Journal of Science and Technology
ISSN: 2456-5660 Volume 8, Issue 12 (Dec -2023)
www.jst.org.in *DOI:https://doi.org/10.46243/jst.2023.v8.i12.pp219-229*

PRESERVING PRIVACY IN THE ERA OF BIG DATA: A MACHINE LEARNING-BASED ANONYMIZATION FRAMEWORK FOR SPATIOTEMPORAL TRAJECTORY DATASETS

C. Gazala Akhtar[1], G. Lahari[2], P. Shruthi[2], V. Akshitha[2]

[1]Assistant Professor,[2]UG Students, Department of Cyber Security Engineering.
[1,2]Malla Reddy Engineering College for Women, Maisammaguda, Dhulapally, Kompally,
Secunderabad-500100, Telangana, India.

To Cite this Article

C. Gazala Akhtar, G. Lahari, P. Shruthi, V. Akshitha , "PRESERVING PRIVACY IN THE ERA OF BIG DATA: A MACHINE LEARNING-BASED ANONYMIZATION FRAMEWORK SPATIOTEMPORAL TRAJECTORY DATASETS" *Journal of Science and Technology, Vol. 08, Issue 12 - Dec 2023, pp219-229*

Article Info

Received: 14-11-2023 Revised: 24 -11-2023 Accepted: 04-12-2023 Published: 14-12-2023

ABSTRACT

Publishing datasets plays an essential role in open data research and promoting transparency of government agencies. However, such data publication might reveal users' private information. One of the most sensitive sources of data is spatiotemporal trajectory datasets. Unfortunately, merely removing unique identifiers cannot preserve the privacy of users. Adversaries may know parts of the trajectories or be able to link the published dataset to other sources for the purpose of user identification. Therefore, it is crucial to apply privacy preserving techniques before the publication of spatiotemporal trajectory datasets. In this paper, we propose a robust framework for the anonymization of spatiotemporal trajectory datasets termed as machine learning based anonymization (MLA). By introducing a new formulation of the problem, we are able to apply machine learning algorithms for clustering the trajectories and propose to use k-means algorithm for this purpose. A variation of k-means algorithm is also proposed to preserve the privacy in overly sensitive datasets. Moreover, we improve the alignment process by considering multiple sequence alignment as part of the MLA. The framework and all the proposed algorithms are applied to T-Drive, Geolife, and Gowalla location datasets. The experimental results indicate a significantly higher utility of datasets by anonymization based on MLA framework.

Keywords: Preserving Privacy, Big Bata, Machine Learning Algorithms, Spatiotemporal, Trajectory Datasets.

1. INTRODUCTION

In the contemporary era of big data, the practice of publishing datasets plays a crucial role in advancing open data research and fostering transparency within government agencies. However, this seemingly beneficial practice raises concerns about the potential compromise of users' private information inherent in the published data. Among the most sensitive data sources are spatiotemporal trajectory datasets, which, when left unprotected, can expose individuals to privacy breaches. The conventional method of merely removing unique identifiers from such datasets proves insufficient to safeguard user privacy. Sophisticated adversaries could discern fragments of trajectories or establish connections between the published dataset and external sources, thereby facilitating user identification. Recognizing this vulnerability, it becomes imperative to employ privacy-preserving techniques prior to the dissemination of spatiotemporal trajectory datasets. In response to this challenge, this paper introduces a robust anonymization framework specifically tailored for spatiotemporal trajectory datasets, termed as the

Machine Learning-Based Anonymization (MLA) framework. The novelty of this approach lies in its innovative formulation of the problem, allowing for the application of machine learning algorithms to cluster trajectories effectively. The proposed framework leverages the widely-used k-means algorithm for trajectory clustering, presenting a refined variation to ensure privacy preservation in datasets with heightened sensitivity. Additionally, the alignment process is enhanced through the incorporation of multiple sequence alignment as an integral component of the MLA framework. To validate the efficacy of the proposed framework and algorithms, extensive experiments are conducted on prominent spatiotemporal trajectory datasets, namely T-Drive, Geolife, and Gowalla. The experimental results notably demonstrate a substantial enhancement in the utility of the datasets achieved through anonymization based on the MLA framework. This research not only contributes to the growing body of knowledge in privacy preservation but also offers a practical and effective solution to a critical concern in the era of big data and open data initiatives. The contemporary landscape of open data research and government transparency, marked by the widespread publication of datasets, introduces a critical problem concerning the compromise of users' privacy. This issue is particularly pronounced in spatiotemporal trajectory datasets, which, if not properly protected, pose a substantial risk to individual privacy. The conventional practice of merely removing unique identifiers from these datasets proves inadequate, as sophisticated adversaries can exploit partial trajectory information or link the published data to external sources for the explicit purpose of identifying users. Consequently, there exists a compelling need for pre-emptive privacy-preserving measures to be applied before the dissemination of spatiotemporal trajectory datasets. This problem statement underscores the urgency of developing effective techniques to safeguard user privacy in the context of open data initiatives, especially in datasets characterized by the intricate and sensitive nature of spatiotemporal trajectories. The motivation behind this research is rooted in the growing significance of open data research and government transparency juxtaposed against the increasing concerns regarding user privacy. With the widespread practice of publishing datasets, particularly spatiotemporal trajectory datasets, the inherent risk of privacy breaches has become a paramount issue. The conventional methods of data de-identification, such as removing unique identifiers, have proven insufficient in protecting user privacy, leaving room for adversaries to exploit trajectory information or cross-reference datasets for user identification. In response to these challenges, the research is propelled by a compelling motivation to pioneer robust privacy-preserving techniques specifically tailored for spatiotemporal trajectory datasets. By addressing the vulnerabilities associated with conventional anonymization methods, this research aims to contribute practical and effective solutions that enhance the privacy protection mechanisms in the era of open data. The overarching goal is to strike a balance between the imperative of transparent information sharing and the critical need to safeguard individual privacy, fostering a more secure and responsible landscape for open data initiatives.

