Differential Privacy in Applied Social Science Settings

Thomas Chen

University of California

Mentor: Ruobin Gong

June 4, 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 $S \subset \mathcal{M}$ 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.

PSID Dataset

- Main focus will be on The Panel Study of Income Dynamics (PSID)
- It is the longest running longitudinal household survey in the world, frequently used in many social science studies.
- Currently, there has not been implementation of a differentially private synthetic data generator on the database, and there has been no formal study on it

Goals for the summer

- Implement a differentially private synthetic data generator for the Panel Study of Income Dynamics
- Formalize and quantify the privacy on this dataset and evaluate how effective it will be
- Our first goal will be to apply a synthetic data generator called PrivBayes, which was used in the NIST PSCR Differential Privacy Synthetic Data Challenge in 2019

Thank You

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

Acknowledgements

This work was carried out as a part of the 2024 DIMACS REU program at Rutgers University, supported by NSF grant CNS-2150186