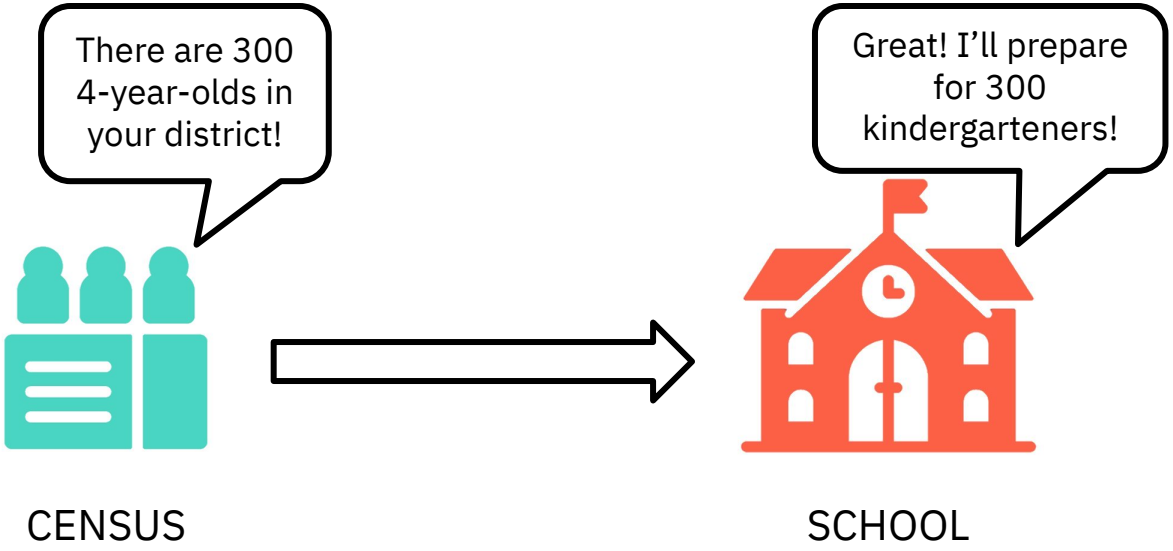
Counting Kindergartners: De-Identification's Impact on Student Representation

Sarah Radway, sarah.radway@tufts.edu
Miranda Christ, mchrist@cs.columbia.edu

Census data use:

Planning/Funding For Incoming Students



PRIVACY of Census Data

- Census data contains sensitive information about children and their families.
- Participation rates are impacted by privacy concerns.
- 13 U.S.C. §9(a)(2): Census data must not be personally identifiable

Ensuring privacy is integral to census participation.

PRIVACY of Census Data

291 kindergarten-age children

Age	Count
…	…
4	291
…	…

data de-identification

Age	Count
…	…
4	300
…	…

SCHOOL

UTILITY of Census Data

291 kindergarten-age children

Age	Count
…	…
4	291
…	…

data de-identification

Age	Count
…	…
4	180
…	…

SCHOOL

UTILITY of Census Data

“We don’t need to hire more teachers.”

“We don’t need more funding.”

“We don’t need more classrooms.”

UTILITY of Census Data

"We don't need to hire more teachers."

"We don't need more funding."

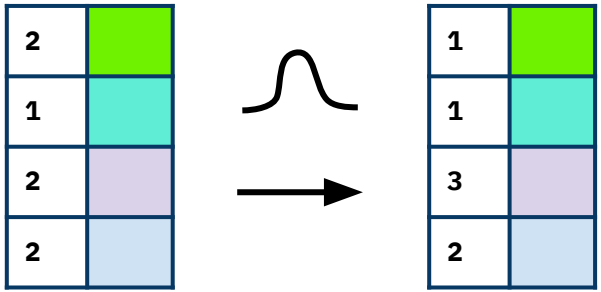
"We don't need more classrooms."

Inaccurate data may lead to insufficient allocation of resources.

How can we balance data **utility** and **privacy**?

How can we balance data **utility** and **privacy**?

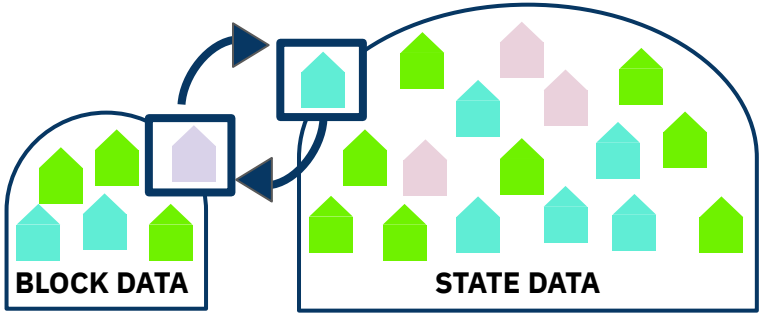
2020
Differential Privacy



DP & The Comparison Problem

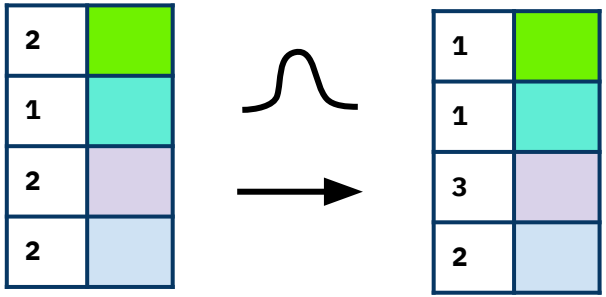
- As we know, DP isn't perfect.
- Many works have critiqued the ability of differential privacy to preserve data utility [NCAI 2019][Ruggles et al. 2019][Wezerek, Van Riper 2020]
- However, we need to contextualize the results of DP-data into those of alternatives [Christ et al. 2022].

2010
Swapping



- Exchange of data about individuals between groups
- Prioritizes unique entries
- **swap rate**: proportion of data to be swapped; estimated 2-4% nationally, with uneven geographical distribution

2020
Differential Privacy



- Add random noise (parameterized by ε)
- Privacy guarantee: changing one person's data only changes the de-identified data "a little bit"
  - **Higher ε:** higher accuracy; lower privacy

RESEARCH QUESTION:

How do these de-identification methods impact granular data, like single-year-of-age?

Our Approach

- Generate synthetic data for 8 tracts **of varying sizes** from synthetic state data.

- Create census-like swapping and DP algorithms, and run them on the synthetic tract data.

- Compare the accuracy of data produced by DP and swapping **at reasonable parameter values**.

Swapping

To de-identify a tract dataset:

- Select households within the tract to swap, **prioritizing unique households.**
- For each household to swap:
  - Find a similar household in the state population to swap. Similarity is based on:
    - Household Size
    - # of Household Members: Age >=18
    - # of Household Members: Age <18
- **Swap rate:** percentage of households that are swapped

Differential Privacy

To de-identify a tract dataset:

- Sum across queries to create table of counts.