2. LITERATURE SURVEY

Publication of data by different organizations and institutes is crucial for open research and transparency of government agencies. Just in Australia, since 2013, over 7000 additional datasets have been published on ’data.gov.au,’ a dedicated website for the publication of datasets by the Australian government. Moreover, the new Australian government data sharing legislation encourage government agencies to publish their data, and as early as 2019, many of them will have to do so [2]. Unfortunately, the process of data publication can be highly risky as it may disclose individuals’ sensitive information. Hence, an essential step before publishing datasets is to remove any uniquely identifiable information from them. However, such an operation is not sufficient for preserving the privacy of users. Adversaries can re-identify individuals in datasets based on common attributes called quasi-identifiers or may have prior knowledge about the trajectories traveled by the users. Such side information enables them to reveal sensitive information that can cause physical, financial, and reputational harms to people.

One of the most sensitive sources of data is location trajectories or spatiotemporal trajectories. Despite numerous use cases that the publication of spatiotemporal data can provide to users and researchers, it poses a significant threat to users' privacy. As an example, consider a person who has been using GPS navigation to travel from home to work every morning of weekdays. If an adversary has some prior knowledge about a user, such as the home address, it is possible to identify the user. Such an inference attack can compromise user privacy, such as revealing the user's health condition and how often the user visits his/her medical specialist. Therefore, it is crucial to anonymize spatiotemporal datasets before publishing them to the public. The privacy issue gets even more severe if the adversary links identified users to other databases, such as the database of medical records. That is the very reason why nowadays most companies are reluctant to publish any spatiotemporal trajectory datasets without applying an effective privacy preserving technique.

A widely accepted privacy metric for the publication of spatiotemporal datasets is k-anonymity. This metric can be summarized as ensuring that every trajectory in the published dataset is indistinguishable from at least k − 1 other trajectories. The authors in [3], adopted the notion of k-anonymity for spatiotemporal datasets and45d5vvv proposed an anonymization algorithm based on generalization. Xu et al. [4] investigated the effects of factors such as spatiotemporal resolution and the number of users released on the anonymization process. Dong et al. [5] focused on improving the existing clustering approaches. They proposed an anonymization scheme based on achieving kanonymity by grouping similar trajectories and removing the highly dissimilar ones. More recently, the authors in [6] developed an algorithm called k-merge to anonymize the trajectory datasets while preserving the privacy of users from probabilistic attacks. Local suppression and splitting.
Lack of a well-defined method to cluster trajectories as there is not an easy way to measure the cost of clustering when considering the distances among trajectories rather than simply the locations. • The existing literature focuses on pairwise sequence alignment, which results in a high amount of information loss [3], [6], [8]–[10]. • There is no unified metric to evaluate and compare the existing anonymization methods.

In this paper, we address the mentioned problems by proposing an enhanced anonymization framework termed machine learning based anonymization (MLA) to preserve the privacy of users in the publication of spatiotemporal trajectory datasets. MLA consists of two interworking algorithms: clustering and alignment. We have summarized our main contributions in the following bullet points.

By formulating the anonymization process as an optimization problem and finding an alternative representation of the system, we are able to apply machine clustering algorithms for clustering trajectories. We propose to use k 0 -means 1 algorithm for this purpose, as part of the MLA framework.

We propose a variation of k 0 -means algorithm to preserve the privacy of users in the publication of overly sensitive spatiotemporal trajectory datasets. • We enhance the performance of sequence alignment in clusters by considering multiple sequence alignment instead of pairwise sequence alignment.

We propose a utility metric to evaluate and compare the anonymization frameworks. MLA and all algorithms associated with it are applied on two real-life GPS datasets following different distributions in time and spatial domains. The experimental results indicate a significantly higher utility levels while maintaining k-anonymity of trajectories.

3. PROPOSED SYSTEM

The above techniques are not reliable as malicious users can identify how to crack groups and noise data to know user location. To overcome from this problem author has introduce Machine Learning based data privacy preserving technique which consists of 3 models and these 3 models will provide more security and anonymize or generalized which cannot be easily understand or crack.

1) Clustering model: in this model user locations will be clusters by using KMEANS algorithm and then calculate loss value. Loss value indicates difference between correct value and predicted value and the lesser the loss the better is the algorithm. The loss value will be saved to compare with Dynamic Sequence Alignment Loss and this Dynamic Sequence is called as Heuristic Clustering Algorithm.
2) Dynamic Sequence Alignment: In this module or algorithm we will take location form cluster member and then take random locations from original dataset and both this records will be aligned to get location which has minimal loss.
3) Data Generalization: in this module user location will be generalized or anonymised by summing up location with loss values.

