International Journal of Gender, Science and Technology

ISSN: 2040-0748

UGC Care Group I Journal

Vol-12 Issue-02 Dec 2023

PRIVACY PRESERVING LOCATION DATA PUBLISHING - A MACHINE LEARNING APPROACH

Mrs.T.Venkata Lavanya[1].,Mokshit.A[2].,Akshay Sagar Mekkonda[3].,Manoj Kumar Kindidhoddi[4].,5Soujanya Manne.

[1] Assistant Professor, [2,3,4,5]StudentsB.Tech-IT.

MallaReddyInstituteofTechnologyandScience.,Maisammaguda.,Medchal.,Telangana,India

- *[1]lavanya.tanneru91@gmail.com,[2] mokshitailum17@gmail.com, akshaysagar2002@email.com,[4]manojmannu899@gmail.com,[5]mannesoujanya 98@gmail.com.*

ABSTRACT

An important part of open data research and helping to make government bodies more transparent is publishing datasets. However, users' private information may be exposed if such data were to be made public. When it comes to data sources, spatiotemporal trajectory datasets are among the most delicate. Removing unique identifiers is not enough to protect consumers' privacy, unfortunately. In order to identify users, attackers can know partial trajectories or find ways to connect the publicly available information to other sources. Hence, privacy-preserving methods must be used prior to the release of spatiotemporal trajectory datasets. Machine learning based anonymization (MLA) is a strong approach that we provide in this study to anonymize spatiotemporal trajectory datasets. We suggest using the k-means technique to cluster the trajectories, which is possible thanks to a reformulated issue that allows us to utilize machine learning algorithms. We also suggest a variant of the k-means method that may protect privacy in really sensitive information. Additionally, by including multiple sequence alignment into the MLA, we enhance the alignment process. Using the T-Drive, Geolife, and Gowalla location datasets, the framework and all of the suggested algorithms are put into action. Based on the experimental findings, datasets anonymized using the MLA framework are far more useful.

1. INTRODUCTION

Open research and government agency openness depend on several groups and institutions publishing data. The Australian government's dedicated website for publishing information, "data.gov.au," has seen the inclusion of over 7,000 datasets since 2013. Furthermore, several government institutions in Australia may be required to disclose their data as early as 2019 according to new data sharing laws [2]. Data publishing is inherently hazardous as it could lead to the exposure of personally identifiable information. Datasets must be stripped of any personally identifying information before they may be made public. Nonetheless, user privacy cannot be adequately protected by such an operation. An adversary may know the users' paths in advance or be able to re-identify people in datasets using shared characteristics termed quasi-identifiers. They are able to expose sensitive information that might hurt individuals physically, financially, and reputationally thanks to this supplementary data.

Information pertaining to a person's whereabouts and the passage of time is one of the most delicate types of data. Publishing spatiotemporal data has many potential benefits for both users and academics, but it also seriously compromises individuals' privacy. Think of someone who, on weekday mornings, has been used GPS guidance to go from their house to their workplace. An attacker may potentially identify a user if they have any background information about them, including their home address. The user's health status and the frequency of trips to the user's specialist may be revealed by such an inference attack, compromising user privacy. Hence, before making spatiotemporal datasets public, it is essential to make them anonymous. If the attacker connects identifiable users to other databases, like the medical

records database, the privacy problem becomes even worse. Because of this, most businesses today are wary about making public any spatiotemporal trajectory statistics unless they have implemented a robust privacy preservation strategy.

For the release of spatiotemporal datasets, k-anonymity is a commonly used privacy measure. Simply put, this measure checks whether there are at least k − 1 additional trajectories in the published dataset that are indistinguishable from each other. In [3], the authors presented a generalization-based anonymization method that used the idea of k-anonymity for spatiotemporal datasets. The impact of variables like spatiotemporal resolution and the quantity of released users on the anonymization process was examined by Xu et al. [4]. The goal of Dong et al. [5] was to enhance the current methods of clustering. Achieving kanonymity by clustering comparable trajectories and excluding those that are very divergent was their suggested anonymization method. To protect users' anonymity from probabilistic assaults and anonymize trajectory datasets, the developers of [6] created a technique named k-merge. Suppression and splitting on a local level.