  Queries used:

  - Age
  - Age & Sex

- Add noise to counts using geometric mechanism.
  - Minimal post processing: setting negative values to 0

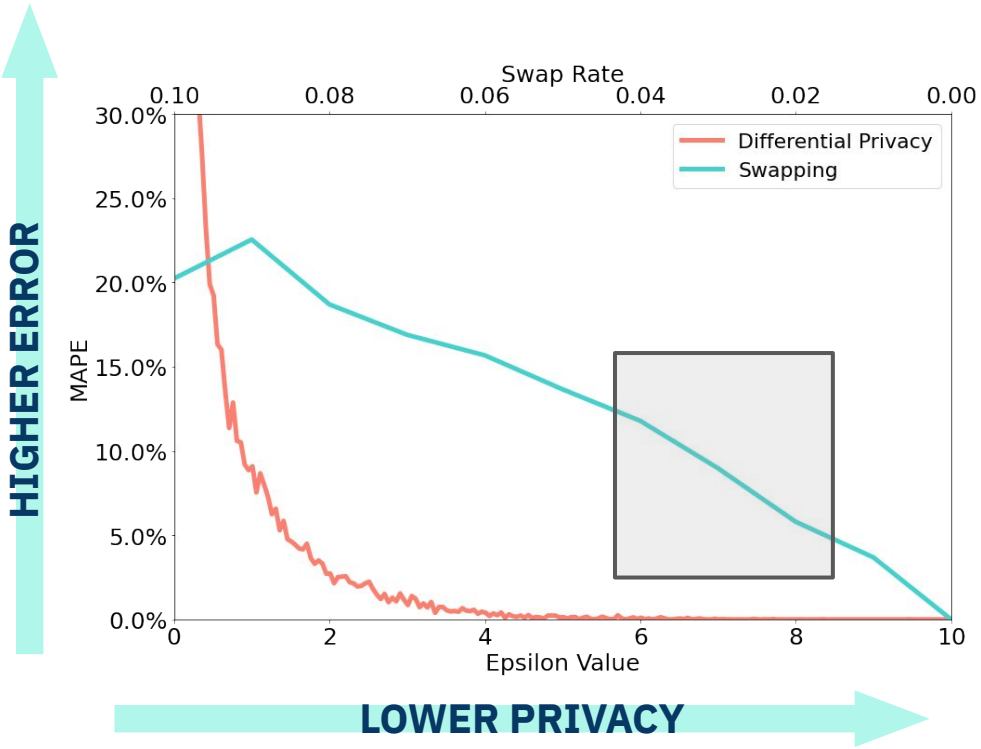
Accuracy

- Create histograms of single-year-of-age counts for ground truth data and de-identified data.
- Use ground truth histogram and de-identified histogram to compute Mean Absolute Percentage Error (MAPE).

  Look specifically at:

  - Total Population
  - Population Ages <18 Years
  - Population Ages 4-5 Years

Figures



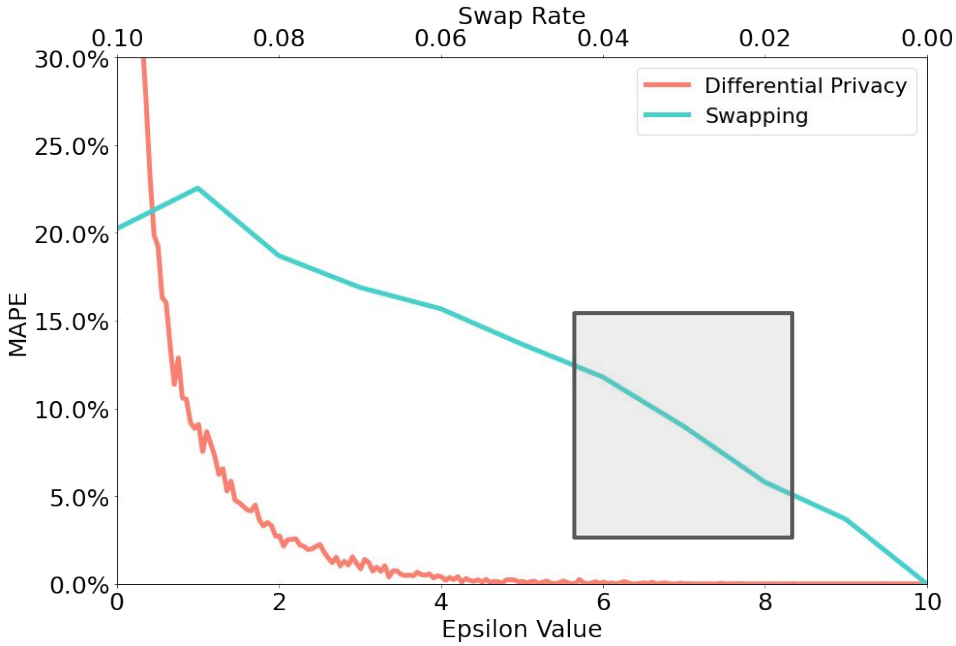
= National estimated swap rate

RI 51000 - **Population ~1,500**

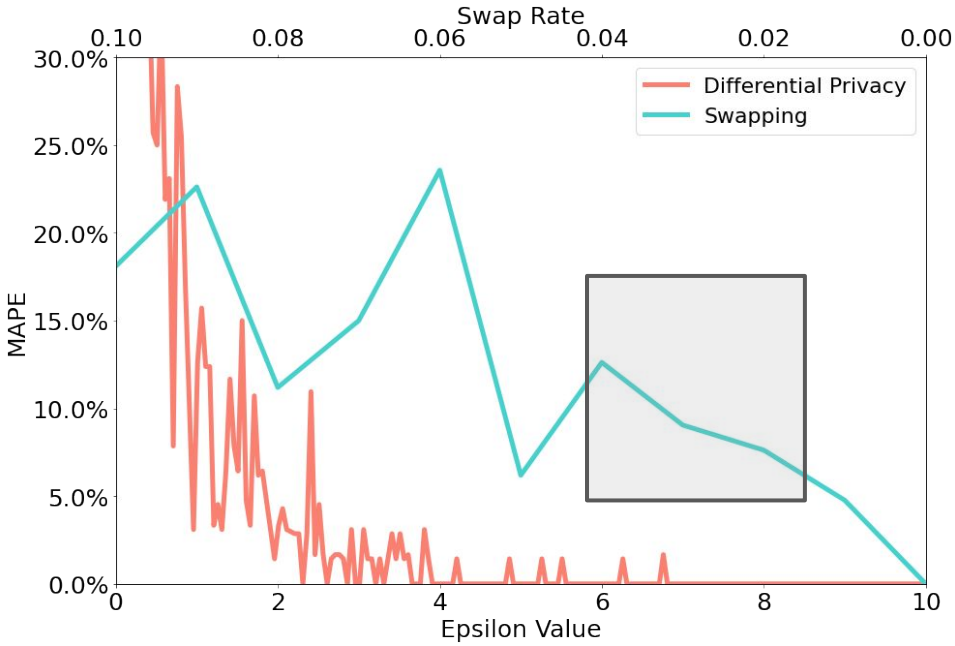
Results

= National estimated swap rate

Total Population



Ages 4-5

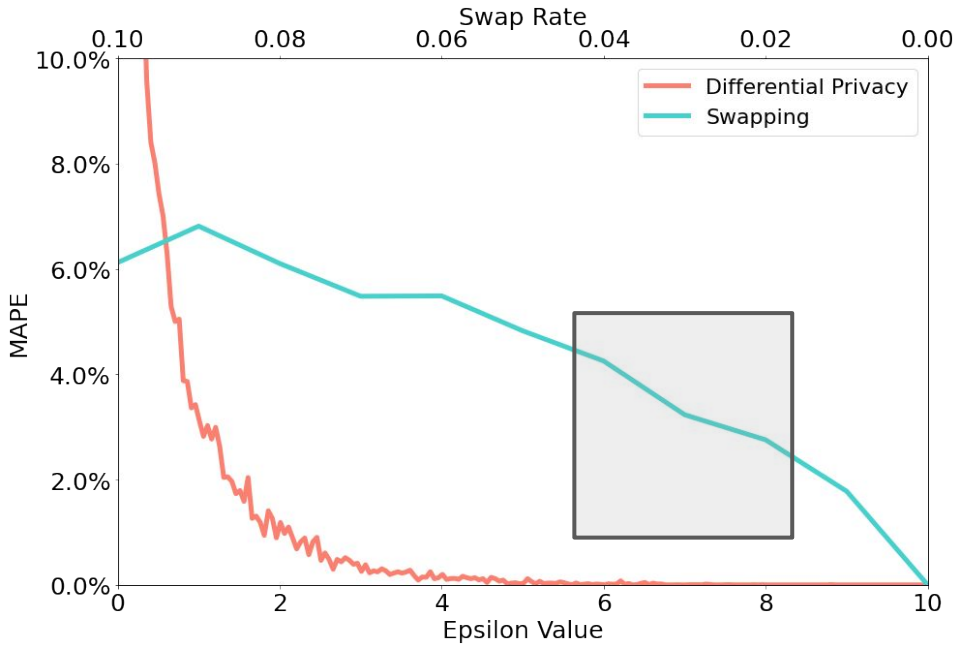


RI 15800 - **Population ~3,800**

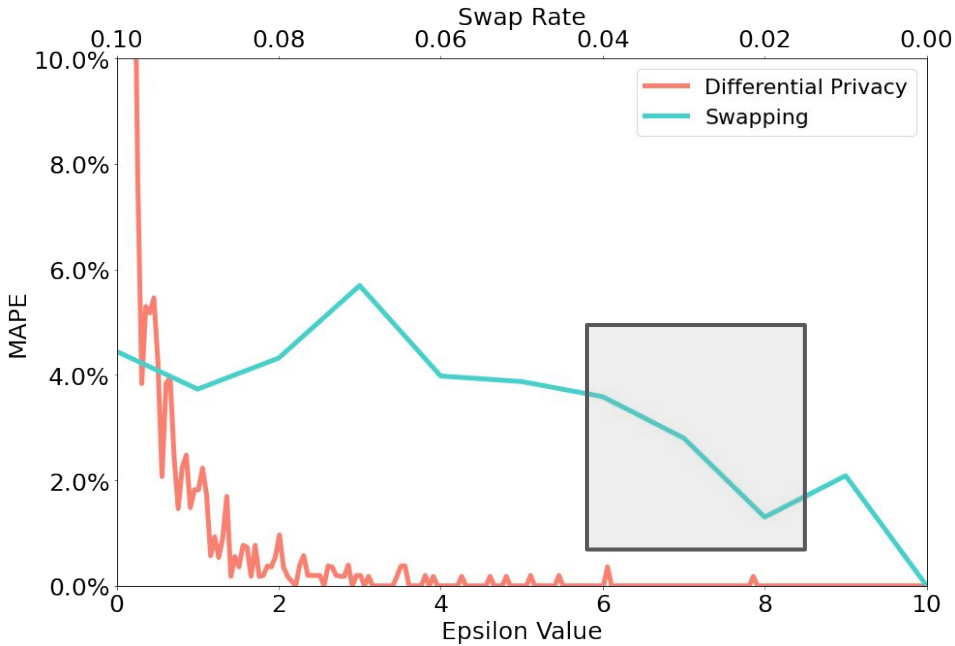
Results

= National estimated swap rate

Total Population



Ages 4-5

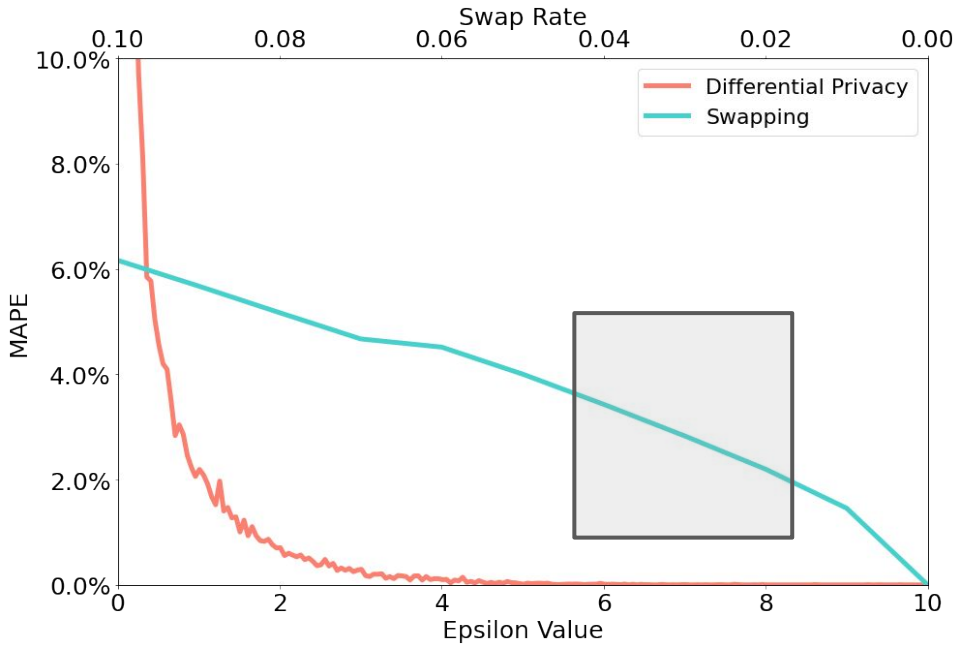


AL 5400 - **Population ~4,000**

