Differential Privacy in Applied Social Science Settings

Thomas Chen

University of California

Mentor: Ruobin Gong
In collaboration with Kelly Xu

July 18, 2024

Data Privacy

- We live in a world where data, especially public datasets is used for a multitude of purposes, from research studies to ML and AI.
- However, there is increasing concerns over the risk of sharing public data.
- Malicious agents can use public databases to find out sensitive information about specific individuals.
- Traditionally, Statistical disclosure control (SDC), or limitation (SDL) have been used to limit privacy risk against certain attacks

Differential Privacy

- Differential privacy (DP) is a framework to quantify the amount of privacy provided by a algorithm.
- Given ϵ, a sanitation algorithm $\mathcal{M}$ is $\epsilon - DP$ for all $S \subset range(\mathcal{M})$ and for all X and X' that differ by one record, it fulfills the following equation:

$$\frac{\Pr(\mathcal{M}(X) \in S)}{\Pr(\mathcal{M}(X') \in S)} \leq \exp(\epsilon)$$

- We want algorithms that satisfy $\epsilon - DP$, while also preserving as much useful statistics from the original dataset
- One of our main sanitation algorithms will be synthetic data generation: Generating artificial dataset from the original dataset.

Main problem

- Implement a differentially private synthetic data generator on different datasets
- Learn how to generate synthetic datasets that effectively preserve usability while satisfying differential privacy.
- We looked at 3 different studies from 2 different datasets, but we primarily focused on a study using the Panel Study of Income Dynamics
- It is the longest running longitudinal household survey in the world, frequently used in many social science studies.

Synthetic Data Generation

- We first implemented a synthetic data generator called Datasynthesizer, which is based on a method called PrivBayes that was used in the NIST PSCR Differential Privacy Synthetic Data Challenge in 2019
- The method is based on Bayesian Networks, a probablistic model that represents the distribution of the variables but also the dependencies between them.
- The algorithm first generates a Bayesian Network based on the variables and creates a probability distribution
- Then it generates the synthetic data using the probability distribution.
- Noise is injected in both processes to satisify DP.

Bayesian Network

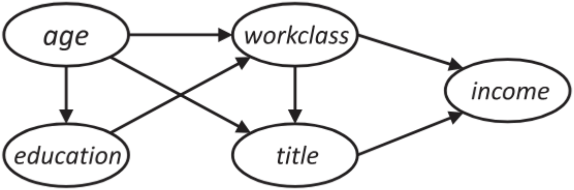


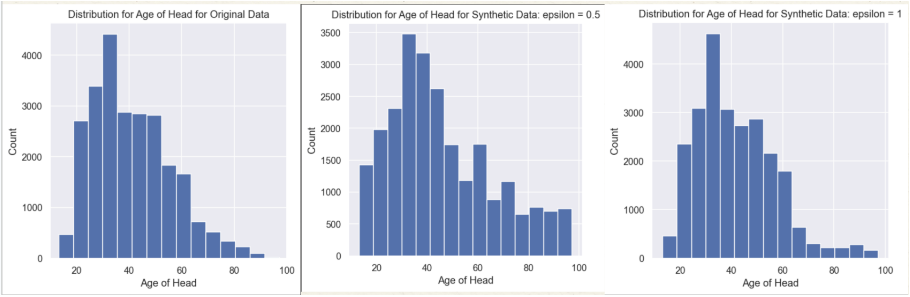
Fig. 1. A Bayesian network $\mathcal{N}_1$ over five attributes.