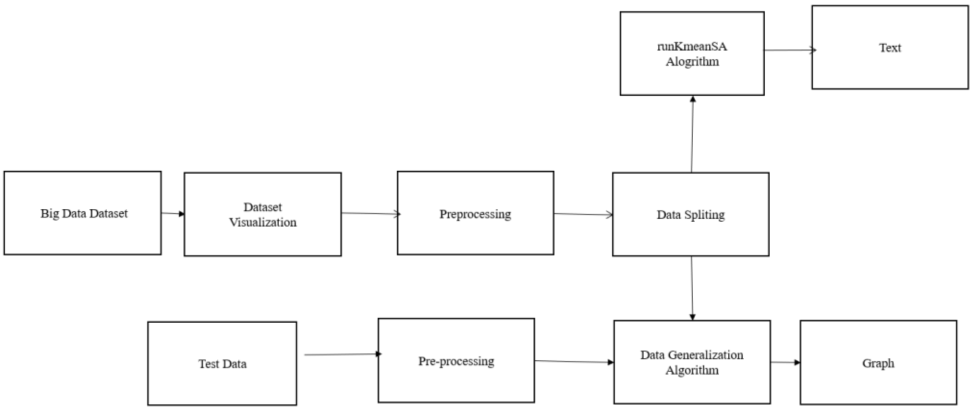


Figure 1 : Architecture diagram of proposed Model

Future work

Unfortunately, merely removing unique identifiers of users cannot protect their privacy, as databases can be linked to each other based on their quasi-identifiers. Doing so, adversaries can reveal sensitive information about the users and compromise their privacy. In this section, we review the existing approaches for the anonymization of spatiotemporal datasets.

Module implementation

Full-domain generalization: This technique emphasizes on the level that each value of an attribute is located in the generalization tree. If a value of an attribute is generalized to its parent node, all values of that attribute in the dataset must be generalized to the same level.

• Subtree generalization: In this method, if a value of an attribute is generalized to its parent node, all other child nodes of that parent node need to be replaced with the parent node as well .

• Cell generalization: This generalization technique considers each cell in the table separately. One cell can be generalized to its parent node while other values of that attribute remain unchanged**.**

Overview of the MLA Framework

demonstrates the overview of our proposed framework. The original dataset and the value of k are the inputs of the framework, and the output is the anonymized dataset preserving the privacy of users. The MLA framework consists of three mechanisms working together to anonymize spatiotemporal datasets, i.e., clustering, alignment, and generalization. A short description of each mechanism is provided as follow.

Clustering: At the highest level of the MLA framework, clustering is applied to seek for the most suitable grouping of trajectories that minimizes information loss. We propose to use k 0 -means clustering algorithm and a variation of it for overly sensitive datasets. Moreover, to have a baseline for comparison purposes, we develop a heuristic approach to cluster datasets. Our proposed clustering approaches are elaborated in Section

K-means Clustering

K-means is a widely used clustering algorithm that partitions a dataset into distinct groups, or clusters, based on similarity between data points. The algorithm's objective is to minimize the within-cluster sum of squared distances, effectively grouping data points that are close to each other while maintaining a certain number of clusters, denoted by 'k'. The working principle of K-means involves iteratively assigning each data point to the cluster whose mean (centroid) is closest and then recalculating the centroids. This process continues until the centroids no longer change significantly or a specified number of iterations is reached.

The mathematical representation of K-means includes defining the objective function that the algorithm aims to minimize.

Let X={x1,x2,...,xn} be the dataset with 'n' data points, and

C={c1,c2,...,ck} be the set of cluster centroids.

The objective function J is defined as

$$J(C) = \sum_{i=1}^{n} \min_{j=1}^{k} \|x_i - c_j\|^2$$

Here

$$\|x_i - c_j\|^2$$

represents the squared Euclidean distance between data point xi and centroid cj. The algorithm iterates through two steps: the assignment step, where each data point is assigned to the nearest centroid, and the update step, where the centroids are recalculated based on the current assignments. The process continues until convergence, ensuring the centroids represent the mean of the data points within each cluster.

The assignment step is defined as follows:

$$S_i = \arg\min_j \|x_i - c_j\|^2$$

This equation assigns each data point *Xi* to the cluster *Si* whose centroid *cj* minimizes the squared distance.

The update step calculates the new centroids based on the current assignments:

$$c_j = \frac{1}{|S_j|} \sum_{i \in S_j} x_i$$

Here, |Sj| represents the number of data points assigned to cluster j.

The key idea behind K-Means is to minimize the sum of squared distances between data points and their respective cluster centers.

Here's how it works:

Step 1: Initialization:

- Randomly select K initial cluster centroids (centers), where K is the predetermined number of clusters.
- These initial centroids can be chosen from the data points themselves or using other methods for better convergence.

Step 2: Assignment Step (Expectation Step):

- For each data point, calculate its distance to each of the K cluster centroids.
- Assign the data point to the cluster whose centroid is closest (typically using Euclidean distance).
- This step forms K clusters based on the initial centroids.

Step 3: Update Step (Maximization Step):

- Recalculate the cluster centroids by taking the mean of all data points assigned to each cluster.
- These new centroids represent the "center of mass" for each cluster.