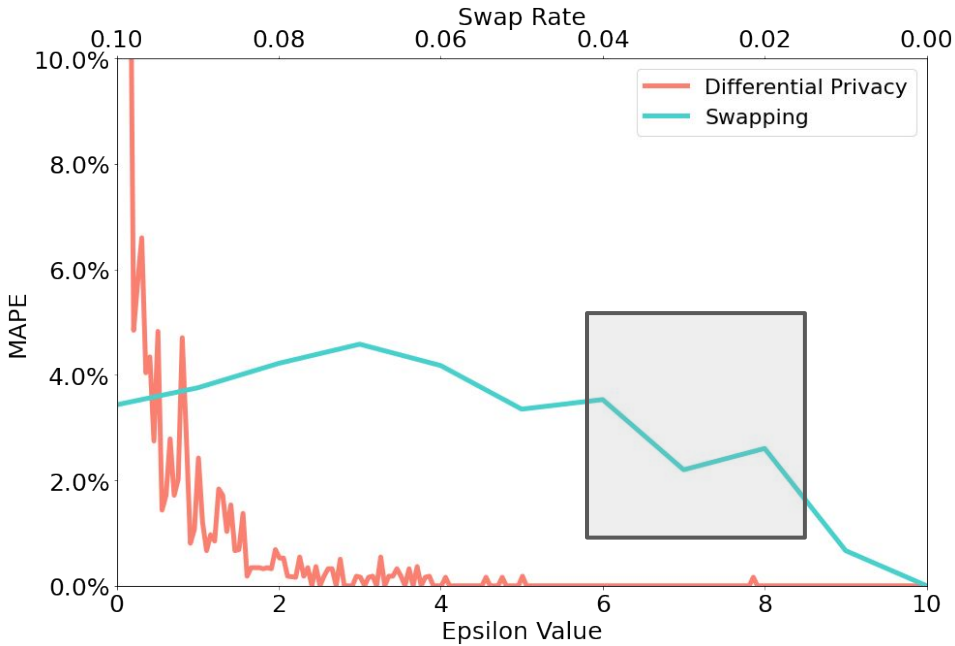
Results

= National estimated swap rate

Total Population



Ages 4-5

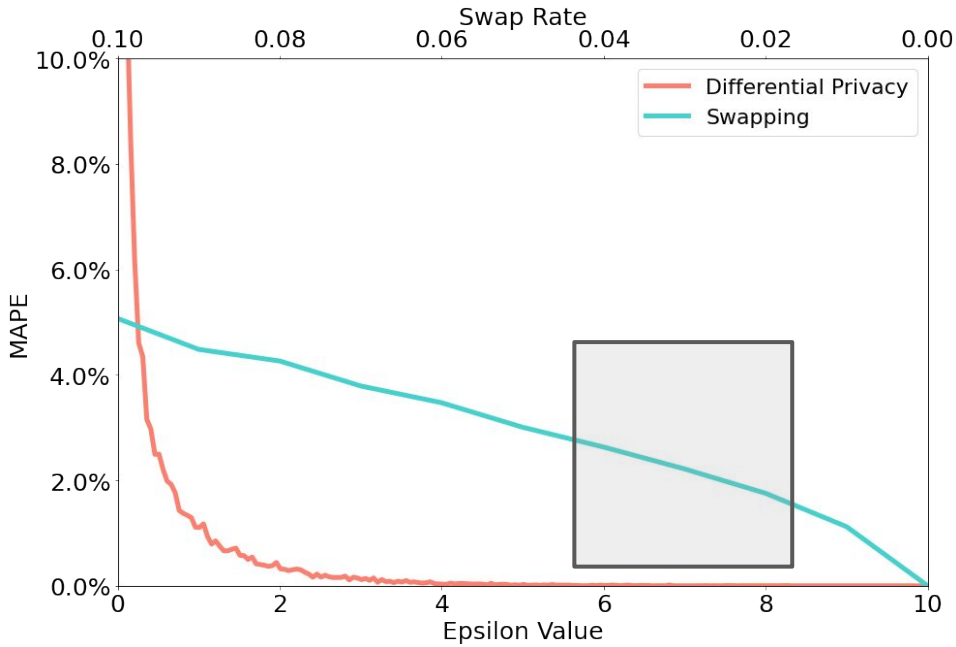


AL 10001 - **Population ~8,000**

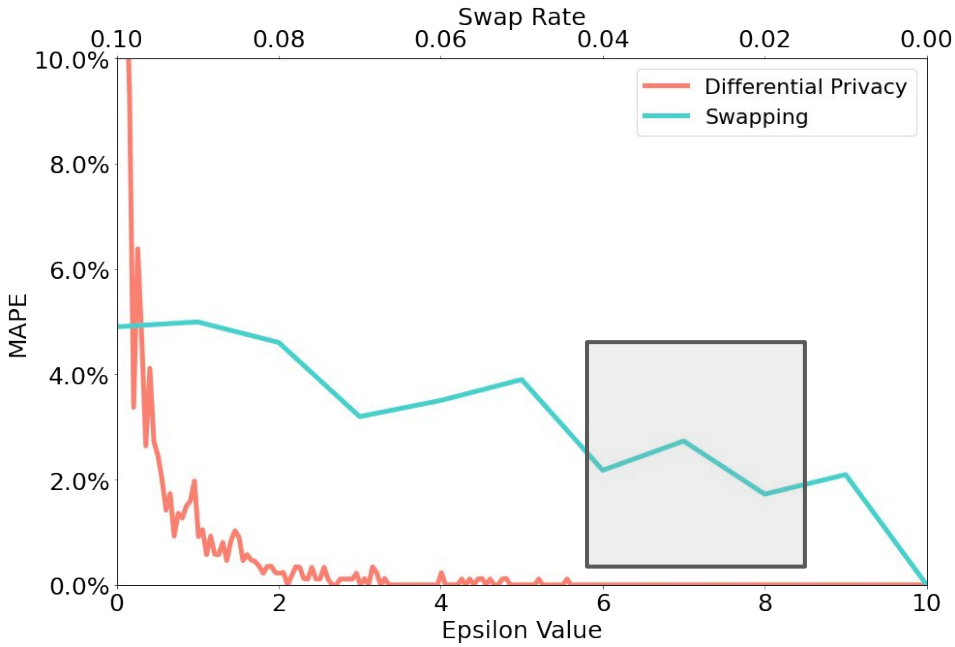
Results

= National estimated swap rate

Total Population



Ages 4-5

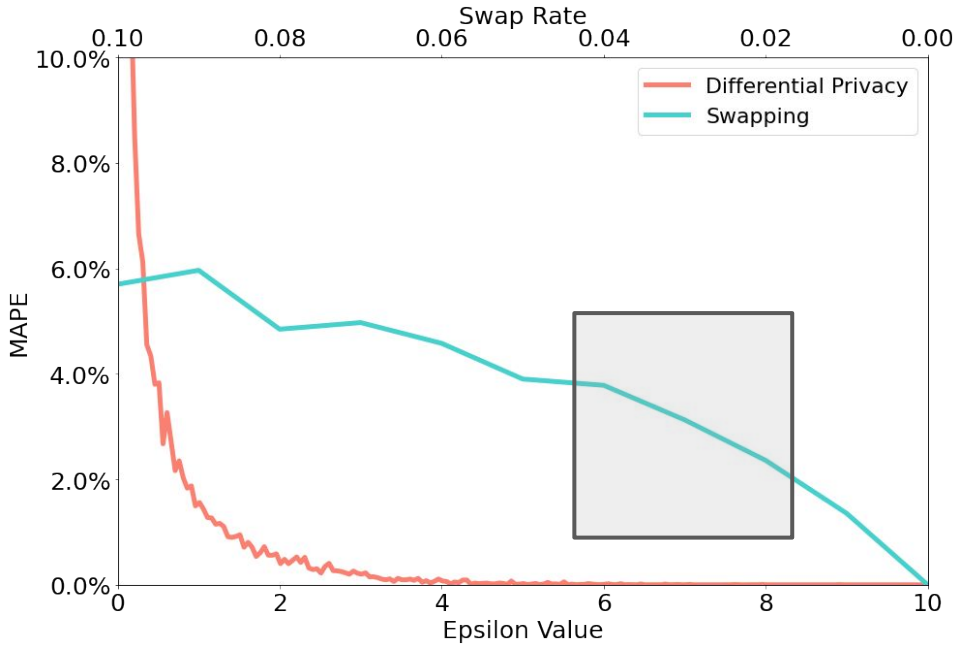


RI 20101 - **Population ~8,500**

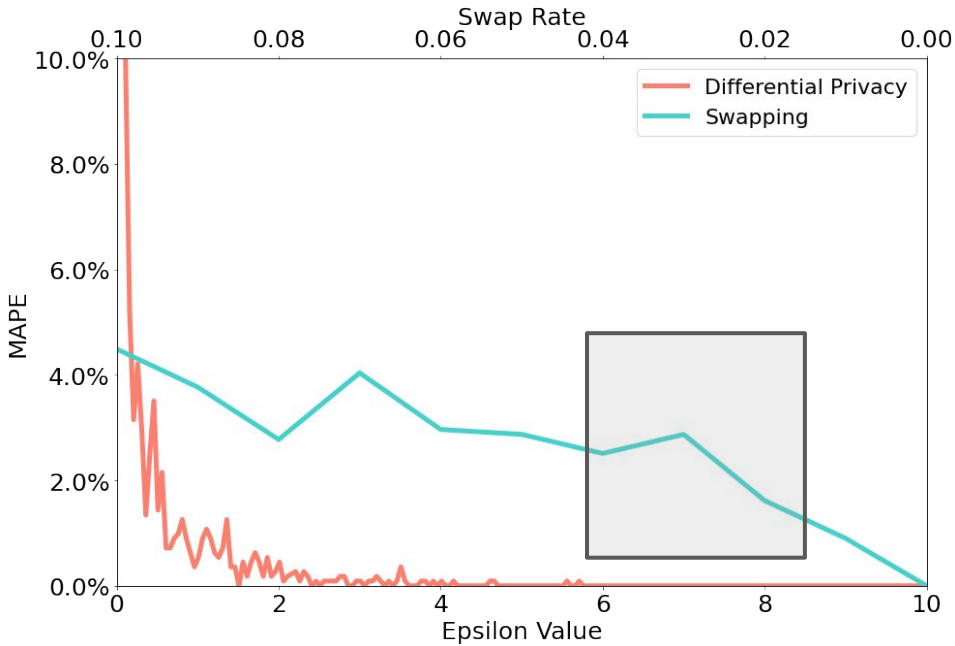
Results

= National estimated swap rate

Total Population



Ages 4-5

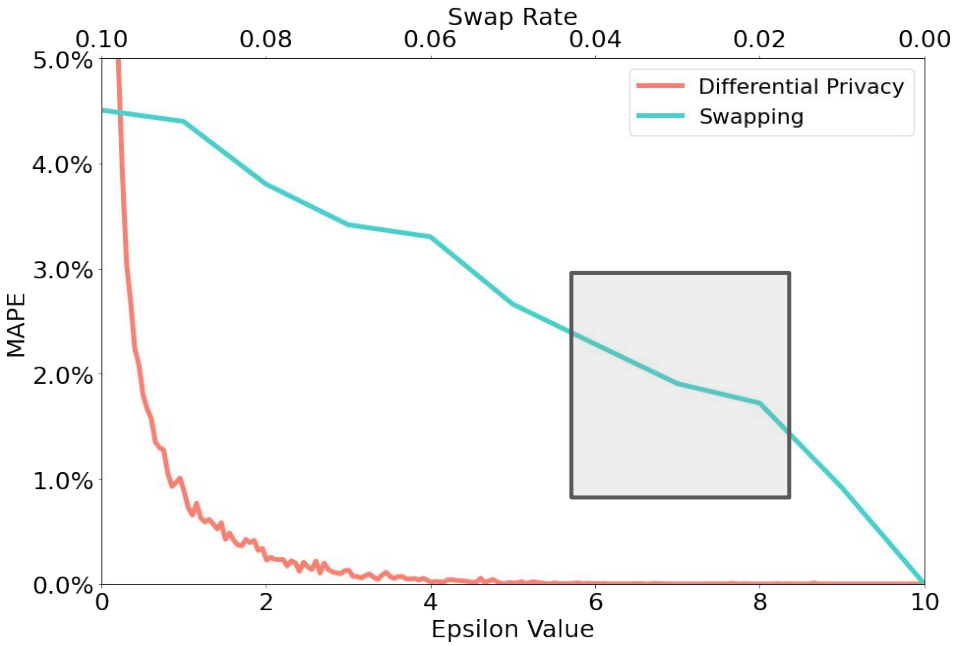


AL 100 - **Population ~12,250**