Figure: Taken from PrivBayes. Zhang J. et.al (2017)

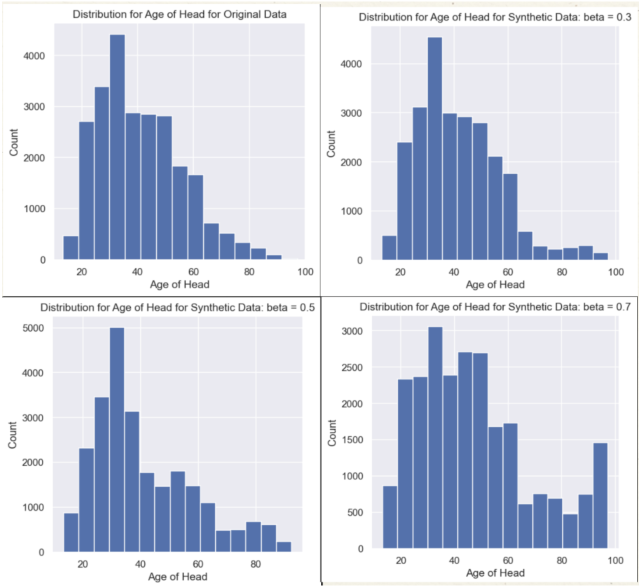
Variables to Consider

- ϵ, the amount of data privacy budget we have. The more budget, the better the synthetic data since we are injecting less noise
- β, How much privacy to allocate to building the network vs generation
- Maximum degree of the network. Bigger/complex networks could better describe the relationships between variables, but the trade-off is that more noise is injected to the network as a whole to ensure DP

Marginal Distributions of Head Age based on $\epsilon(\beta = 0.3)$



Marginal Distributions of Head Age based on $\beta(\epsilon = 1)$



Main Study

- The main study that we did analysis on is the "New Estimates of the Sandwich Generation in the 2013 Panel Study of Income Dynamics" by Friedman, et.al (2014).
- The study looks at the transfer of wealth and time in people who have parents and children.
- We replicated four of the tables in the study based on the original data, then compared to results to when we used DP synthetic data.
- We examined how changing the variables affected the effectiveness of the analysis.

Table 1 in the Sandwich Study

Table 1. Percent of Women and Men with Children and Parents, by Age (PSID 2013 Family Roster & Transfer Module)

	Women				Men			
	Overall	35–49	50–64	65–75	Overall	35–49	50–64	65–75
Both	44.9	41.6	59.0	17.7	44.3	31.0***	61.9†	32.1***
Child(ren) Only	26.9	3.6	28.7	75.7	20.9***	2.7	21.9***	61.6***
Parent or In-law only	24.6	52.9	8.0	0.9	31.1***	63.8***	11.7**	1.4
None	3.6	1.9	4.4	5.7	3.8	2.5	4.5	5.0
Married (%)	67.6	70.5	67.7	60.9	77.0***	75.4**	77.3***	80.3***
N	4,688	2,106	2,008	574	3,952	1,768	1,658	526

Note: Weighted using 2013 individual weights. Unweighted *N*.
†$p < .10$; *$p < .05$; **$p < 0.01$, ***$p < 0.001$.

