Scalable Local-Recoding Anonymization using Locality Sensitive Hashing for Big Data Privacy Preservation

Khotso Bore
University Of Pretoria
u19180642@tuks.co.za

Innocentia Ledimo
University Of Pretoria
u21556564@tuks.co.za

Busisiwe Vemba
University Of Pretoria
u22928678@tuks.co.za

Siphesihle Khumalo
University Of Pretoria
u25759257@tuks.co.za

ABSTRACT

This report details the implementation and evaluation of a scalable local recoding anonymization approach using locality sensitive hashing (LSH), primarily based on the research paper "Scalable Local-Recoding Anonymization using Locality Sensitive Hashing for Big Data Privacy Preservation" [3]. This approach proposed to enhance the scalability and efficiency of k-anonymity in big data environments. Using the UCI Adult dataset [1] the method integrates a new provenance-set-based semantic distance metric with LSH(specifically MinHash) for efficient data partition followed by a parallelized, recursive agglomerative k-member clustering algorithm(Beta-AC) [3].

1. EXPLORATORY DATA ANALYSIS (EDA)

1.1 Data Inspection

The dataset used is a preprocessed version of the Adult dataset from the UCI Machine Learning Repository [1]. It contains over 30000 records across 10 relevant attributes including 8 quasi-identifiers and 2 numerical features implicitly discretised by the methodology.

- Initial Data Count: The raw dataset had 32561 entries.
- column dropping: The analysis began with loading the Adult dataset consisting of 15 columns initially. Several columns were considered non-quasi-identifying or as non-sensitive like capital-gain, capital-loss, fnlwgt, education-num, and income; in order to focus on the key quasi-identifier(QI) attributes.
- Final Quasi identifiers set: The resulting dataset contained 10 columns consisting of 8 categorical attributes(workclass, education, marital-status, occupation, relationship, race, sex, native-country), and two numerical attributes(age, hours-per-week) as tabulated in table 1 that would be discretized by the methodology. Work Class is used as the sensitive attribute.
- Missing Data: The initial inspection revealed missing data represented by a '?' string. These records were removed, reducing the total to 30162 clean records.

The table features the quasi-identifier (QI) attributes used in the analysis, inclusive of their type and the calculated percentage of missing values which are marked as '?' of which were removed during preprocessing. The total rows removed due to missing values was 2399 out of 32561, resulting in a 7.37% for features containing these entries.

Table 1: Dataset Overview

Feature	Type	Missing (%)
Age	Numerical	0.0
Workclass	Categorical	7.37
education	Categorical	0.0
marital-status	Categorical	0.0
occupation	Categorical	7.37
relationship	Categorical	0.0
race	Categorical	0.0
sex	Categorical	0.0
hours-per-week	Numerical	0.0
native-country	Categorical	7.37

1.2 Visualisations

Here are some key visualizations from the EDA phase. We explore the distributions of the numerical features age and hours worked, and their correlations.

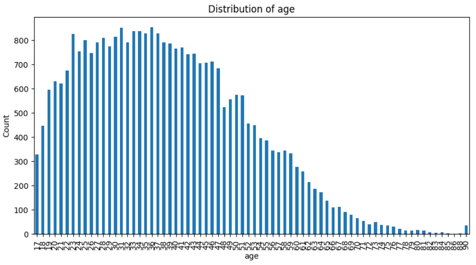


Figure 1: Age Distribution

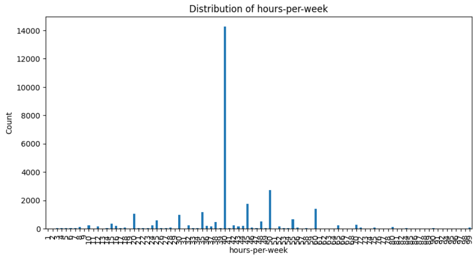


Figure 2: Hours Per Week Distribution

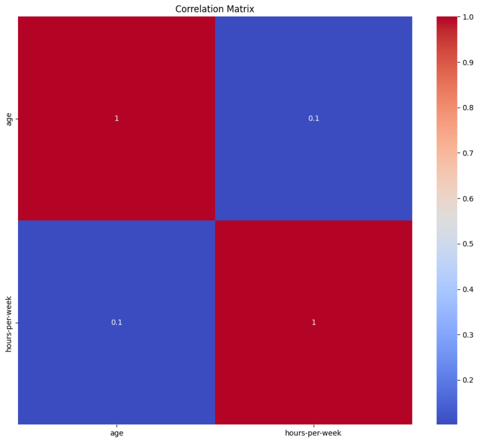


Figure 3: Correlation Matrix of Numerical Features

1.3 Insights

The correlation between age and hours worked suggests that while there may be slight tendencies such as younger or older employees working somewhat more or fewer hours, the relationship is not strong or consistent across the group.

A large portion of employees work standard full-time hours (around 40 hours per week), with fewer employees working significantly more or less than this amount. This is highly likely due to most records being from the United States, with a smaller representation from other countries.

