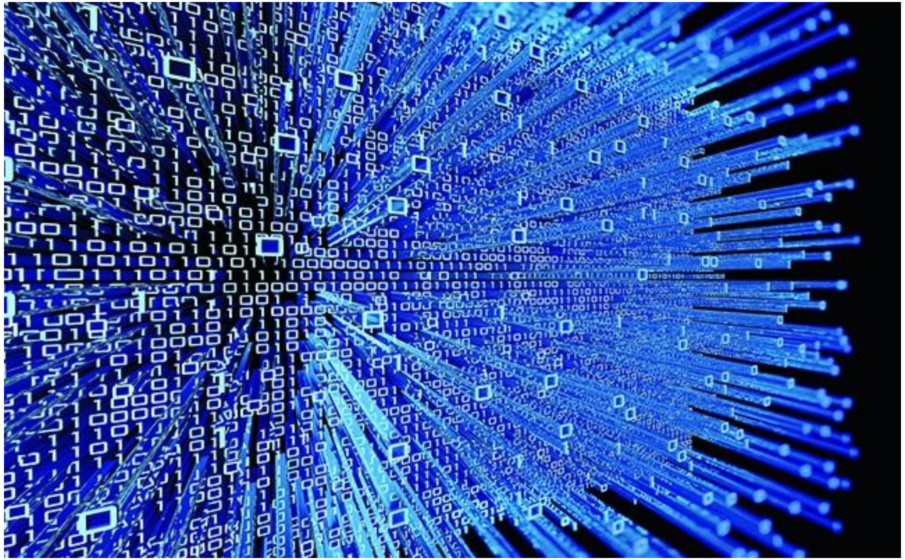
Data Privacy in the 21st Century

Student: Samarth Pusegaonkar
Mentor: Peng Shao

Introduction

- Lots of technology
- Lots of data
- Lots of applications for data
- Lots of personal information
- Lots of threats



How do companies protect us?

- California Consumer Privacy Act (CCPA)
- General Data Protection Regulation (GDPR, European)
- Regulations on the sale of data
- How is data modified to protect consumers/data subjects?

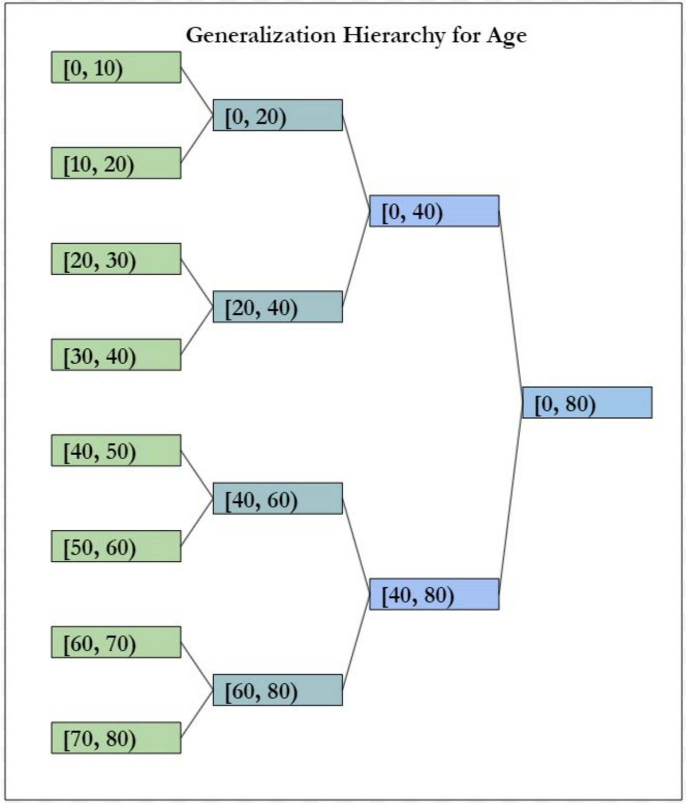
What is anonymization?