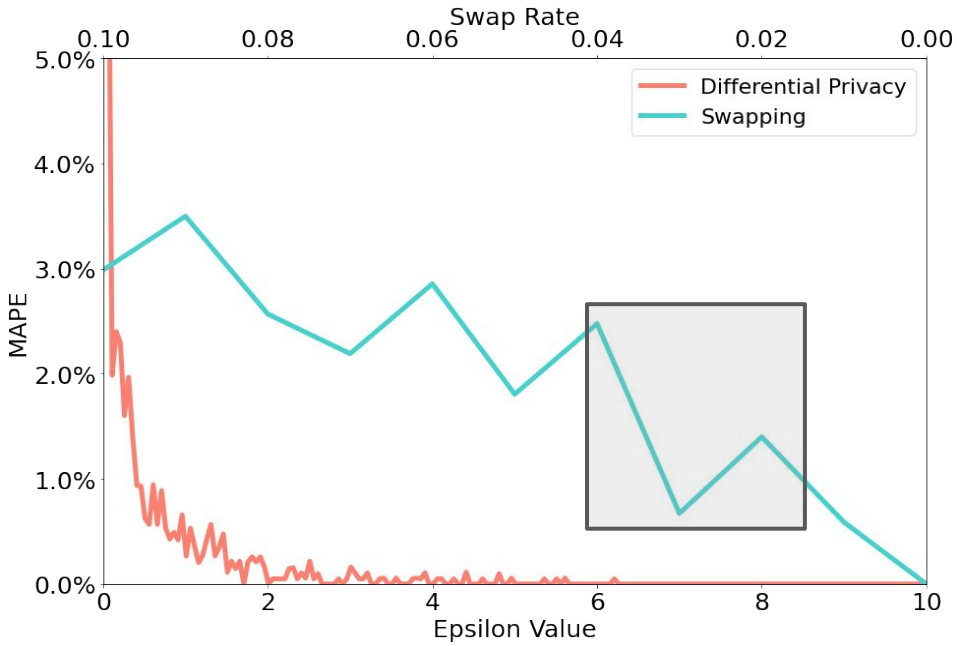
Results

= National estimated swap rate

Total Population



Ages 4-5

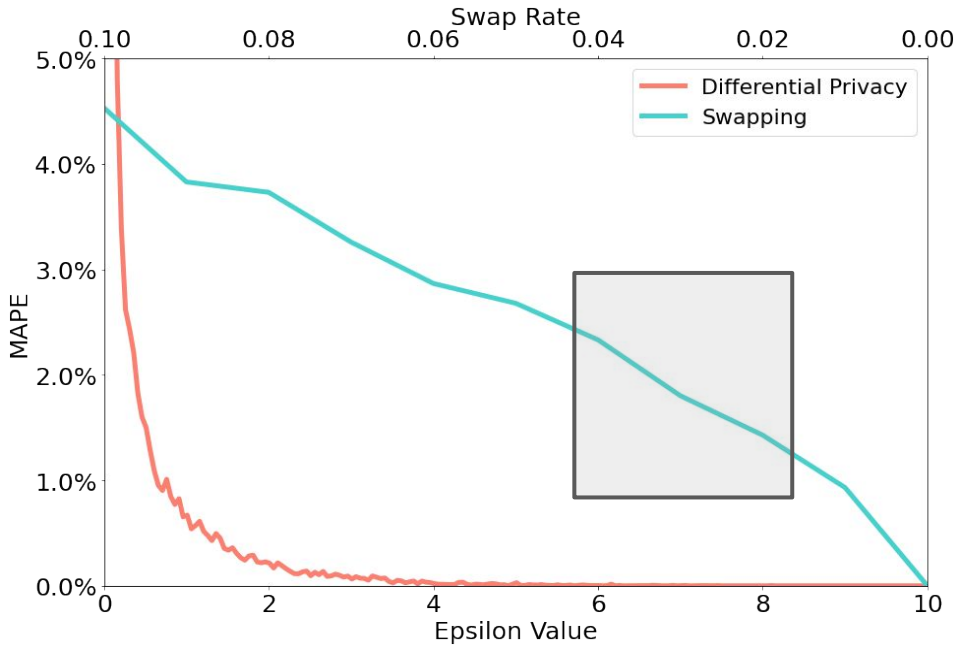


AL 10704 - **Population ~19,500**

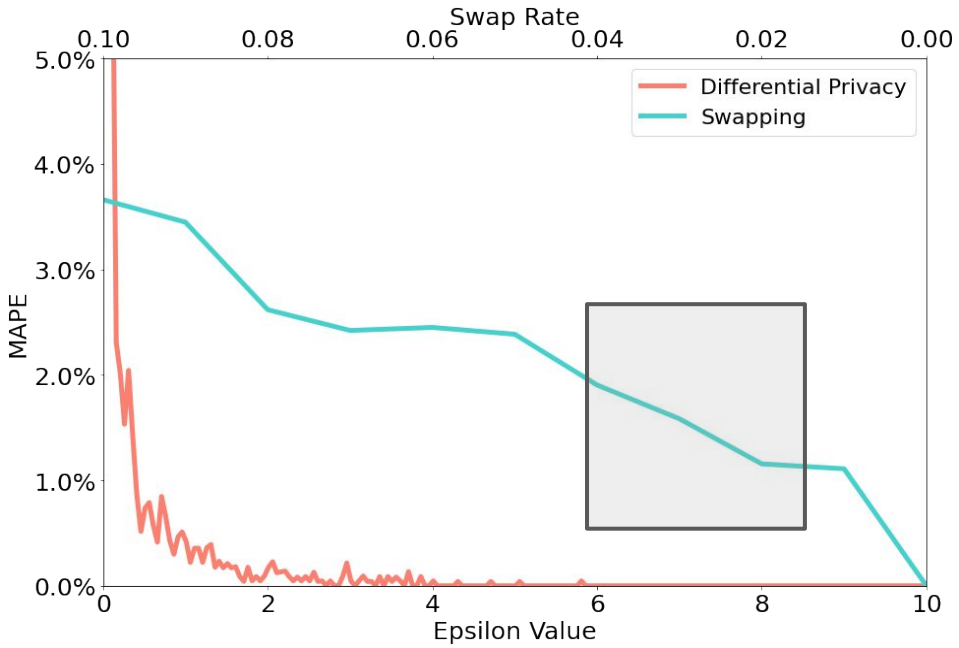
Results

= National estimated swap rate

Total Population



Ages 4-5

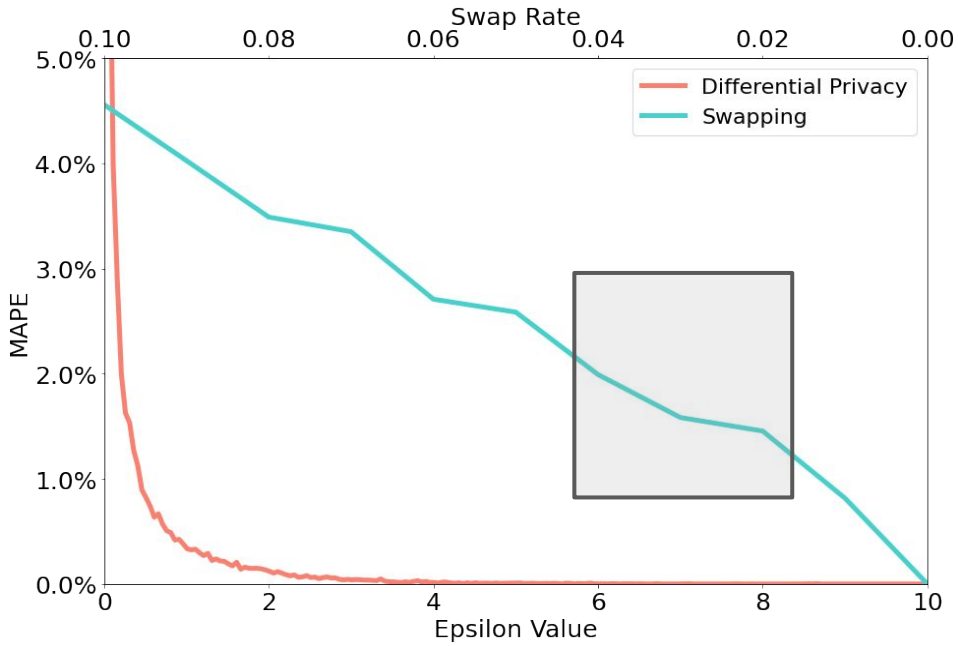


AL 400 - **Population ~23,750**

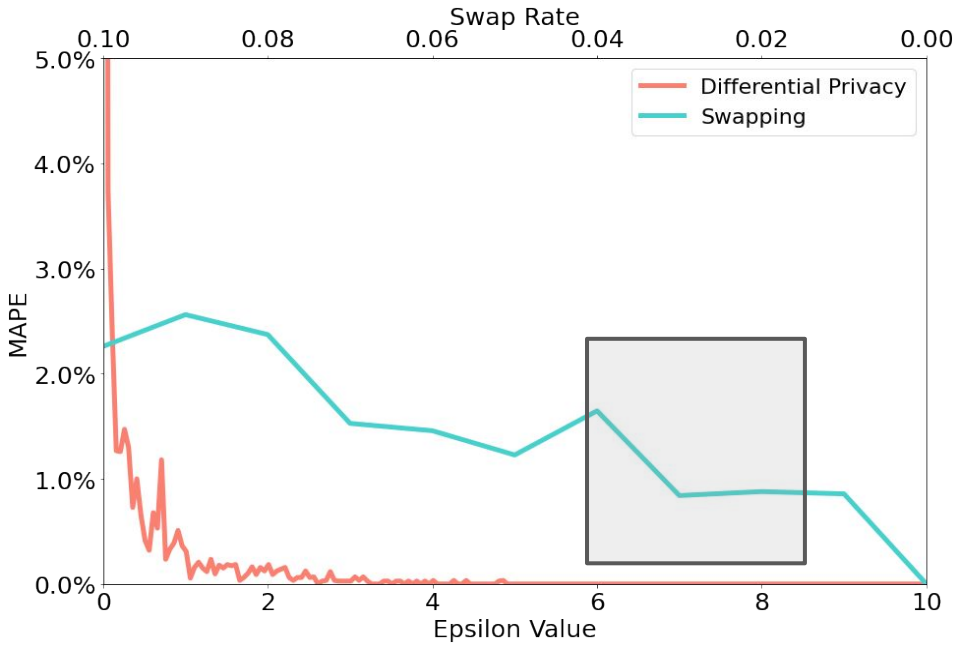
Results

= National estimated swap rate

Total Population



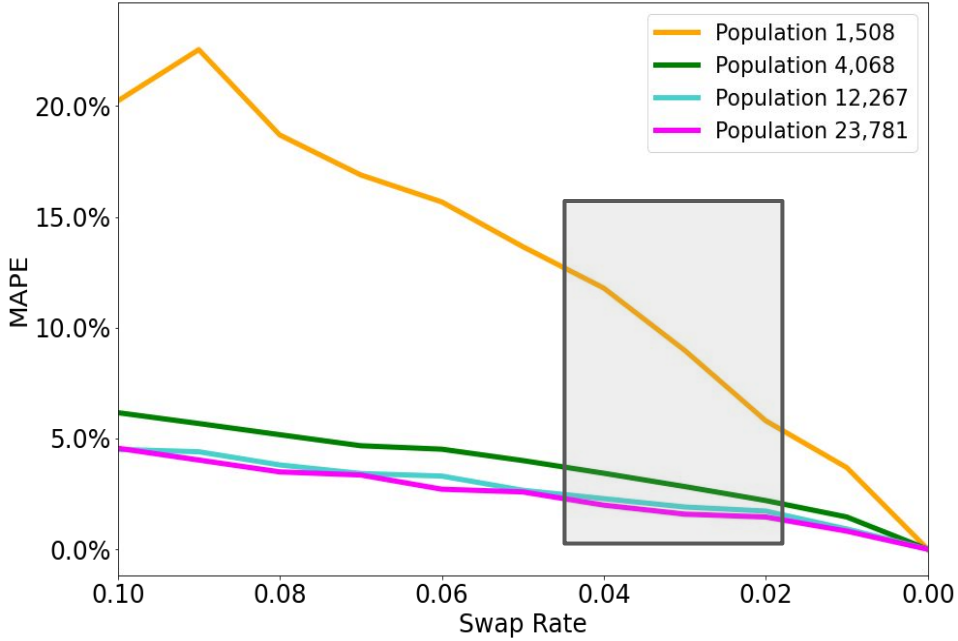
Ages 4-5



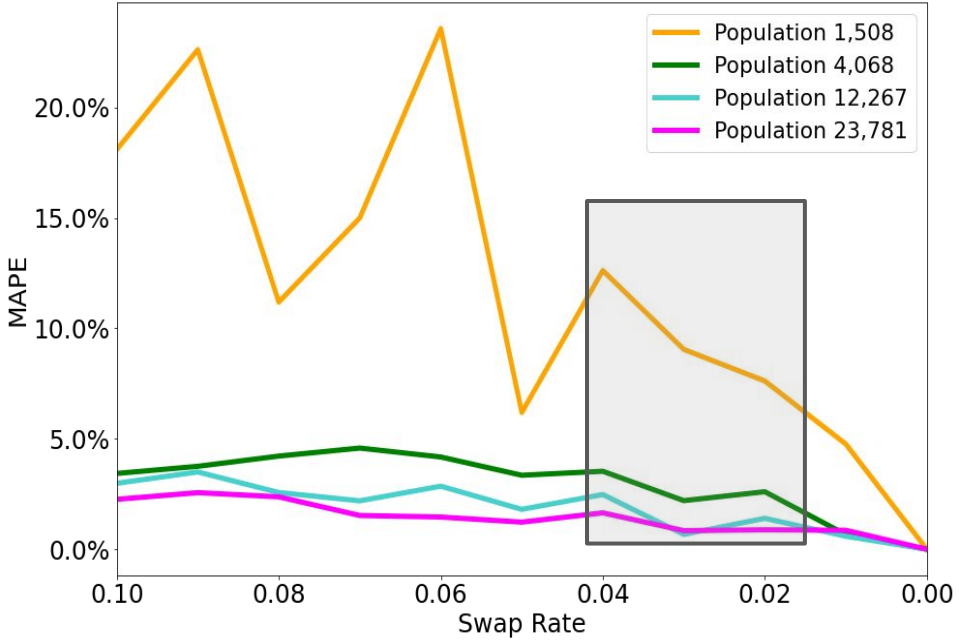
Swapping