Table Distribution Comparisons: Male Proportions

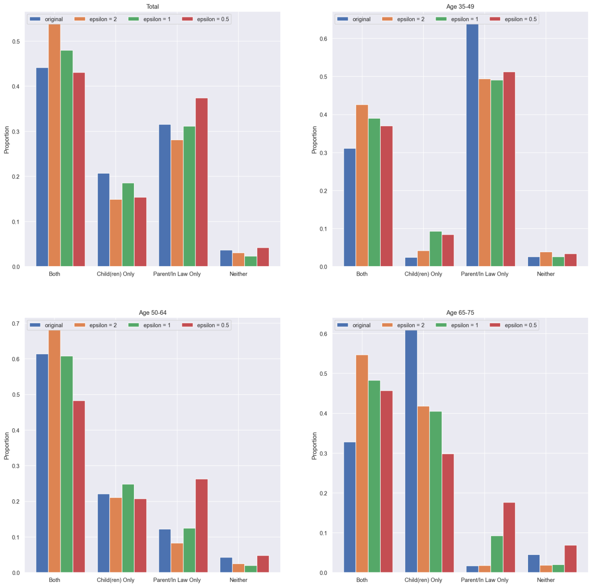
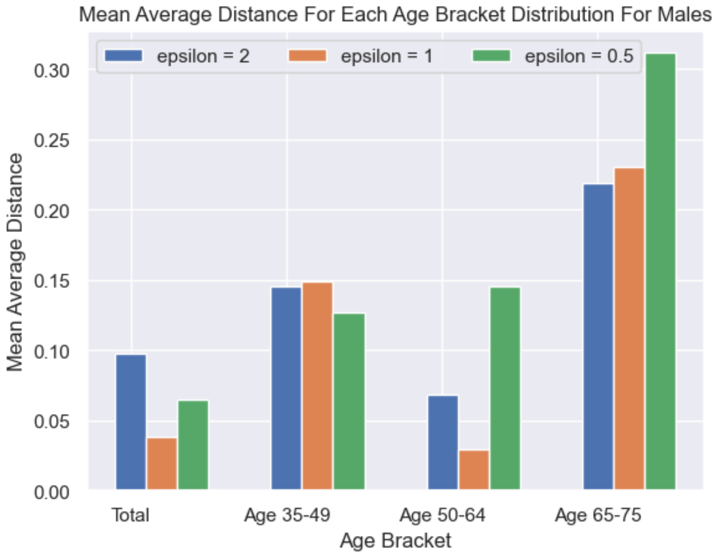
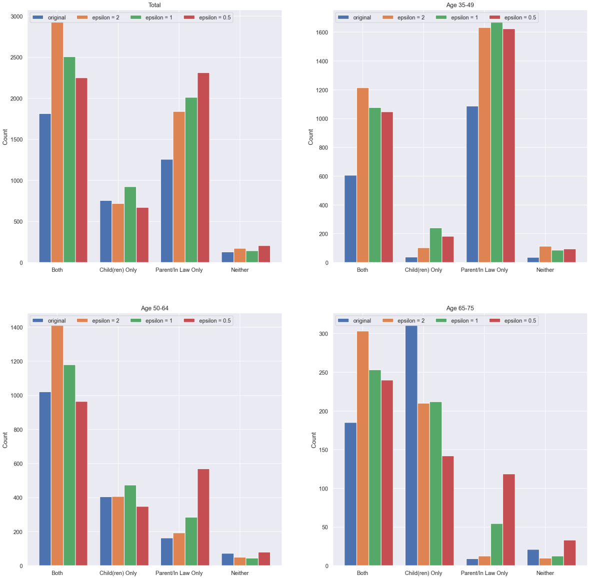


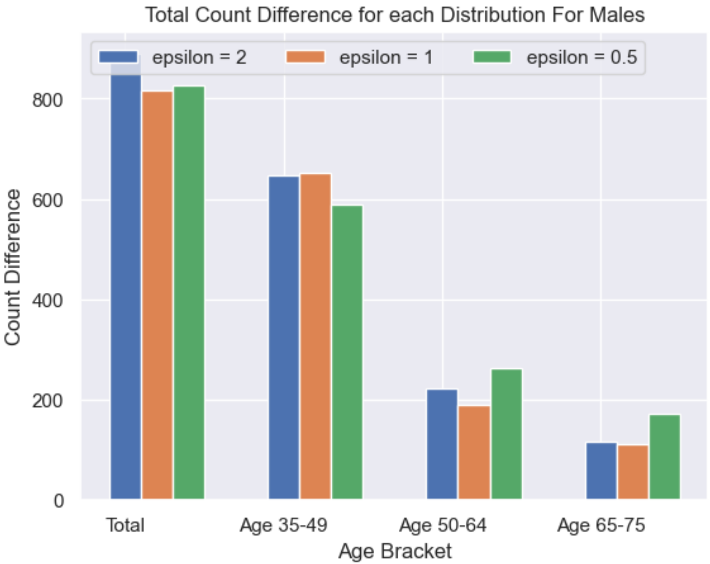
Table Distribution Comparisons: Male Proportions Mean Average Error