Step 4: Repeat Steps 2 and 3:

- Iterate between the Assignment and Update steps until one of the stopping criteria is met:
    - Convergence: When the centroids no longer change significantly.
    - A predefined number of iterations is reached.

Step 5: Output:

- The final output of the K-Means algorithm is K clusters, where each data point belongs to one of the clusters.

Step 6: Interpretation:

- Once the clustering is complete, we can analyze and interpret the results. Each cluster represents a group of similar data points, and we can examine the characteristics and behaviors of each cluster.

4. **RESULTS AND D10ISCUSSION**

4.1 Implementation description

The Python script using the Tkinter library to create a graphical user interface (GUI) for a framework aimed at preserving privacy in the era of big data. The main functionalities include uploading a taxi trajectory dataset, preprocessing the dataset, running a KMeans clustering algorithm with a DynamicSA

(Dynamic Social Anonymization) algorithm, performing data generalization, and displaying a loss comparison graph.

Let's break down the implementation description:

- Import Statements: The script imports necessary modules and libraries, including Tkinter for GUI, NumPy for numerical operations, Pandas for data manipulation, BioPython for bioinformatics-related functionalities, and scikit-learn for machine learning operations.
- Global Variables: Global variables are declared to store information about the dataset, training data, cluster labels, and various loss metrics.
- GUI Setup: The Tkinter GUI is set up with a main window titled "Preserving privacy in the era of big data a ML based anonymization framework." Buttons are created for uploading the dataset, preprocessing, running the KMeans with DynamicSA algorithm, data generalization, and displaying a loss comparison graph.
- Function Definitions:
  uploadDataset: Prompts the user to upload a taxi trajectory dataset, reads the data using Pandas, and displays the first few records in the Tkinter Text widget.
  processDataset: Handles the preprocessing of the dataset by filling NaN values and extracting relevant columns.
  dynamicSA: Implements the DynamicSA algorithm, which involves comparing sequences and choosing records based on alignment scores.
  runKmeansSA: Applies the KMeans clustering algorithm to the dataset, evaluates the accuracy, and performs dynamic social anonymization. It calculates and displays KMeans and heuristic losses.
  dataGeneralization: Performs generalization on the processed trajectory data by adjusting latitude and longitude values based on stored losses.
  graph: Displays a bar graph comparing heuristic and KMeans losses.
- GUI Elements:
  Labels, buttons, and text widgets are created and configured with specific fonts, sizes, and positions in the Tkinter window.
- Main Loop:
  The Tkinter main loop (main.mainloop()) is initiated, which keeps the GUI running and responsive to user interactions.
- Description:
  The script essentially provides a graphical interface for a privacy-preserving framework applied to taxi trajectory data. It incorporates KMeans clustering and a heuristic-based approach to anonymize sensitive information in location data, and it allows users to visualize the results through the GUI.

4.2 Results description:

The figure 2 depicts the main interface of the application, providing an overview of the tool for anonymization framework. It includes various features and options for users to interact with the application.

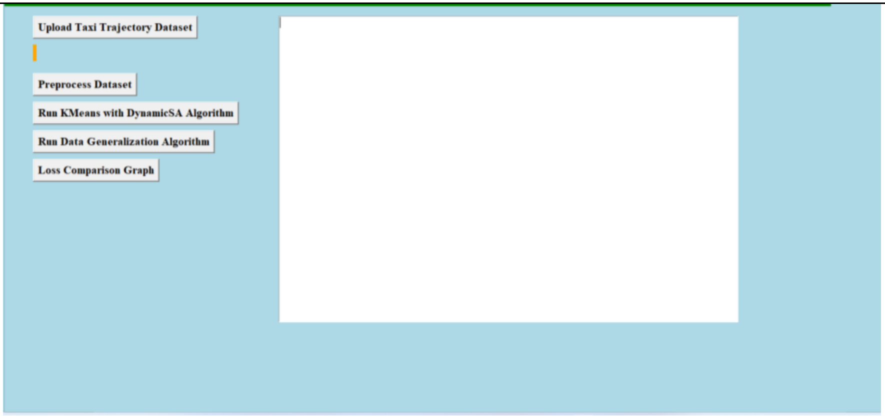
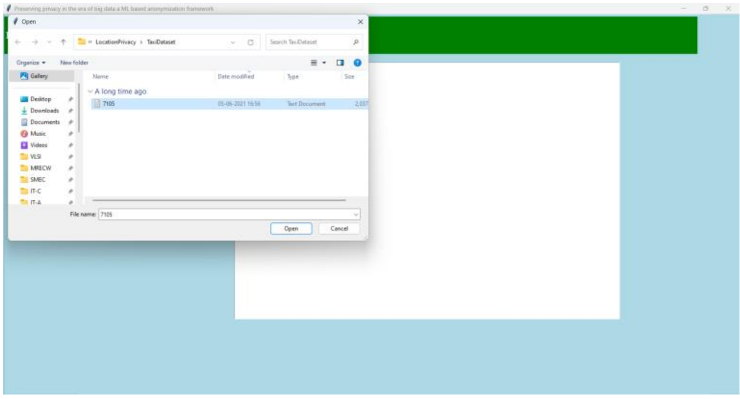

Upload Taxi Trajectory Dataset
Preprocess Dataset
Run KMeans with DynamicSA Algorithm
Run Data Generalization Algorithm
Loss Comparison Graph

Figure 2: Displays the GUI of ML based anonymization framework.

