Sanitizing Clinical Data Using Big Data Frameworks

Hongzhu Jiang[1]

[1]**ShanghaiTech University, Shanghai, China**

Introduction

In the era of big medical data, datasets often contain quasi-identifiers—combinations of non-sensitive attributes like age, gender, and ZIP code—that, when combined, can uniquely identify individuals. Even in anonymized datasets, these quasi-identifiers can lead to privacy breaches when adversaries link them with external data sources. Meanwhile, sensitive attributes such as income or disease status must be explicitly protected.

A widely adopted solution is k-anonymity[1], which ensures that each individual in the dataset is indistinguishable from at least $k-1$ others in terms of quasi-identifiers. However, this protection has limitations. For instance, if all individuals in a k-anonymous group share the same sensitive value, an attacker can still infer that sensitive value with certainty. This leads to attribute disclosure risks.

To address this, l-diversity improves upon k-anonymity by requiring that each equivalence class contains at least l different sensitive attribute values. While this mitigates exact inference, it does not prevent probabilistic inference when a sensitive value dominates within the group.

To further limit inference risks, t-closeness requires that the distribution of sensitive values in each group be statistically similar to the distribution in the entire dataset. This ensures that even if an adversary knows an individual belongs to a group, the information gain about the sensitive attribute remains low.

However, enforcing stronger privacy guarantees often reduces data utility and increases computational cost, especially on large-scale clinical datasets such as EHRs. Therefore, there is a pressing need for scalable, distributed privacy-preserving algorithms that can balance privacy and utility while supporting regulatory compliance (e.g., HIPAA, GDPR).

This project aims to implement and evaluate such techniques—k-anonymity, l-diversity, and t-closeness—on clinical data using Apache Spark, with a particular focus on scaling to real-world datasets and measuring re-identification risks using both traditional and game-theoretic methods.

Pilot Project

To assess the feasibility of scalable clinical data anonymization, I conducted a pilot study using a small synthetic dataset from the SyntheticMass (Synthea) platform. This dataset consists of 116 simulated patient records and mimics real-world EHR structure, making it ideal for initial testing before scaling up to larger datasets such as MIMIC-III.

Table 1. Raw data preview before anonymization.

ID	Birthdate	Gender	ZIP	Income
1	1993/11/11	M	02644	71625
2	1987/8/2	F	02472	542941
3	1989/3/5	F	02081	886745
4	1993/2/2	F	02169	35850
5	1999/10/27	F	00000	149942

The dataset included key quasi-identifiers: birthdate, gender, and zip, and I used income as the sensitive attribute. I first converted birthdate into age, and truncated ZIP codes to their first three digits (ZIP3) to reflect geographic generalization. I also cleaned up invalid ZIP codes (e.g., “00000”) by labeling them as “UNK” to

avoid distortion in groupings. To evaluate the privacy protection achieved through this method, I verified that every group in the resulting anonymized dataset contained no fewer than five records, thereby satisfying the formal definition of k-anonymity. I then calculated the l-diversity of each group by counting the number of distinct income values present. This analysis enabled me to evaluate the degree to which the sensitive attribute (income) exhibited sufficient variability within groups, thereby mitigating the risk of attribute inference.

In addition to quantitative assessments, I generated a scatter plot of age versus ZIP3, with data points color-coded by income, to visually inspect the distribution of sensitive attributes in the original dataset prior to anonymization. The visualization shown in Figure 1 revealed several clusters with limited income diversity, underscoring the necessity of applying stronger privacy-preserving techniques such as l-diversity and t-closeness.