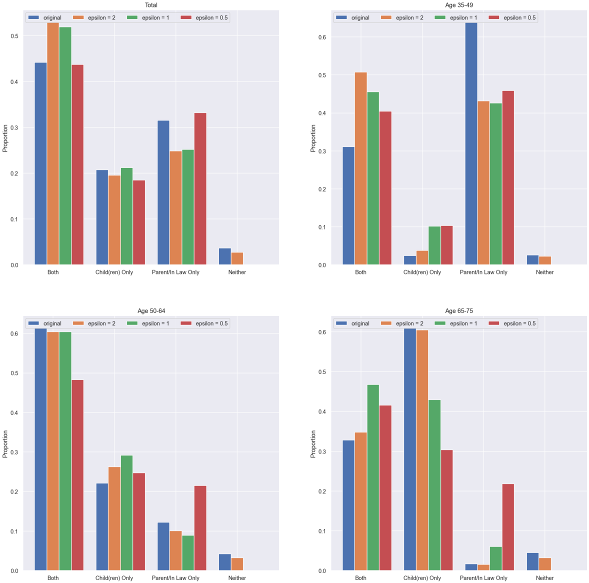
Distribution Comparisons: Male Counts



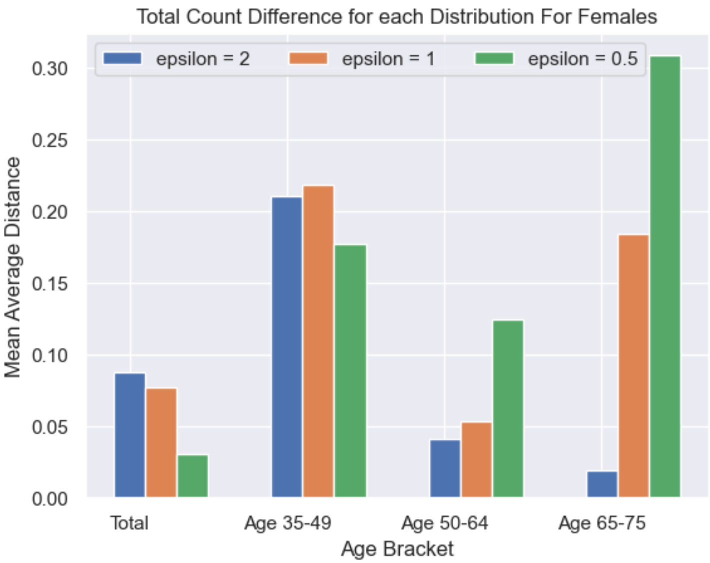
Distribution Comparisons: Male Counts Total Difference



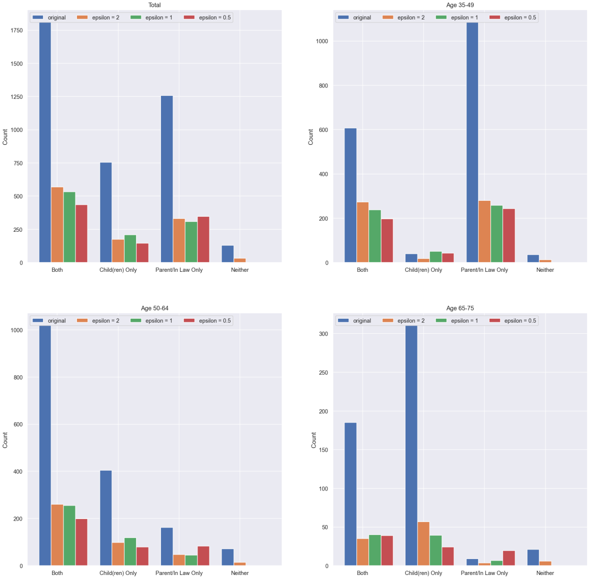
Distribution Comparisons: Female Proportions



