Achieving K-Anonymity for Sensitive Data

Team member: Mawell Adu, Hyomin Kim, Weigen Chen, Surafel Tsadik

Introduction

This project aims to achieve k-anonymity for structured datasets, such as healthcare records, to reduce the risk of re-identifying individuals within the data. K-anonymity is a privacy-preserving technique that ensures each individual in a dataset is indistinguishable from at least k−1 other individuals based on their quasi-identifiers (e.g., age, gender, ZIP code). By employing techniques such as generalization (substituting specific values with broader categories) and suppression (removing certain data points), the project seeks to balance the trade-offs between privacy and data utility. Tools like ARX and other similar anonymization frameworks will be utilized to systematically implement and evaluate these techniques.

Overall Goal

The project's main objective is to reduce the risks of re-identification in sensitive datasets while preserving the data's usefulness for analysis and querying. This entails:

- Implementing k-anonymity using techniques of generalization and suppression.
- Assessing the trade-offs between privacy protection and data utility.
- Ensuring that the anonymized dataset continues to be functional for database queries and analytical purposes.

What are we doing?

When working with datasets, it's important to identify quasi-identifiers and sensitive attributes that need protection. Quasi-identifiers can lead to re-identification when combined with external data, while sensitive attributes contain private information. Effective data anonymization strategies, such as using tools like ARX for generalization and suppression, achieve k-anonymity, ensuring individuals can't be distinguished from at least k-1 others.

Evaluating the effectiveness of these methods by measuring k-anonymity levels and query accuracy is essential for maintaining data utility while safeguarding privacy. A trade-off analysis helps balance privacy and utility, leading to optimal configurations for

various use cases. Finally, documenting the entire process is crucial for future data privacy initiatives and maintaining transparency in managing sensitive data.

Why Is This Important?

Protecting privacy is of paramount importance when dealing with sensitive datasets, such as those related to healthcare records. These datasets often contain personal information that, if re-identified, could result in privacy breaches, discrimination, or misuse. Techniques like k-anonymity are employed to safeguard individuals' identities by ensuring that their data cannot be easily distinguished from that of others in the dataset. Simultaneously, it is crucial to maintain the utility of the data, allowing it to remain beneficial for research, policy-making, and analysis. Striking the right balance between privacy and utility ensures that sensitive information is protected without compromising the dataset's value.

Moreover, achieving k-anonymity assists organizations in complying with stringent data protection regulations, such as GDPR and HIPAA, which require the safeguarding of personal information. Compliance not only mitigates legal risks but also cultivates trust among individuals and organizations that share and utilize sensitive data. By implementing robust privacy measures, organizations demonstrate their dedication to protecting personal information, thereby strengthening trust and promoting responsible data sharing and usage.

Background

Tools:

The ARX data anonymization tool, developed by Fabian Prasser and his team at TUM (Technical University of Munich, Germany), has been selected for this project due to its scalability, flexibility, and extensive support for privacy models beyond just k-anonymity. While many anonymization tools focus on a single privacy-preserving technique, ARX provides a comprehensive framework that includes ℓ-diversity, t-closeness, differential privacy, and risk-based anonymization approaches.

This tool is particularly well-suited for handling structured datasets, allowing fine-tuned anonymization strategies while preserving data utility for research and analysis. Additionally, ARX's built-in data transformation, risk assessment, and visualization features make it an efficient choice for balancing privacy protection and usability.

Dataset:
The dataset we will be using is a COVID-19 Public Therapeutic Locator dataset, which shows the location of publicly available COVID-19 Therapeutics. The columns that exist in this dataset include the provider's name, address, city, country, state, zip code, the therapeutic's name, and NDC (National Drug Code), plus other columns. The columns we will mark as sensitive are the address, city, state, country, and zip code.

For this study, the address, city, state, country, and zip code are designated as sensitive attributes, as they could potentially lead to the re-identification of providers or patients. Using ARX, these quasi-identifiers will be generalized or suppressed to achieve an appropriate level of anonymity, ensuring compliance with privacy best practices while maintaining the dataset's usability for public health and research purposes.

Privacy Risks

In this part, we will describe what privacy risk we will address in this project and how these risks map to the Solove and Citron's taxonomy.

Privacy Risks We Want to Addressed in This Project

The k-Anonymity model is designed to prevent re-identification of individuals in datasets while still allowing valuable data to be shared for research and analysis. The key privacy risks addressed in this project include:

1.Re-identification through Quasi-Identifiers: Even after removing explicit identifiers, an attacker can still use quasi-identifiers to uniquely identify an individual. K-Anonymity ensures that each record is indistinguishable from at least k-1 others(Sweeney, 2002).

2. Linkage Attacks: Attackers can link anonymized data with external datasets (e.g., voter registration lists) to identify individuals. K-Anonymity mitigates this by ensuring that groups of at least k individuals share the same quasi-identifiers.Even anonymized datasets are vulnerable when cross-referenced with external databases (Narayanan & Shmatikov, 2008)

3. Data Misuse by Third Parties: If de-anonymization is successful, attackers can misuse personal data for financial gain or social manipulation (Rocher, Hendrickx, & de Montjoye, 2019). K-Anonymity reduces this risk but does not eliminate it completely.

Mapping These Risks to Solove and Citron's Taxonomy

1. Physical Harms and Psychological Harms: If re-identification occurs, individuals may face targeted physical or psychological threats, such as stalking or violence or nuisance calls. k-Anonymity mitigates this by preventing direct re-identification.

