Replication of the Paper: Scalable Local-Recoding Anonymization using Locality Sensitive Hashing for Big Data Privacy Preservation

Lucas Blanc
University of Pretoria
P.O. Box 14622
Student Number: u21436135
u21436135@tuks.co.za

Matt van Coller
University of Pretoria
P.O. Box 14622
Student Number: u22491199
u22491199@tuks.co.za

Matthew Pretorius
University of Pretoria
P.O. Box 14622
Student Number: u21775169
u21775169@tuks.co.za

Denver Saurombe
University of Pretoria
P.O. Box 14622
Student Number: u22653539
u22653539@tuks.co.za

ABSTRACT

The main objective of this paper is to provide an improvement on scalable local recoding anonymization solutions for Big Data Privacy. Current methods that are used for local recoding anonymization have an average time complexity of $O(n^2)$, which makes them difficult to apply on very large datasets. The reason behind this difficulty lies within the high memory usage of these methods, especially when it comes to using cloud computing systems, as many of the companies that utilize cloud computing systems may need to anonymize their sensitive data before third party access. With the increasing use of cloud computing systems, many organizations are facing restrictions from government agencies and regulatory bodies regarding how they can use the personal information of their customers. One of the reasons why organizations are having such difficulty complying with government agency and regulatory body requirements is that traditional anonymization methods do not support scalability with big data. As such, organizations are experiencing bottleneck issues related to both the data utility and the level of privacy protection provided by anonymization methods.

This paper presents a solution to the problem described above through the use of a semantic distance measure and locality-sensitive hashing (LSH) to provide scalability improvements of orders of magnitude while still providing k-anonymity.

1. INTRODUCTION

The explosion of digital data presents a vast opportunity for big data analysis, personalization, and decision making; however, as companies transition their data to cloud computing, they face an increasing concern regarding the protection of private user data. In order to comply with regulations (i.e., the GDPR and POPIA), organizations are required to remove identifiable data before sharing or analyzing it, thus creating a conflict between the organization's need to preserve the usefulness of the data and the need to protect users' identities.

One method used to provide identity protection is k-anonymity. k-anonymity provides that no single piece of an individual's data will be able to distinguish him/her from k-1 other individuals. Although traditional local recoding approaches are effective at achieving k-anonymity by generalizing data within a cluster (thus preventing identification), these approaches have limited applicability due to their lack of scalability. The primary limitations of local recoding approaches include computational complexity (typically $O(n^2)$) and high memory requirements

Due to the size of most big data sets, local recoding approaches are generally unfeasible in practice.

In addition to the aforementioned limitations, many organizations rely on third party providers for data processing and analysis, thus requiring the use of anonymization techniques prior to sending the data outside of their control. As data volumes grow and data dimensions increase, standard anonymization techniques have difficulty balancing privacy with usability, resulting in excessive processing overhead and diminished information value.

In response to the above limitations, this research seeks to replicate the work of Zhang et al. (2016), who developed a scalable local-recoding anonymization technique using Locality Sensitive Hashing (LSH) and a new semantic distance metric. Zhang et al.'s (2016) LSH approach allows for improved scalability, relative to previous approaches, by reducing the computational overhead associated with clustering, by grouping similar data records into the same partitions and then applying MapReduce or a similar distributed framework to process the data in parallel. Zhang et al.'s (2016) approach demonstrates improved scalability by several orders of magnitude, while still providing k-anonymity protection and minimal information loss.

Therefore, the purpose of this project is to replicate Zhang et al.'s (2016) scalable anonymization technique using a real world data set (the Yelp 2018 – 2019 data set), and to evalu-

ate the performance of the technique using metrics such as: execution time, information loss and k-anonymity protection. Additionally, this replication will aim to demonstrate that scalable anonymization techniques do not necessarily result in diminished data utility.

2. LITERATURE REVIEW

Research regarding privacy-preserving data publication has been underway for some time now, to find a middle ground between the data's utility and privacy protections. K-anonymity is probably one of the most popular privacy models; k-anonymity states that no single record in a dataset should be distinguishable from another k–1 record(s), through a set of quasi-identifier (QI) attributes, that could link to an individual through indirect means, i.e., gender, location, etc.

Traditionally, k-anonymity was achieved by employing either local or global recoding techniques. Global recoding employs generalizations that apply to all records in the dataset (i.e., replace age with a range). This typically results in significant data loss. On the other hand, local recoding group similar records together into clusters, and then generalize within these clusters. While preserving more data utility than global recoding, local recoding does present scalability issues - especially with very large and multi-dimensional datasets, where clustering these datasets is computationally expensive.

Several earlier algorithms were developed to increase efficiency over exhaustive search methodologies (Greedy Clustering (GC) algorithm, Mondrian Multidimensional k-Anonymity) by increasing the speed of the search - however, they still have $O(n^2)$ time complexities, and large memory requirements. Therefore, they do not lend themselves well to Big Data/Cloud Computing environments where datasets can consist of millions of records.