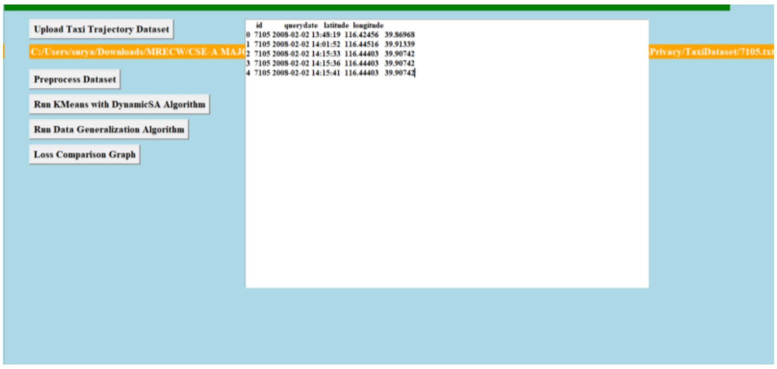
The figure 3 shows the user interface or a section of the application where the user can select the Taxi Trajectory Dataset for analysis. It include options to browse and load the dataset into the system.



File name: 7105

Figure 3: Displays the selection of Taxi Trajectory Dataset

The figure 4 displays a representation of the sample dataset, providing a glimpse of the data's structure, including columns and rows, to give users an overview of the information it contains.

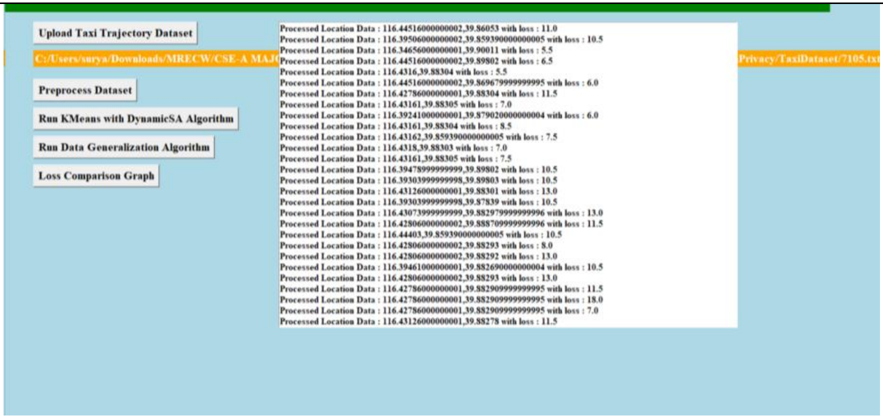

Upload Taxi Trajectory Dataset
Preprocess Dataset
Run KMeans with DynamicSA Algorithm
Run Data Generalization Algorithm
Loss Comparison Graph

C:/Users/surya/Downloads/MRECW/CSE-A MAJ ... Privacy/TaxiDataset/7105.txt

```
   id        querydate  latitude  longitude
0  7105 2008-02-02 13:48:19  116.42456   39.86968
1  7105 2008-02-02 14:01:52  116.44516   39.91339
2  7105 2008-02-02 14:15:33  116.44403   39.90742
3  7105 2008-02-02 14:15:36  116.44403   39.90742
4  7105 2008-02-02 14:15:41  116.44403   39.90742
```

Figure 4: Represents the sample dataset.

The figure 5 shows the result of processing location data using the Kmeans algorithm with dynamicSA Algorithm. It highlights how clustering has been applied to the location data.

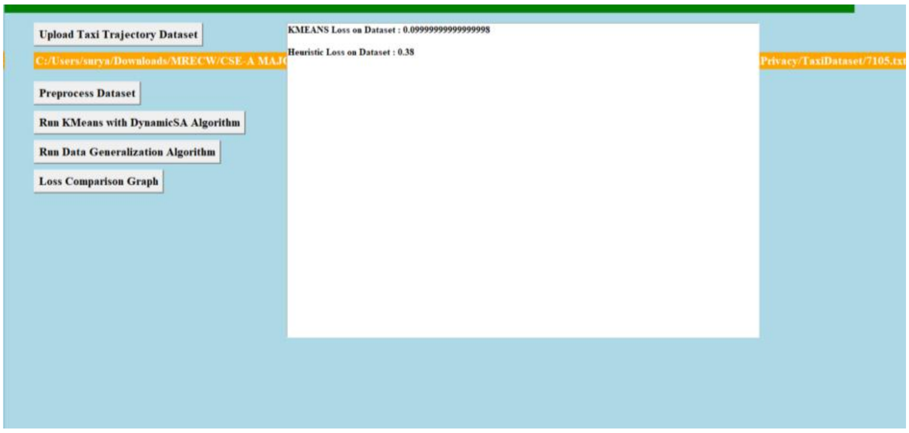

Upload Taxi Trajectory Dataset
Preprocess Dataset
Run KMeans with DynamicSA Algorithm
Run Data Generalization Algorithm
Loss Comparison Graph