A Rigorous Definition of Anonymization

- “remove or perturb data to prevent adversaries from inferring sensitive information while ensuring the utility of the published data” (Beigi and Liu, 2020)
- Remove/Perturb → Methods of achieving anonymization
- Protecting sensitive information
- Ensuring utility

Methods of Achieving Anonymization

- Generalization, Suppression, Adding Noise
  - Laplacian Noise → Differential Privacy
  - Creating Hierarchies
  - Balancing various methods
- Anonymization Algorithms
  - K-anonymity
    - Induced Equivalence Classes
  - L-diversity
    - Well represented Values
  - T-closeness
    - Variational/Kullback Leibler Distance



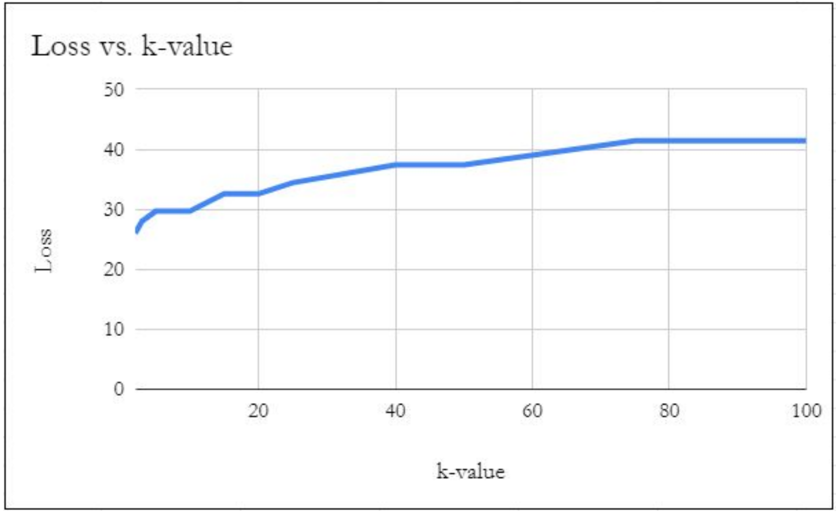
Experimental Design

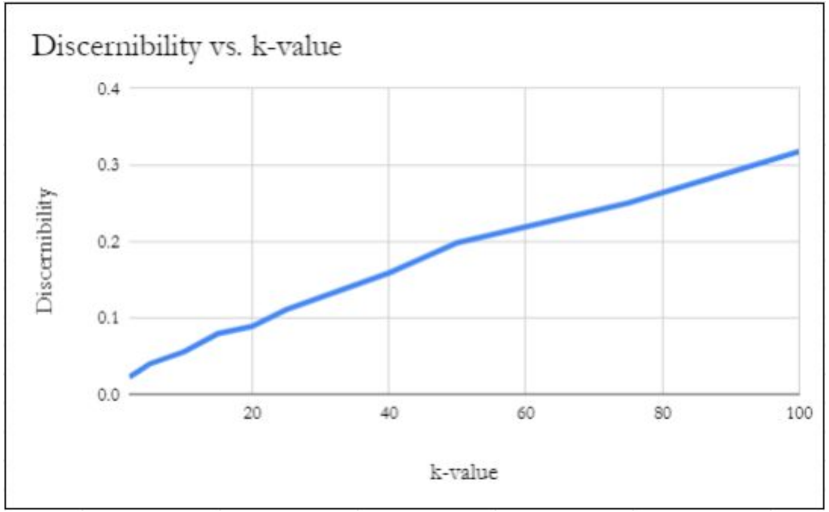
- Experimenting with the addition of randomized fake data and its effects on utility and anonymity of dataset
- Used python randomization library along with noise distribution to simulate
- Using discernibility and loss matrix to measure utility of a dataset

$$C_{DM}(g,\ k) = \sum_{\forall E s.t.\ |E| \geq k} |E|^2 + \sum_{\forall E s.t.\ |E| \geq k} |D||E|$$

Results

- As k-value increases, Loss and discernibility both increase (obviously)
- Loss seems to level out
- Larger induced equivalence classes ➜ lower utility





Conclusions

- It’s hard to tell how fake data affects utility of the data because this is largely dependent on the application
- Anonymization is not affected by fake data
- This presents a future research direction

Related Topics

- Being able to analyze the content of data beyond the hierarchies that are created to support the algorithm
- Using anonymization or de-anonymization in graph based social networks
- Applying data analytics in writing

Thanks