2. DATA PREPROCESSING

2.1 Handling Missing Data

The dataset contained both standard and non-standard forms of missing values, which were handled in two stages:

2.1.1 Initial Removal of Missing Entries:

After loading the data, all records containing standard missing values were removed. This created a clean baseline for subsequent processing.

2.1.2 Handling Non-Standard Missing Indicators:

Some attributes, such as `workclass`, used the symbol "?" to represent missing values instead of a recognized missing data marker, such as `NaN`. These entries were first identified and converted into proper missing values, after which they were removed from the dataset.

This two-step process ensured that all incomplete records were removed. Maintaining a complete dataset is crucial for the LSH-based anonymization pipeline, since missing values can distort similarity measurements and clustering outcomes.

2.2 Feature Engineering

Feature engineering focused on selecting only the attributes relevant for anonymization and preparing them for efficient processing.

2.2.1 Feature Selection:

Several columns were removed from the dataset because they were either:

- Not quasi-identifiers that could link to external data sources,
- Redundant representations of existing information, or
- Sensitive or irrelevant for the anonymization process.

This reduced the dataset from fifteen to ten columns, retaining only the quasi-identifiers such as *age*, *workclass*, *education*, *marital status*, *occupation*, *relationship*, *race*, *sex*, *native country*, and *hours per week*. These attributes were used as the basis for the local-recoding anonymization.

2.2.2 Data Processing Throughout the LSH-Based Anonymization Pipeline:

Additional transformations were applied throughout the pipeline to ensure that data representation was suitable for the various algorithms at every stage of the pipeline:

- Binary vector conversion was used to transform categorical values into formats suitable for MinHash operations.
- Distance calculation functions were defined to measure semantic similarity between records.
- String standardization was applied to ensure consistent formatting across all categorical values.

These preprocessing components established the foundation for the LSH-based anonymization pipeline, which relies on accurate and consistent data representation. Further details on these transformations are discussed in Section 3.

2.3 Standardisation and Normalisation

Unlike numerical datasets that require techniques such as z-score standardisation or min–max scaling [2], this dataset consists mainly of categorical variables. Therefore, normalisation was incorporated directly into the anonymisation pipeline through specialised distance-based approaches:

- **Taxonomy-based distance metrics** that automatically bound all categorical distances within [0, 1] based on tree height and path length.
- **Provenance set-based semantic distances** that ensure comparability across different attributes and taxonomy structures.
- **String trimming** that removes whitespace to ensure consistent categorical value matching.

These methods maintain the original categorical form of the data while ensuring that all similarity and distance calculations are standardised. The details of these normalisation techniques are presented in Section 3.

3. DATA MINING METHODS AND ANALYSIS

3.1 Methods

The anonymization process treats privacy preservation as a k-member clustering problem under k-anonymity. The proposed approach integrates these components:

- **Provenance-Set Semantic Distance:** Defines similarity between records using Jaccard distance of their provenance sets, enabling LSH to approximate semantic closeness.

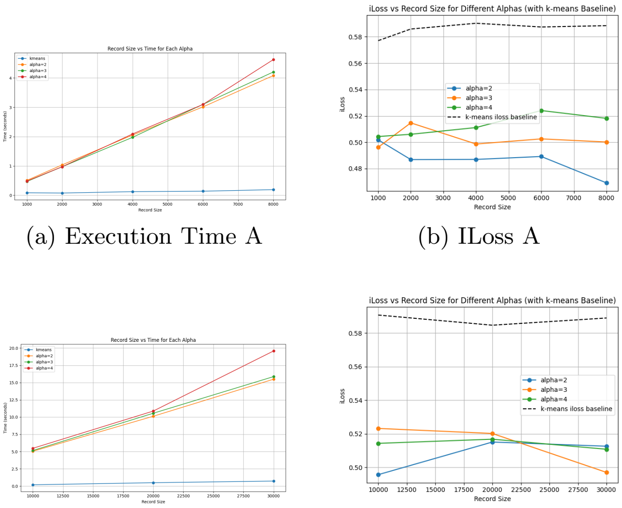
- **MinHash-Based LSH Partitioning:** Transforms each record's provenance vector into MinHash signatures. Similar records are hashed into the same β-clusters (coarse groups).

- **Recursive k-Member Clustering (LSH-RC):** Within each β-cluster, a recursive partitioning and agglomerative clustering algorithm is applied. Combines smaller clusters until each group size $\geq$ k.

- **Local recording:** Within each cluster, categorical attributes are generalized according to attribute taxonomies, ensuring that each cluster meets the k-anonymity requirement while minimizing information loss. The implementation constructs inverted taxonomies to compute necessary generalizations efficiently.

3.2 Results

The performance of the proposed LSH-RC method was evaluated against a baseline k-means clustering approach across varying dataset sizes. The key metrics assessed were execution time and information loss (ILoss) as functions of the number of records. Overall our results indicate that LSH-RC scales much less efficiently than k-means while maintaining comparable ILoss levels.