C:/Users/surya/Downloads/MRECW/CSE-A MAJ Privacy/TaxiDataset/7105.txt

```
Processed Location Data : 116.44516000000002,39.86053 with loss : 11.0
Processed Location Data : 116.39506000000002,39.85939000000005 with loss : 10.5
Processed Location Data : 116.34656000000001,39.90011 with loss : 5.5
Processed Location Data : 116.44516000000002,39.89802 with loss : 6.5
Processed Location Data : 116.4316,39.88304 with loss : 5.5
Processed Location Data : 116.44516000000002,39.86967999999995 with loss : 6.0
Processed Location Data : 116.42786000000001,39.88304 with loss : 11.5
Processed Location Data : 116.43161,39.88305 with loss : 7.0
Processed Location Data : 116.39241000000001,39.87902000000004 with loss : 6.0
Processed Location Data : 116.43161,39.88304 with loss : 8.5
Processed Location Data : 116.43162,39.85939000000005 with loss : 7.5
Processed Location Data : 116.4318,39.88303 with loss : 7.0
Processed Location Data : 116.43161,39.88305 with loss : 7.5
Processed Location Data : 116.39478999999999,39.89802 with loss : 10.5
Processed Location Data : 116.39303999999998,39.89803 with loss : 10.5
Processed Location Data : 116.43126000000001,39.88301 with loss : 13.0
Processed Location Data : 116.39303999999998,39.87839 with loss : 10.5
Processed Location Data : 116.43073999999999,39.88297999999996 with loss : 13.0
Processed Location Data : 116.42806000000002,39.88709999999996 with loss : 11.5
Processed Location Data : 116.44403,39.85939000000005 with loss : 10.5
Processed Location Data : 116.42806000000002,39.88293 with loss : 8.0
Processed Location Data : 116.42806000000002,39.88292 with loss : 13.0
Processed Location Data : 116.39461000000001,39.88269000000004 with loss : 10.5
Processed Location Data : 116.42806000000002,39.88293 with loss : 13.0
Processed Location Data : 116.42786000000001,39.88290999999995 with loss : 11.5
Processed Location Data : 116.42786000000001,39.88290999999995 with loss : 18.0
Processed Location Data : 116.42786000000001,39.88290999999995 with loss : 7.0
Processed Location Data : 116.43126000000001,39.88278 with loss : 11.5
```

Figure 5: Represents the processed location data by applying Kmeans with dynamicSA Algorithm

The figure 6 shows the losses associated with applying the Kmeans and Heuristic algorithms on the dataset. It includes visualizations or numerical representations of the losses over iterations or data points.


Upload Taxi Trajectory Dataset
Preprocess Dataset
Run KMeans with DynamicSA Algorithm
Run Data Generalization Algorithm
Loss Comparison Graph

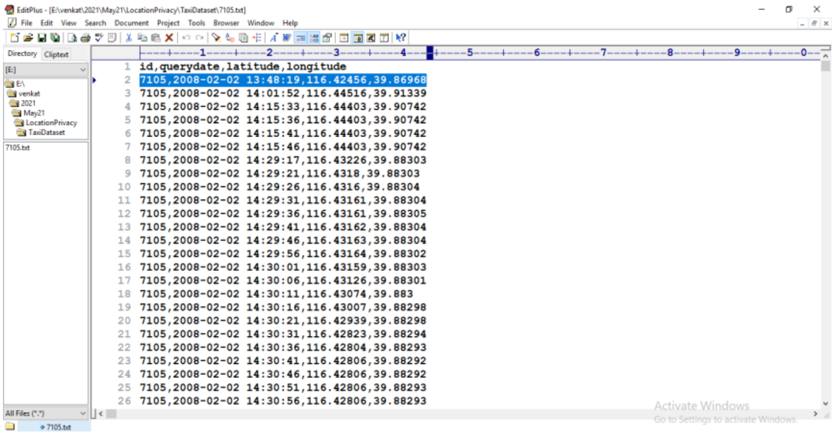
C:/Users/surya/Downloads/MRECW/CSE-A MAJ Privacy/TaxiDataset/7105.txt

```
KMEANS Loss on Dataset : 0.09999999999999998

Heuristic Loss on Dataset : 0.38
```

Figure 6: Displays the Kmeans and Heuristic losses on dataset.

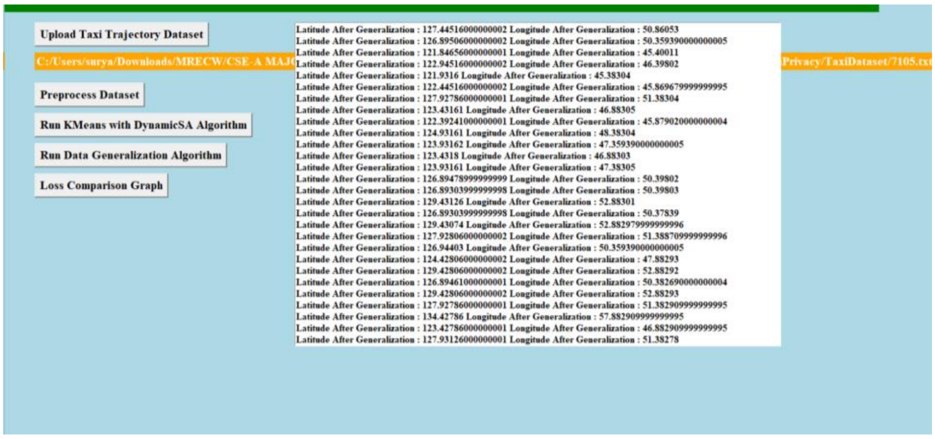
The figure 7 display the actual, real-world location values from the dataset. It is a visualization of the raw data before any processing or clustering has been applied.