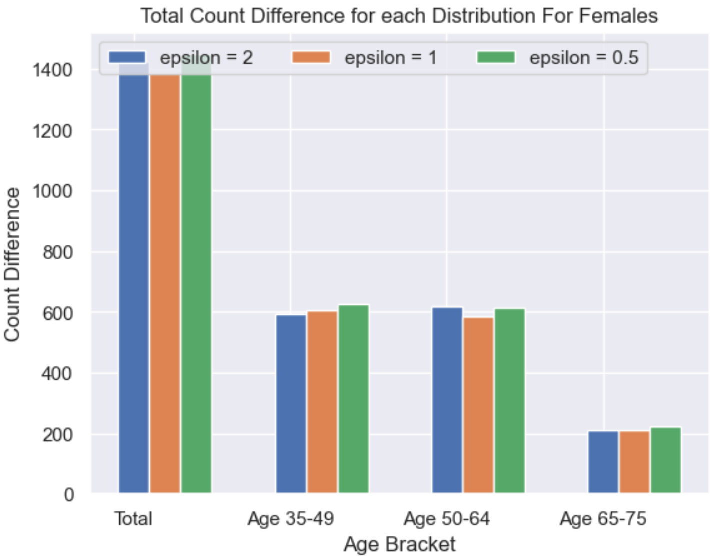
Distribution Comparisons: Female Mean Average Error



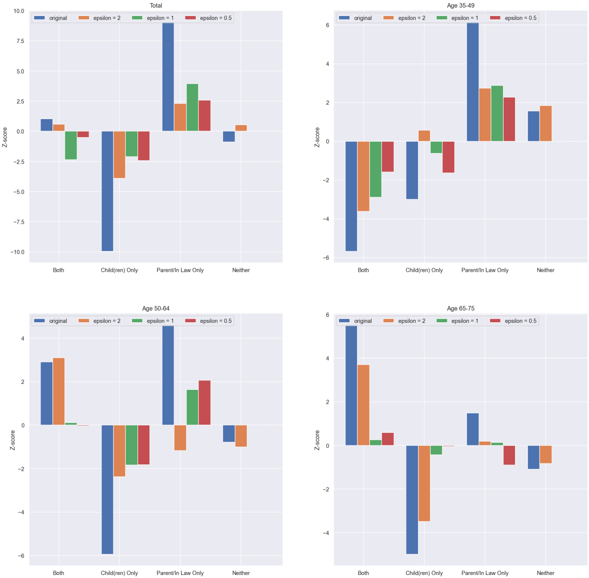
Distribution Comparisons: Female Counts



Distribution Comparisons: Female Counts Total Difference



Z score comparison



Future Goals

- Look at how data synthesis works with longitudinal studies
- Comparing these results to another DP synthetic data Generators
- Examining how other kinds of statistical analyses fare under these synthetic data generators.

Citations

- Bowen, C. M., & Liu, F. (2020). Comparative study of differentially private data synthesis methods. Statistical Science, 35(2). https://doi.org/10.1214/19-sts742
- Bowen, C. M., & Snoke, J. (2021). Comparative study of differentially private synthetic data algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge. Journal of Privacy and Confidentiality, 11(1). https://doi.org/10.29012/jpc.748
- Dwork, C., & Roth, A. (2014). The algorithmic foundations of Differential Privacy. now Publishers Inc.
- PSID Home. (n.d.). https://psidonline.isr.umich.edu/default.aspx
- Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., Xiao, X. (2017). Privbayes. ACM Transactions on Database Systems, 42(4), 1–41. https://doi.org/10.1145/3134428

Citations

- Ping, H., Stoyanovich, J., Howe, B. (2017). Datasynthesizer. Proceedings of the 29th International Conference on Scientific and Statistical Database Management. https://doi.org/10.1145/3085504.3091117
- New Estimates of the Sandwich Generation in the 2013 Panel Study of Income Dynamics
- Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Megan Schouweiler, and Michael Westberry. IPUMS CPS: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D030.V11.0

Acknowledgements

This work was carried out as a part of the 2024 DIMACS REU program at Rutgers University, supported by NSF grant CNS-2150186
Thank you to Prof. Lazaros Gallos and the DIMACS program for giving me the opportunity to research at Rutgers University.
Thank you to my mentor, Professor Roubin Gong for her time, effort, and encouragement in research. Thank you to my partner Kelly Xu for their help throughout the Summer.

Thank You