(a) Execution Time A

(b) ILoss A

(c) Execution Time B

(d) ILoss B

Figure 4: Execution time (a) & (c), and ILoss (b) & (d) w.r.t. #(records).

The result growth was noticeably super-linear in execution time. In comparison, **k-means** scaled more smoothly evident with an increased runtime that is almost linearly with dataset size, demonstrating significantly better computational efficiency.

3.2.1 Information Loss

Despite the higher computational cost, the LSH-RC method achieved similar and slightly lower iLoss than the k-means baseline for the evaluated dataset sizes. The provenance-set semantic similarity combined with local recoding allowed LSH-RC to group semantically coherent records into clusters that require less aggressive generalization, thereby reducing ILoss.

The following empirical patterns were observed:

- LSH-RC and k-means ILoss curves followed the same general trend.

- At larger dataset sizes, LSH-RC achieved comparable ILoss with smaller variance, indicative of a more stable anonymization behavior.

- The higher number of intermediate micro-clusters in LSH-RC contributed positively to maintaining semantic locality within groups.

Overall, LSH-RC method delivers competent anonymization quality, although at a cost of much poorer runtime scalability.

4. DISCUSSION

The experiment revealed several important insights regarding the behavior as well the suitability of LSH-RC for large-scale anonymization.

4.1 Patterns and Group Behavior

- The LSH produced coarse partitions that often significantly differed in size, with numerous very small clusters. This behavior is expected from MinHash-LSH because it prioritizes similarity over size balance.

- The recursive k-member agglomeration merged these clusters into valid groups, but the heavy imbalance increased the number of merge operations which then contributed to longer runtimes.

- The final anonymized clusters exhibited properties of being semantically coherent, especially for attributes with strong categorical structure.

4.2 Appropriateness of Methods

The LSH-based approach is theoretically suitable for high-dimensional categorical data since MinHash approximate Jaccard similarity efficiently, provenance-set-based distance well aligns with taxonomy-based generalization, and that local recoding reduces unnecessary global generalization. However, in practice, the method showed limited computational efficiency and scalability on the adult dataset even after preprocessing. K-Means, although less semantically grounded, offered substantially better execution times.

4.3 Limitations

Limitations and anomalies that have emerged include:

- Increasing leftovers records: Some iterations produced more leftover (<k) clusters than expected, suggesting that provenance sets may be sparse for certain records.
- Cluster fragmentation: For larger subsets, LSH produced many singletons, which then required multiple merge steps, inflating runtime.
- Runtime instability: Execution time occasionally jumped sharply between nearby dataset sizes, likely due to hash collisions and varying β-cluster fragmentation.

These behaviours indicate that while semantically strong, the current implementation is not suitable for and optimized for large-scale performance.

5. CONCLUSION AND REFLECTION

This project implemented and evaluated a complete LSH-based local recoding anonymization pipeline using provenance-based semantic similarity, MinHash-LSH partitioning, and recursive k-member clustering. Several conclusions and reflections emerged from the study.

5.1 Key Insights

The LSH-RC technique successfully produced k-anonymized clusters with low information loss, verifying its effectiveness at preserving data utility. MinHash-based partitioning enabled semantically meaningful grouping even in high-dimensional categorical spaces. Local recoding based on attribute taxonomies yielded less aggressive generalization than global methods.

5.2 Challenges Encountered

- Scalability was the primary challenge: The recursive merging step became computationally expensive due to many small β-clusters and large cluster-size variance.
- Parameter tuning for α (number of hash bands) required iterative experimentation, as small changes had disproportionately large effects.
- Implementing provenance-set representations and distance metrics required careful handling of categorical hierarchies and string normalization.
- Debugging recursive clustering and generalization steps demanded detailed logging and validation.

Despite these difficulties, the pipeline ultimately produced valid anonymization results comparable in quality to the k-means baseline.

6. FUTURE WORK

Several improvements could enhance this method significantly:

- Parallelizing LSH and recursive clustering to reduce execution time.
- Dynamic tuning of α based on dataset characteristics rather than fixed values.
- Hybrid anonymization approaches: LSH for coarse partitioning and k-means or hierarchical clustering within β-clusters.
- Improved taxonomies to support richer semantic generalization.
- GPU-accelerated MinHash for fast signature computation at scale.

Through the expansion of these techniques, LSH-RC could be made viable for true big-data environments.

7. ADDITIONAL AUTHORS

8. REFERENCES

[1] R. K. Barry Becker. Adult, 1996.

[2] S. Kappal et al. Data normalization using median median absolute deviation mmad based z-score for robust predictions vs. min–max normalization. *London Journal of Research in Science: Natural and Formal*, 19(4):39–44, 2019.

[3] X. Zhang, C. Leckie, W. Dou, J. Chen, R. Kotagiri, and Z. Salcic. Scalable Local-Recoding Anonymization using Locality Sensitive Hashing for Big Data Privacy Preservation. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 1793–1802, New York, NY, USA, Oct. 2016. Association for Computing Machinery.