= National estimated swap rate

Total Population

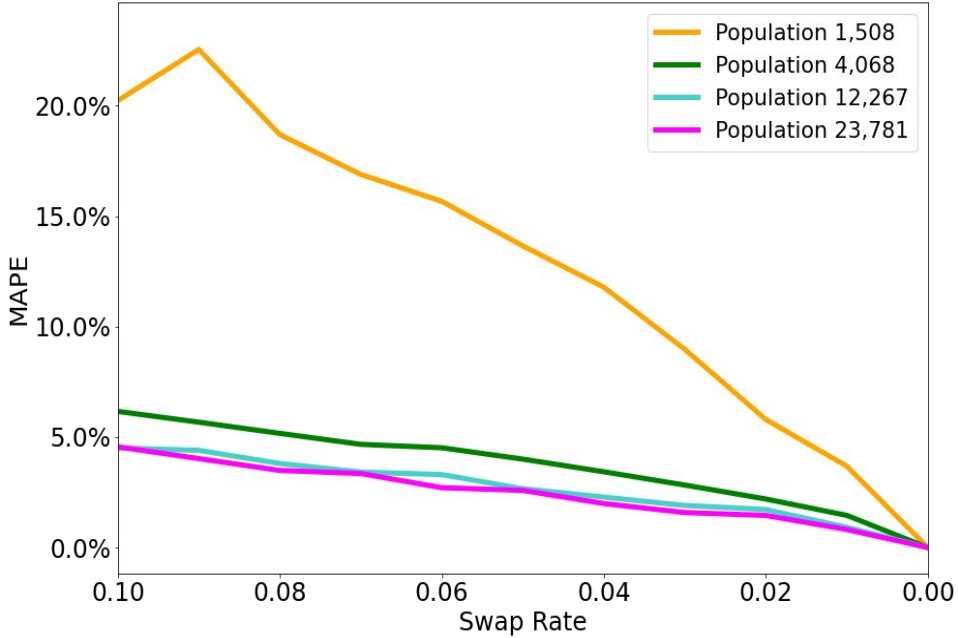


Ages 4-5

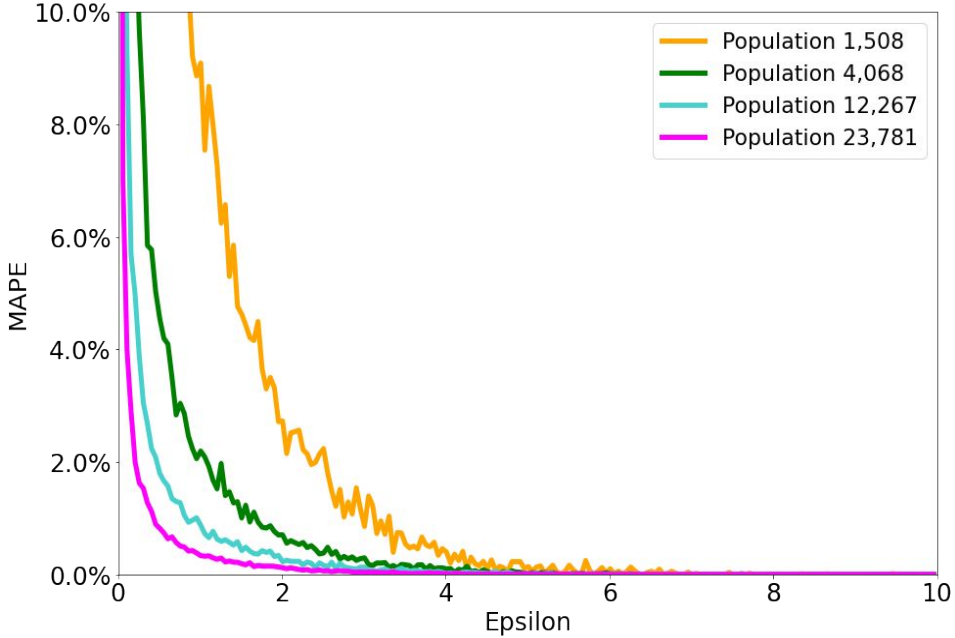


Not so different after all...

Swapping



Differential Privacy



Takeaways

Both mechanisms have higher error for smaller populations.

Swapping Mechanisms:

- Accuracy varies more between parameter values, especially for small populations.

DP Mechanisms:

- Performs more predictably across different population sizes.
- Provides a predictable relationship between utility and privacy.

References

