Project Proposal

Boaz Kaufman
Tanishk Suvarna
Kosta Ligris
Karl-Johan Westhoff

Introduction

In the era of big data, organizations and researchers often share or publish datasets for various purposes—ranging from academic research to public transparency. However, even when such datasets appear to be "anonymized," they can often contain hidden privacy risks. This is especially true when data attributes can be correlated with external sources (e.g., social media) that reveal identifying information.

This project will explore a publicly-leaked dataset, whose publisher has claimed there are "no privacy issues" associated with its release. Using well-established privacy metrics and techniques—such as k-anonymity, l-diversity, and t-closeness—we aim to demonstrate that an adversary could, in fact, re-identify individuals in the dataset when combined with other publicly available information. Our work will illustrate the potential dangers of insufficient anonymization practices and highlight how privacy engineering concepts can be systematically applied to measure and mitigate these risks.

Survey of related work.

- https://www.worldscientific.com/doi/abs/10.1142/S0218488502001648 (Class reading from week 5)
- Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1, 1 (March 2007), 3–es. https://doi.org/10.1145/1217299.1217302
- TBA

Feasibility

We regard the basic idea of analysing a dataset using k-anonymity, l-diversity, and t-closeness as feasible. The main challenge is to find suitable datasets that are both ethical/legal to use and suitable in terms of attributes and format. A fallback plan could be to synthesize a dataset or disregard the premise of data being from a breach.

Contributions

¼ Workload per group member, workload to be determined in retrospect ;-)

Milestones

Proposed (internal) milestones until the end of the semester.

1) Acquire Dataset
    a) Dataset pre processing, washing, formatting
2) Revisit Problem Definition, update if needed based on available data
3) Baseline Processing
    a) Calculate k-anonymity
    b) Evaluate l-diversity
    c) Assess t-closeness
4) Cross link to additional dataset
    a) Data processing - joining the datasets
5) Re-evaluate
    a) k-anonymity
    b) l-diversity
    c) T-closeness
6) Compare before and after results
7) Propose mitigation
8) Reporting