There isn't a clear approach to cluster trajectories since it's hard to quantify the cost of grouping when looking at the distances between trajectories instead of just their positions. There is a lot of information loss since the current research is centered on pairwise sequence matching [3, 6, 8–10]. Presently, there is no standard measure by which to assess and contrast different anonymization techniques.

To solve these issues and protect users' privacy while publishing spatiotemporal trajectory information, we provide an improved anonymization method called machine learning based anonymization (MLA). The two algorithms that make up MLA, clustering and alignment, operate together. In the following bullet points, we have outlined our primary contributions.

We may use machine clustering techniques for clustering trajectories by recasting the anonymization process as an optimization problem and discovering a different way to describe the system. For this, we suggest incorporating the k0-means-1 algorithm into the MLA architecture.

• To protect users' anonymity while sharing critical spatiotemporal trajectory datasets, we suggest a variant of the k-means method. • By shifting our focus from pairwise to multiple sequence alignment, we improve cluster sequence alignment performance.

We provide a utility measure for comparing and contrasting the anonymization systems. We use MLA and all its related techniques to two real-world GPS datasets that follow distinct temporal and space distributions. Maintaining k-anonymity of trajectories leads to much greater utility levels, according to the experimental data.

2. PROPOSED SYSTEM

On the other hand, bad actors may still figure out how to breach groups and noisy data, making the aforementioned methods useless for pinpointing user locations. The author has introduced a three-model machine learning-based data privacy preservation approach to address this issue; these models will increase security, generalize the data so it is difficult to interpret or break, and anonymize the data.

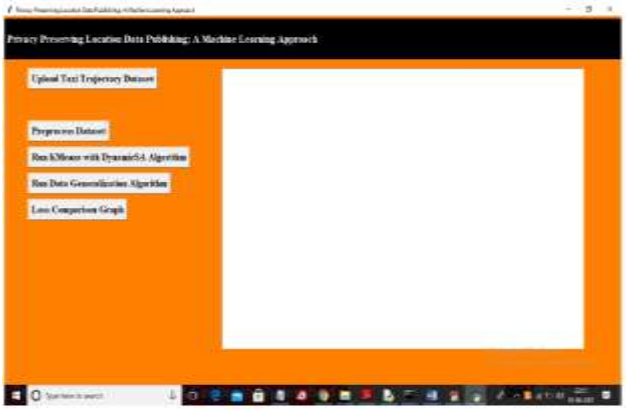
1. A model for clustering: this model uses the K-means method to group user locations into clusters, and then it calculates the loss value. The loss number shows how much of a discrepancy there is between the actual and anticipated values; a smaller loss suggests that the algorithm is doing better. We will save the loss figure for future comparisons.

using a Dynamic Sequence, which is known as a Heuristic Clustering Algorithm, and a Dynamic Sequence Alignment Loss.

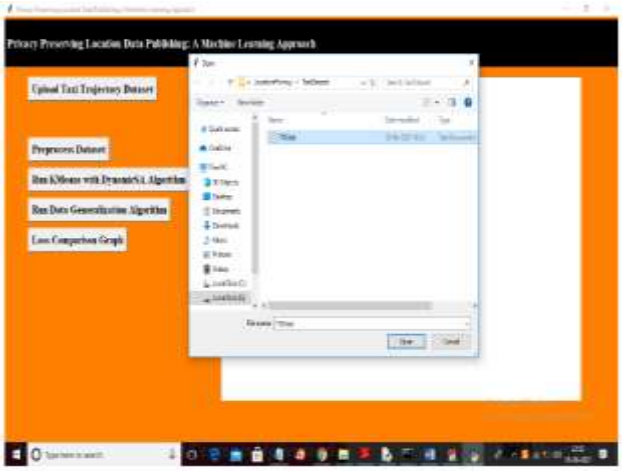
2. Dynamic Sequence Alignment: This method or module will align records with little loss by first taking locations from cluster members and then randomly selecting locations from the original dataset.

3. Generalizing Data: This module will combine user location with loss values to make their location less specific or anonymous.