Researchers have addressed scalability concerns related to anonymizing large datasets by utilizing Distributed and Parallel frameworks, specifically MapReduce and Spark, to break down datasets into smaller parts and run these parts concurrently to reduce processing times. Unfortunately, naive partitioning may reduce the quality of anonymized output, if similar records are assigned to different processing nodes.

In order to improve both scalability and anonymity guarantees, Zhang et al. (2016) proposed an innovative solution to combine Locality-Sensitive Hashing (LSH) with Semantic Distance Measurement. Initially, they define a provenance-set-based semantic distance using the Jaccard distance on taxonomy trees to measure similarity among categorical attributes. Next, using MinHash-based LSH, similar records are efficiently clustered into β-clusters. Finally, a recursive k-member clustering step is applied to ensure that each β-cluster satisfies the k-anonymity requirement, while minimizing information loss (ILoss).

Since the work of Zhang et al. (2016), subsequent research and adaptations of LSH in privacy preservation have validated this concept, showing that approximate similarity search methods can scale anonymization to larger-scale, more heterogeneous data. Specifically, the LSH-based local recoding method provides an effective trade-off between computational scalability, data utility, and privacy strength - providing a practical basis for secure Big Data Analytics in Distributed Systems.

3. METHODOLOGY

The Methodology section will discuss how certain aspects of the project were completed.

3.1 Exploratory Data Analysis

3.1.1 Data Inspection

This research utilizes the Yelp Dataset (Yelp Data 2018–2019). It includes review data by users, business information and metadata about the reviews like ratings, categories and geographical location of businesses. As part of this project we created a sub-dataset including those attributes that are relevant to anonymizing the data and clustering - i.e., user_id, business_id, city, category and stars. We chose these attributes because they included identifiable or semi-identifiable information (also called quasi-identifiers) that could potentially be used to identify a particular user when linked together. The size of this dataset is substantial (millions of records) which makes it an excellent testbed for assessing the ability of our proposed LSH-based anonymization technique to scale well. Initial exploratory analysis of the data demonstrated that there were several common types of errors in the data that would need to be addressed prior to utilizing the data - specifically, missing values in either the city or category field(s); inconsistent capitalization in the textural data; and several businesses had missing or incomplete metadata. In addition to the diversity of categories and cities in the data, the presence of many unique values in each of the fields also creates high cardinality in the data. This level of diversity will provide us with an opportunity to assess how well the clustering algorithm performs in grouping records based on similarities without losing too much utility from the data. Overall, the Yelp dataset provides a realistic and heterogeneous environment that mirrors the complexity of performing privacy preserving operations in large scale, cloud based environments.

3.1.2 Visualisations

3.1.3 Insights

The Anonymization Model was heavily influenced by the results of Exploratory Data Analysis. The majority of the Reviews found in this data set come from a few locations where there is an abundance of activity (i.e. Las Vegas, Phoenix, and Toronto), thus introducing a geographic bias into the data. Furthermore, the categories (i.e. Business Categories) in this data set are also biased towards Restaurants, Nightlife, and Shopping, etc., thereby creating an uneven distribution among these categorical variables. This will affect the Clustering Behavior since categories that have a high number of instances (i.e. those with "dense" representations), will most likely create large clusters of similar items, while the categories that have less instances (i.e. those that are "sparse" or "niche"), will most likely create small, fragmented clusters.

An additional piece of information found during the exploration process, deals with the wide range of Star Ratings found within this data set, and how they may be related to Category and Location. In general, most of the ratings found in this data set fall somewhere between 3 and 5 Stars, creating a skewness in the distribution of ratings, and therefore affecting how distances are calculated when combining

numerical and categorical attributes. The High Dimensionality of the data, along with the Categorical Imbalance, and the Mixed Data Types, all indicate why traditional Global or Greedy Anonymization techniques are generally not scalable for handling such large datasets. As such, the use of Locality Sensitive Hashing (LSH) was used to quickly and efficiently approximate similarity among records, yet at a level of granularity that allows for preservation of Data Utility.

3.2 Data Preprocessing

3.2.1 Handling Missing Data

The pre-processing steps of the data set were completed before the application of the anonymisation model to clean the missing and inconsistent data from the dataset. The initial step of the cleaning involved reviewing the records in which the quasi-identifier attributes (such as City, Category etc.) were missing or blank. All completely missing entries in these quasi-identifier attributes would be deleted if they made up less than one percent of the total data to avoid skewing the distribution of the data. A generic "unknown" label was created to replace entries in the quasi-identifier attribute category when they were missing to preserve the quantity of each record, but not to introduce artificial bias into the data.