EditPlus - [E:\venkat\2021\May21\LocationPrivacy\TaxiDataset\7105.txt]
File Edit View Search Document Project Tools Browser Window Help
Directory Cliptext
[E:]
E:\
venkat
2021
May21
LocationPrivacy
TaxiDataset
7105.txt
All Files (*.*)
7105.txt
Activate Windows
Go to Settings to activate Windows.

```
id,querydate,latitude,longitude
7105,2008-02-02 13:48:19,116.42456,39.86968
7105,2008-02-02 14:01:52,116.44516,39.91339
7105,2008-02-02 14:15:33,116.44403,39.90742
7105,2008-02-02 14:15:36,116.44403,39.90742
7105,2008-02-02 14:15:41,116.44403,39.90742
7105,2008-02-02 14:15:46,116.44403,39.90742
7105,2008-02-02 14:29:17,116.43226,39.88303
7105,2008-02-02 14:29:21,116.4318,39.88303
7105,2008-02-02 14:29:26,116.4316,39.88304
7105,2008-02-02 14:29:31,116.43161,39.88304
7105,2008-02-02 14:29:36,116.43161,39.88305
7105,2008-02-02 14:29:41,116.43162,39.88304
7105,2008-02-02 14:29:46,116.43163,39.88304
7105,2008-02-02 14:29:56,116.43164,39.88302
7105,2008-02-02 14:30:01,116.43159,39.88303
7105,2008-02-02 14:30:06,116.43126,39.88301
7105,2008-02-02 14:30:11,116.43074,39.883
7105,2008-02-02 14:30:16,116.43007,39.88298
7105,2008-02-02 14:30:21,116.42939,39.88298
7105,2008-02-02 14:30:31,116.42823,39.88294
7105,2008-02-02 14:30:36,116.42804,39.88293
7105,2008-02-02 14:30:41,116.42806,39.88292
7105,2008-02-02 14:30:46,116.42806,39.88292
7105,2008-02-02 14:30:51,116.42806,39.88293
7105,2008-02-02 14:30:56,116.42806,39.88293
```

Figure 7: The screen has the real location values of dataset.

The figure 8 represents the latitude and longitude values after applying a data generalization algorithm. Data generalization is often used to protect privacy or reduce granularity in location data


Upload Taxi Trajectory Dataset
C:/Users/surya/Downloads/MRECW/CSE-A MAJ
Privacy/TaxiDataset/7105.txt
Preprocess Dataset
Run KMeans with DynamicSA Algorithm
Run Data Generalization Algorithm
Loss Comparison Graph

```
Latitude After Generalization : 127.44516000000002 Longitude After Generalization : 50.86053
Latitude After Generalization : 126.89506000000002 Longitude After Generalization : 50.359390000000005
Latitude After Generalization : 121.84656000000001 Longitude After Generalization : 45.40011
Latitude After Generalization : 122.94516000000002 Longitude After Generalization : 46.39802
Latitude After Generalization : 121.9316 Longitude After Generalization : 45.38304
Latitude After Generalization : 122.44516000000002 Longitude After Generalization : 45.869679999999995
Latitude After Generalization : 127.92786000000001 Longitude After Generalization : 51.38304
Latitude After Generalization : 123.43161 Longitude After Generalization : 46.88305
Latitude After Generalization : 122.39241000000001 Longitude After Generalization : 45.879020000000004
Latitude After Generalization : 124.93161 Longitude After Generalization : 48.38304
Latitude After Generalization : 123.93162 Longitude After Generalization : 47.359390000000005
Latitude After Generalization : 123.4318 Longitude After Generalization : 46.88303
Latitude After Generalization : 123.93161 Longitude After Generalization : 47.38305
Latitude After Generalization : 126.89478999999999 Longitude After Generalization : 50.39802
Latitude After Generalization : 126.89303999999998 Longitude After Generalization : 50.39803
Latitude After Generalization : 129.43126 Longitude After Generalization : 52.88301
Latitude After Generalization : 126.89303999999998 Longitude After Generalization : 50.37839
Latitude After Generalization : 129.43074 Longitude After Generalization : 52.882979999999996
Latitude After Generalization : 127.92806000000002 Longitude After Generalization : 51.388709999999996
Latitude After Generalization : 126.94403 Longitude After Generalization : 50.359390000000005
Latitude After Generalization : 124.42806000000002 Longitude After Generalization : 47.88293
Latitude After Generalization : 129.42806000000002 Longitude After Generalization : 52.88292
Latitude After Generalization : 126.89461000000001 Longitude After Generalization : 50.382690000000004
Latitude After Generalization : 129.42806000000002 Longitude After Generalization : 52.88293
Latitude After Generalization : 127.92786000000001 Longitude After Generalization : 51.382909999999995
Latitude After Generalization : 134.42786 Longitude After Generalization : 57.882909999999995
Latitude After Generalization : 123.42786000000001 Longitude After Generalization : 46.882909999999995
Latitude After Generalization : 127.93126000000001 Longitude After Generalization : 51.38278
```

Figure 8: Displays the latitude and longitude values after running data generalization algorithm.