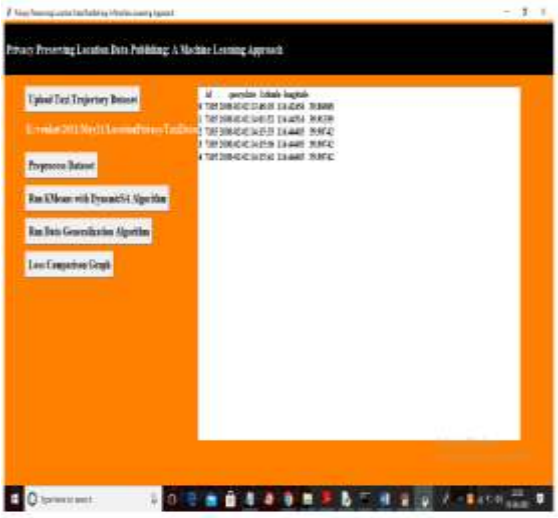
3. RESULTS



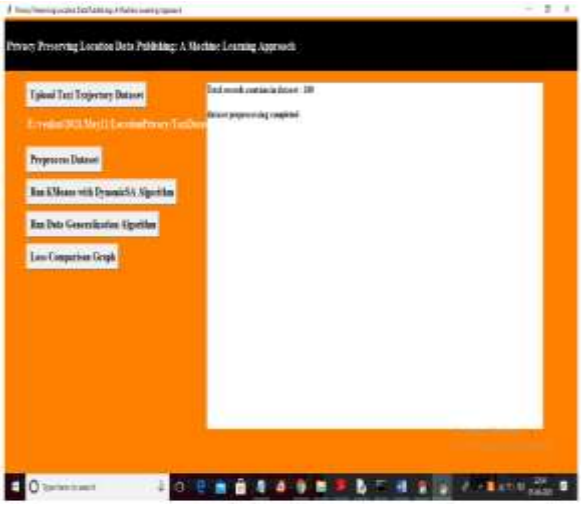
Press the "Upload Taxi Trajectory Dataset" button on the previous page to submit the dataset.



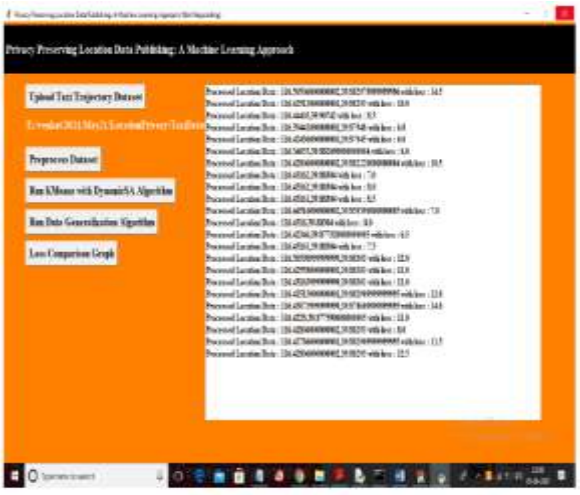
After you've chosen and uploaded your taxi trajectory file, click the "Open" button to load the dataset. Then, you'll see the screen below.



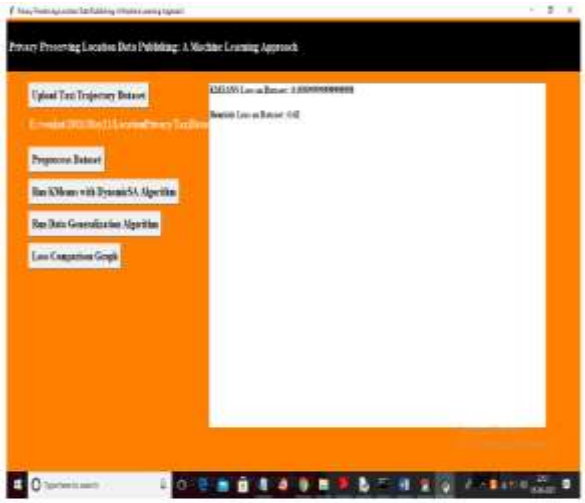
The dataset is loaded on the above page; to eliminate empty values and extract latitude and longitude locations, click the "Preprocess Dataset" button.