[Christ et al. 2022]: M. Christ, S. Radway, and S. Bellovin. "Differential Privacy and Swapping: Examining De-Identification's Impact on Minority Representation and Privacy Preservation in the US Census." 2022 IEEE Symposium on Security and Privacy (SP). IEEE Computer Society, 2022.

[NCAI 2019]: National Congress of American Indians, “Differential privacy and the 2020 U.S. Decennial Census: Impact on American Indian and Alaska Native data.” [Online]. Available: https://www.ncai.org/prc/2020 Census and AIAN data FINAL 9 11 2019.pdf

[Ruggles et al. 2019]: S. Ruggles, C. Fitch, D. Magnuson, and J. Schroeder, “Differential privacy and census data: Implications for social and economic research,” in AEA Papers and Proceedings, vol. 109, 2019, pp. 403–08

[Wezerek, Van Riper 2020]: G. Wezerek and D. Van Riper, “Changes to the census could make small towns disappear,”The New York Times, Feb 2020

Thank You! Any Questions?

Sarah Radway, sarah.radway@tufts.edu
Miranda Christ, mchrist@cs.columbia.edu

We thank Steven M. Bellovin for his help.

Figure Credits:
Student by Adrien Coquet from NounProject.com
Census by Solomakhina Maria from NounProject.com
School by IconHome from NounProject.com