The figure 9 provides a comparison of the losses incurred by the Heuristic and Kmeans algorithms. It help in evaluating the effectiveness of these methods in the context of the specific dataset and task.

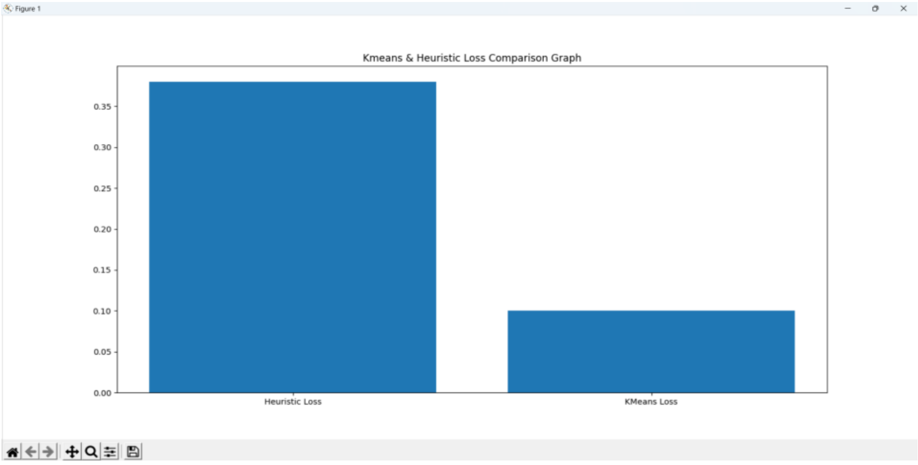


Figure 9: Shows the comparison of heuristic and Kmeans

5. CONCLUSION

In this paper, we have proposed a framework to preserve the privacy of users while publishing the spatiotemporal trajectories. The proposed approach is based on an efficient alignment technique termed as progressive sequence alignment in addition to a machine learning clustering approach that aims at minimizing the incurred loss in the anonymization process. We also devised a variation of k 0 -means algorithm for guaranteeing the k-anonymity in overly sensitive datasets. The experimental results on real-life GPS datasets indicate the superior spatial utility performance of our proposed framework compared with the previous works

REFERENCES

[1] S. Shaham, M. Ding, B. Liu, Z. Lin, and J. Li, “Machine learning aided anonymization of spatiotemporal trajectory datasets,” arXiv preprint arXiv:1902.08934, 2019.

[2] A. Government, "New australian government data sharing and release legislation," 2018.

[3] A. Tamersoy, G. Loukides, M. E. Nergiz, Y. Saygin, and B. Malin, "Anonymization of longitudinal electronic medical records," IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 3, pp. 413–423, 2012.

[4] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data," in Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017, pp. 1241–1250.

[5] Y. Dong and D. Pi, "Novel privacy-preserving algorithm based on frequent path for trajectory data publishing," Knowledge-Based Systems, vol. 148, pp. 55–65, 2018.

[6] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, "Towards privacy-preserving publishing of spatiotemporal trajectory data," arXiv preprint arXiv:1701.02243, 2017.

[7] M. Terrovitis, G. Poulis, N. Mamoulis, and S. Skiadopoulos, "Local suppression and splitting techniques for privacy preserving publication of trajectories," IEEE Trans. Knowl. Data Eng, vol. 29, no. 7, pp. 1466–1479, 2017.

[8] M. E. Nergiz, M. Atzori, and Y. Saygin, "Towards trajectory anonymization: a generalization-based approach," in Proc. of the SIGSPATIAL ACM GIS. ACM, 2008, pp. 52–61.

[9] S. Gurung, D. Lin, W. Jiang, A. Hurson, and R. Zhang, "Traffic information publication with privacy preservation," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 5, no. 3, p. 44, 2014.

[10] R. Yarovoy, F. Bonchi, L. V. Lakshmanan, and W. H. Wang, "Anonymizing moving objects: How to hide a mob in a crowd?" in Proc. of the 12th International Conference on Extending Database Technology: Advances in Database Technology. ACM, 2009, pp. 72– 83.

[11] B. Liu, W. Zhou, T. Zhu, L. Gao, and Y. Xiang, "Location privacy and its applications: A systematic study," IEEE Access, vol. 6, pp. 17 606–17 624, 2018.

[12] G. Poulis, G. Loukides, S. Skiadopoulos, and A. GkoulalasDivanis, "Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints," Journal of biomedical informatics, vol. 65, pp. 76–96, 2017.

[13] T. Takahashi and S. Miyakawa, "Cmoa: Continuous moving object anonymization," in Proceedings of the 16th International Database Engineering & Applications Sysmposium. ACM, 2012, pp. 81–90.

[14] X. Zhou and M. Qiu, "A k-anonymous full domain generalization algorithm based on heap sort," in International Conference on Smart Computing and Communication. Springer, 2018, pp. 446–459.

[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, 2005, pp. 49–60.