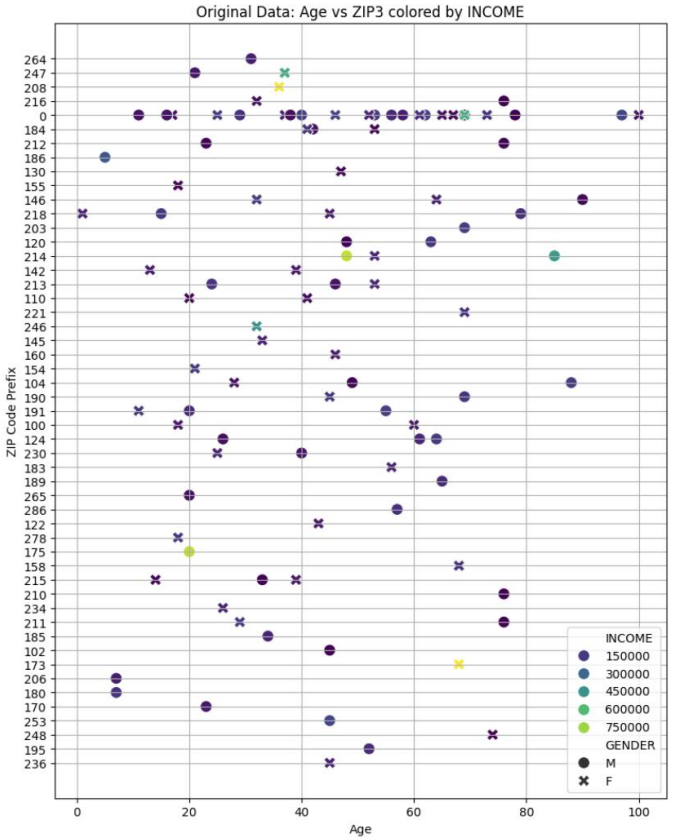

Figure 1. Distribution of sensitive attribute (income) by age and ZIP3 in the raw dataset.

Then I implemented a recursive partitioning algorithm inspired by the Mondrian multidimensional *k*-anonymity method[2] using Python. The algorithm works by iteratively splitting the dataset based on the quasi-identifiers—age, gender, and ZIP3—while ensuring that each resulting group contains at least k records. At each step, I computed the span of each attribute (range for numerical values, and cardinality for categorical ones) and selected the attribute with the widest span to guide the partition. The partitioning process stopped when no further splits could be made without violating the *k*-anonymity condition. I set *k*=5 for this initial test.

Summary of groups (k-anonymity and l-diversity):

	AGE	GENDER	ZIP3	total_count	l_diversity
0	13-17	F,M	0,142,215,218	6	6
1	18-25	F,M	0,154,155,191,212,213,247	7	7
2	18-25	F,M	100,110,170,175,230,265,278	8	7
3	1-11	F,M	0,180,186,191,206,218	6	6
4	26-34	F,M	104,124,185,211,215,234	6	6
5	29-36	F,M	0,145,146,208,216,246,264	7	7
6	37-45	F,M	0,110,142,184,190,247	9	9
7	39-45	F,M	102,122,215,218,230,236,253	7	7
8	46-53	F,M	0,120,130,213,214	9	8
9	46-55	F,M	104,160,184,191,195	5	5
10	56-61	F,M	0,100,124,183,286	8	7
11	62-67	F,M	0,120,124,146,189	8	8
12	68-73	F,M	0,158,190,203,221	8	7
13	68-76	F,M	173,210,211,212,216,248	7	3
14	78-90	M	0,104,146,214,218	7	5
15	97-100	F,M	0	8	2

Figure 2. Summary of total counts and l-diversity in each *k*-anonymized group.

Figure 2 shows that all partitions met the k-anonymity requirement, and most groups showed l-diversity ≥ 5, which indicates a relatively low risk of attribute disclosure. However, a few groups had skewed income distributions, highlighting the potential need to implement t-closeness in the extended phase. The entire process completed in seconds on a single machine, confirming that the approach is lightweight and suitable for scaling in distributed environments.

Extended Project

To extend the practical scope of this study, the project proposes a scalable framework for anonymization and privacy risk analysis using real-world datasets and distributed systems. The MIMIC-III dataset will serve as a benchmark to evaluate scalability and effectiveness. Spark-based implementations of k-anonymity, l-diversity, and t-closeness will be developed, enabling efficient distributed anonymization.

A key component of the extension is the quantification of re-identification risk using a formal game-theoretic framework, based on the model proposed in "A Game Theoretic Framework for Analyzing Re-Identification Risk". Unlike Bayesian game approaches, this framework formulates the interaction between an attacker and a data publisher as a strategic game with explicitly defined payoffs, strategies, and equilibrium analysis. This model will be implemented in Spark to enable large-scale simulations of adversarial behavior and to evaluate how anonymization strategies influence the risk of re-identification.

The system's performance will also be evaluated in terms of execution time, memory usage, and fault tolerance under varying partitioning strategies and privacy levels. Key risks include the computational intensity of game-theoretic models, uncertainty in parameter tuning, and low diversity in real data distributions. To mitigate these, we will leverage Spark's in-memory optimization, conduct sensitivity analyses across strategy spaces, and employ diversity-aware partitioning with utility-preserving suppression mechanisms.

Timeline

Date Range	Milestone Description
2025/5/17-2025/5/23	Data Preparation; Development of Distributed Anonymization Modules
2025/5/24-2025/5/30	Game-Theoretic Model Development; Risk Quantification Analysis
2025/5/31-2025/6/6	System Optimization; Result Analysis; Report Writing

References

1. Sweeney L. k-anonymity: A model for protecting privacy[J]. International journal of uncertainty, fuzziness and knowledge-based systems, 2002, 10(05): 557-570.
2. LeFevre K, DeWitt D J, Ramakrishnan R. Mondrian multidimensional k-anonymity[C]//22nd International conference on data engineering (ICDE'06). IEEE, 2006: 25-25.