Additionally, all text based fields were standardized by using lower case letters and trimming whitespace to treat equivalent entries (i.e. restaurant vs restaurants) uniformly to eliminate artificial differences in categories to compute similarities between records. In addition, duplicate reviews and redundant records were removed to eliminate duplication in the clustering portion of the anonymization process. The final cleaned data set contained the vast majority of the original records and had been transformed into a format that could support both scalable anonymization and distance-based grouping.

3.2.2 Feature Engineering

The feature engineering process focused on representing the semi-structured nature of the dataset's categorical and numerical fields as a format that could accommodate semantic distance calculations and hashing. The quasi-identifier fields of every record (i.e., City, Category, Stars) were encoded to represent their respective taxonomy based on their hierarchy of generalized values. As an example, the City field was encoded to represent as City $\rightarrow$ State $\rightarrow$ Country. Categories such as "Restaurants" or "Cafes", were categorized as a broader category such as "Food and Beverage". Encoding the fields in this hierarchical structure allowed for the use of Jaccard Distance to calculate semantic similarities between all of the records and account for both exact and generalized matches.

In order to prepare for the Locality Sensitive Hashing (LSH) process, each provenance set was encoded as a binary vector using a MinHash signature. MinHash signatures are able to capture the similarity of all of the records by encoding them as a low dimensional, hashable representation, which allows for the efficient grouping of near-duplicates or related records. Derived features, including the average review rating per business and category frequency, were also created in order to assist the clustering process with reflecting both the content and the distribution of the data. Collectively, these engineered features increased the model's ability to generate coherent, privacy preserving clusters from the anonymized data, while maintaining the interpretability and utility of the data.

3.2.3 Standardisation / Normalisation

Due to both categorical and numerical attributes in the dataset, a combination of normalization strategies was employed to provide a balanced impact on each characteristic to be used in the clustering process. A normalization technique called Min-Max Scaling was utilized for the numerical attributes, such as stars to normalize their values into the interval [0,1]. Thus, the size of the rating would not overwhelm the calculation of distances due to the magnitudes of the ratings.

The normalization of the attributes was important because it allowed for the integration of both the continuous and categorical data within the semantic distance function. Uniform weighting was applied to the categorical data to prevent a bias in favor of attributes which have a larger number of categories or taxonomic breadth. In some instances, multiple levels of provenance exist (i.e., City $\rightarrow$ State $\rightarrow$ Country). Therefore, in these instances, the higher level generalizations were lightly de-weighted to preserve the detail in the relationships between the various attributes.

Ultimately, this normalization provided a harmonized feature space where both the semantic and numeric similarities of the features can be appropriately compared. Additionally, this normalization strategy helped to stabilize the clustering process and improve the sensitivity of the LSH hashing stage, as the attribute variance reflected true data diversity as opposed to an uncontrolled variability created through inconsistent scaling.

3.3 Data Mining Methods & Analysis

There were three primary elements in the anonymization pipeline: semantic distance measurement, LSH-based clustering, and recursive k-member grouping. First, a semantic distance measure was used to determine how similar two records are by computing the Jaccard distance between taxonomy-based provenance sets; this allowed for the identification of similar categories - such as Cafes and Restaurants - as related rather than separate entities.

Next, Locality Sensitive Hashing (LSH) was employed to cluster records with similar characteristics into the same partition via MinHash signatures. By employing MinHash, the number of pairwise comparisons is reduced from quadratic to nearly linear time.

Finally, the clusters within each partition were further refined through recursive k-member grouping, where small clusters were merged with their nearest neighbor until all clusters contained $k \geq 10$ members. The result of the process provided significant improvements to scalability, while limiting the amount of information lost and providing adequate levels of privacy protection.

4. DATASET

The data set utilized in this research project was the Yelp Open Dataset (2018 – 2019). This open data set consists of millions of user-generated reviews, descriptions of businesses and related metadata. The Yelp data set has been chosen because it has both numerous categorical variables and many numeric variables which will allow the researchers

to evaluate various anonymization approaches while balancing the trade-offs between preserving semantic meaning in the data and retaining some form of data utility. There are approximately 1.2 million businesses, 8 million reviews, and 1.9 million users contained within the data set; it spans across multiple U.S. cities and international locations.

5. RESULTS & CONCLUSION

6. REFERENCES

APPENDIX

A. INTRODUCTION

B. LITERATURE REVIEW

C. METHODOLOGY

C.1 Exploratory Data Analysis

C.1.1 Data Inspection

C.1.2 Visualisations

C.1.3 Insights

C.2 Data Preprocessing

C.2.1 Handling Missing Data

C.2.2 Feature Engineering

C.2.3 Standardisation / Normalisation

C.3 Data Mining Methods & Analysis

D. DATASET

E. RESULTS & CONCLUSION

F. REFERENCES