K-Anonymity in Privacy Engineering

Weigen Chen, Hyomin Kim, Max Adu, Surafel Tsadik

Supervisor: Prof. Dr. Hana Habib

abstruct

This project focuses on achieving k-anonymity in structured datasets, such as census or healthcare records, to minimize re-identification risks. Using tools ARX, the goal is to apply techniques like generalization and suppression while systematically evaluating the trade-offs between privacy and data utility in database query contexts.

Background

k-anonymity protects privacy by ensuring that each individual in a dataset cannot be distinguished from at least ***k-1 others*** based on quasi-identifiers like ***ZIP code, gender, and birthdate***. While these values may seem harmless alone, they can lead to ***re-identification riks*** when combined.

A well-known real-world case proves this risk. The Massachusetts GIC released anonymized health records, removing names but keeping quasi-identifiers. ***Dr. Latanya Sweeney(who invents the K-anonymity)*** bought a voter registration list for $20 and used it to re-identify Governor William Weld's medical record by matching ***just*** his ZIP, gender, and birthdate.

Her earlier research also showed that 87% of the U.S. population could be uniquely identified by the same three fields. This demonstrates that anonymity cannot rely on removing names alone.

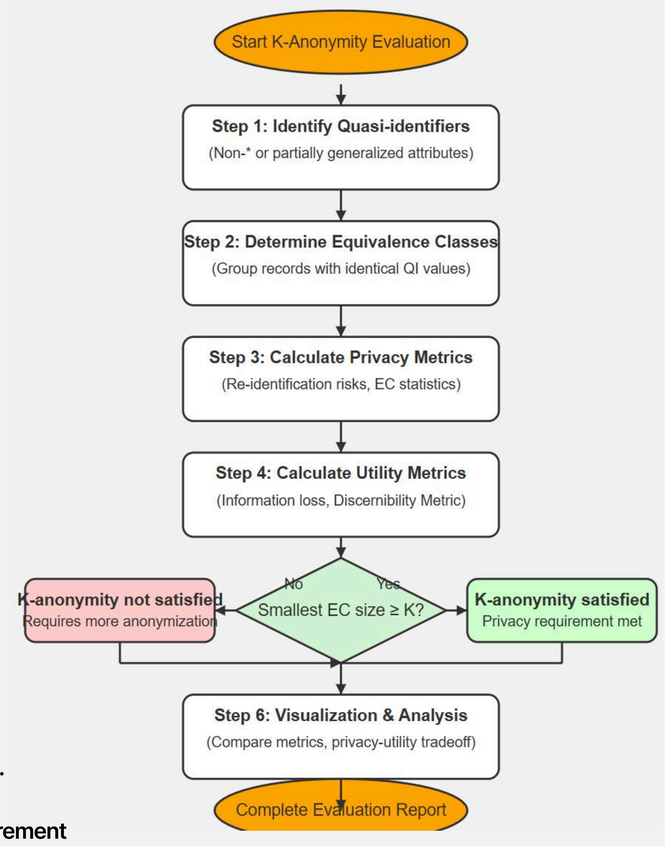
This project explores how to apply k-anonymity and related techniques like ***suppression and generalization*** to minimize re-identification risks while preserving data utility particularly in healthcare and other sensitive domains.

We use the ***ARX*** data anonymization tool, developed by Fabian Prasser and his team at TUM (Technical University of Munich, Germany. The dataset is COVID-19 Public Therapeutic Locator dataset, which shows the location of publicly available COVID-19 Therapeutics

Evaluation

The goal of our evaluation is two-fold:

1. **Minimize re-identification risk by applying techniques like generalization, suppression, and perturbation.**
2. **Preserve data utility to ensure the anonymized dataset remains usable for analysis and research rather than blindly masking all information.**



So, we apply the following measurement

1. Manually calculate DM and NCP. Discernibility Metric (DM), which penalizes datasets for each indistinguishable group (equivalence class) – larger groups or heavy suppression increase the penalty. Another metric is Normalized Certainty Penalty (NCP), which sums up how much each value was generalized relative to its domain. Regardless of the metric, the general trend is that higher k leads to higher distortion

2. Use ARX tools. ARX provide re-identification risk report

Method

Generalization: Converts specific values into broader categories.

- Zip Code: 34207-4105 → 34207 → 342
- Latitude: 27.434454 → 27.43 (rounded to 2 decimal places)
- Date: 2024-12-06 → 2024

  (keep only year for temporal generalization)

Suppression: Removes certain high-risk data points.

- Remove "Provider Note" if only 1% of rows have it
- Suppress exact Address field (e.g., "6003 14TH ST W") when location is rare
- Drop small chains or independent pharmacies that appear only once in dataset

Perturbation (e.g., Differential Privacy)

Adds noise to protect individual contributions in aggregate data.

- Latitude: 40.712776 → 40.71

  (rounding or injecting random noise to obscure exact location)
- COVID test count: 87 → 91

  (add Laplace noise to prevent precise patient tracking)
- Income: $1,250 → $1,170

  (perturbed value retains statistical value but hides individual billing)

Conclusion and Future Development

k-anonymity is a foundational technique in data privacy that ensures each record in a dataset is indistinguishable from at least k-1 others based on a set of quasi-identifiers. It helps prevent re-identification by applying methods such as generalization and suppression. The approach is intuitive, easy to implement(via ARX), and widely applicable to structured data like healthcare records.

However, k-anonymity has notable limitations:

- It does not protect against homogeneity attacks (when all records in a group share the same sensitive value).
- It is vulnerable to background knowledge attacks.
- It often results in information loss as k increases, impacting data utility.

So future work can focus on:

- Incorporating stronger models like l-diversity and t-closeness, which enhance resistance to attribute inference.