2. Economic Harms: Re-identified data can lead to financial consequences, such as increased insurance premiums, denial of loans, or employment discrimination. For example, if a person's medical history indicating a chronic illness is re-identified, an insurance company might charge higher premiums or deny coverage.

3. Reputational Harms: If sensitive information such as medical conditions, criminal records, or financial hardships is re-identified, it can lead to reputational damage. For instance, an individual may face public embarrassment, social stigma, or professional consequences if such information is leaked.

4. Autonomy Harms: When data is misused, individuals' autonomy and freedoms may be compromised. For example, tracking someone's political or voting behavior could lead to manipulation, coercion, or restrictions on freedom of expression.

5. Discrimination Harms: Even when individual identification is avoided, data representing a group with shared attributes can lead to group-based discrimination. For instance, if a dataset reveals that a particular ZIP code has a high incidence of a medical condition, residents of that area might face higher insurance rates or other forms of bias.

6. Relationship Harms: Leaked sensitive information can damage personal relationships(Zendata, 2022). For example, revealing an individual's undisclosed medical condition or personal habits to family, friends, or colleagues could lead to broken trust, social isolation, or conflict.

Requirements

Use Case 1: Healthcare Data Anonymization

In healthcare data anonymization, patient records must be anonymized using a use-case-specific approach to ensure compliance with privacy regulations while preserving data utility. The anonymization process should apply k-anonymity, where quasi-

identifiers are either removed or generalized to prevent re-identification. The statistical properties of the dataset must be maintained to ensure usability for medical research. Researchers should be able to access anonymized data without exposing patient identities, and a validation process must be in place to measure privacy effectiveness while minimizing data distortion.

Use Case 2: Biometric Data Privacy (Face Image Anonymization)

For biometric data privacy, the k-Same algorithm must be implemented to de-identify face images while maintaining their usability for analysis. The de-identified images must be indistinguishable from at least k-1 others to ensure anonymity while preserving essential facial characteristics for research purposes. Accuracy tests should be conducted to assess the effectiveness of the system in preventing re-identification. Additionally, compliance with biometric data privacy regulations and industry best practices must be ensured.

Use Case 3: Educational Data Privacy under FERPA

In educational data privacy under FERPA, student records must be anonymized using k-anonymity techniques to allow de-identified data sharing while ensuring compliance with privacy regulations. Each student's record must be grouped with at least k-1 others to prevent re-identification. Direct identifiers, such as student ID and full names, should be removed or replaced with generalized categories. Role-based access controls must be implemented to restrict access to sensitive student data. The anonymized dataset should remain useful for educational research and policymaking while adhering to FERPA guidelines.

Milestones & Timeline

- Feb 6th: Finish the Project Presentation
- Feb 11th: Very Simple Prototype for Project Done
- Feb 12th-17th: Work on Slides
- Feb 18th-20th: Prepare for Midterm
- Feb 25th-27th: Mid-semester Project Presentation
- April 1st: Mini Project Update (Have majority of work done)
- April 15th: Finish Poster Draft
- April 17th: Quick Group Check-Up
- April 22nd: Final Project Poster
- May 3rd: Final Project Report

Reference

Arvind Narayanan, Vitaly Shmatikov. (n.d.). *Robust de-anonymization of large sparse datasets*. 2008 IEEE Symposium on Security and Privacy (sp 2008). https://ieeexplore.ieee.org/document/4531148

The costs of Anonymization: Case study using clinical data. (2024, April 24). Journal of Medical Internet Research. https://www.jmir.org/2024/1/e49445

COVID-19 public therapeutic locator. (2024, January 31). HealthData.gov. https://healthdata.gov/Health/COVID-19-Public-Therapeutic-Locator/rxn6-qnx8/about_data

Data anonymization tools: The 4 best and the 7 worst choices for privacy. (2024, February 26). MOSTLY AI. https://mostly.ai/blog/data-anonymization-tools

E.M. Newton; L. Sweeney; B. Malin. (2005, February). *Preserving privacy by de-identifying face images*. IEEE Xplore. https://ieeexplore.ieee.org/document/1377174

Estimating the success of re-identifications in incomplete datasets using generative models. (2019, July 23). Nature. https://www.nature.com/articles/s41467-019-10933-3

How easy is it to re-identify data and what are the implications? (2025, February 5). AI Governance & Data Privacy Platform | Zendata. https://www.zendata.dev/post/how-easy-is-it-to-re-identify-data-and-what-are-the-implications

K-anonymity: A model for protecting privacy: International Journal of uncertainty, fuzziness and knowledge-based systems: Vol 10, no 5. (2002, October 1). International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. https://dl.acm.org/doi/10.1142/s0218488502001648

Protecting privacy using K-anonymity. (n.d.). PMC Home. https://pmc.ncbi.nlm.nih.gov/articles/PMC2528029

Re-identification of "Anonymized" data. (2018, September 14). Georgetown Law Technology Review. https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017

U.S. Department of Education. (2015, July). *CASE STUDY #5: Minimizing Access to PII: Best Practices for Access Controls and Disclosure Avoidance Techniques*. https://studentprivacy.ed.gov. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://studentprivacy.ed.gov/sites/default/files/resource_document/file/Case_Study_5_Minimizing_PII_Access_0.pdf