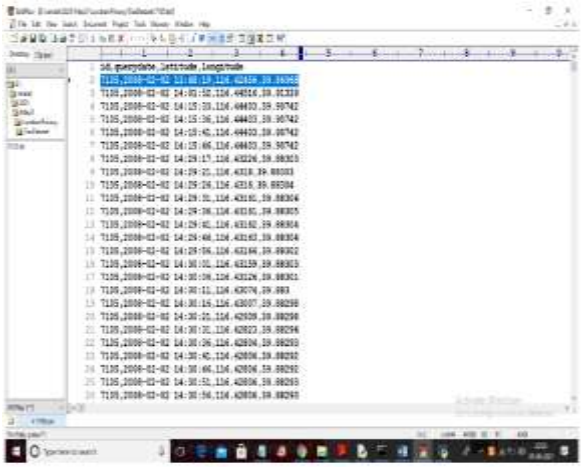
To run KMEANS on the dataset with Dynamic SA, click on the "Run KMeans with DynamicSA Algorithm" button on the top page, which will finish the dataset preprocessing. After clustering all locations that are comparable, this algorithm will run DYNAMIC SA.



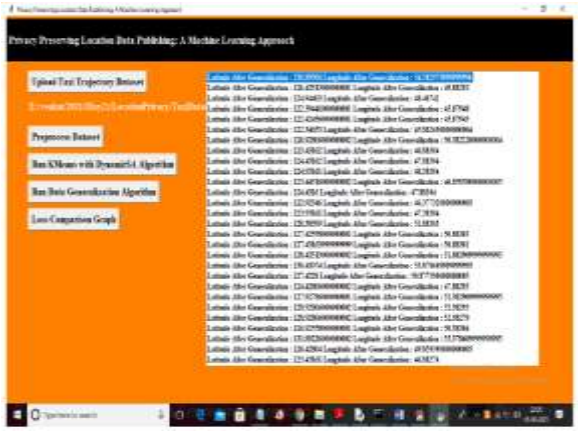
Aligning two locations by selecting the one with the lowest loss occurs on the above screen as a result of processing each location and computing the loss value using dynamic sequence alignment.



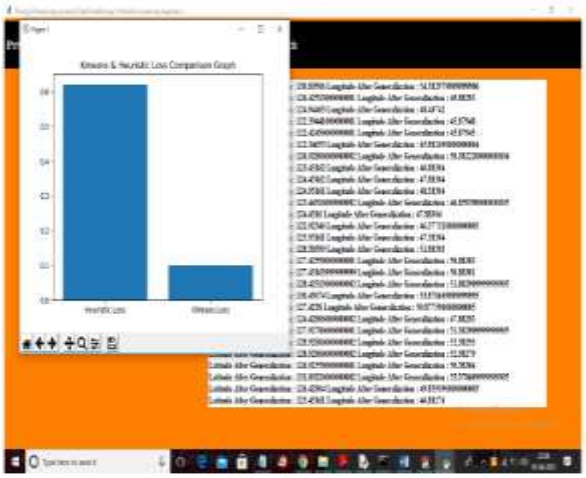
The KMEANS loss is 0.09 while the Heuristic Clustering loss, sometimes called Dynamic SA, is 0.62 on the above screen. Press the "Run Data Generalization Algorithm" button to generate loss-value generalized data. Below, you can see the original dataset location values in the first record, and then, after applying the aforementioned algorithms, you can view the same locations generalized or anonymized on the second screen.



You may see the same location is used with different values in the screen below.



All of the location data are generic in the above screen so that bad actors can't deduce precise locations. To see the following graph, choose "Loss Comparison Graph" from the menu.



In above graph x-axis represents algorithm name and y-axis represents loss values generated for that algorithm and in above graph KMEANS got less loss so KMEANS is better in anonymization.

4. CONCLUSION

In this research, we provide a system that can broadcast users' spatiotemporal trajectories while protecting their anonymity. The suggested method relies on a machine learning clustering strategy that seeks to minimize the loss experienced during anonymization and an effective alignment method known as progressive sequence alignment. Additionally, for really sensitive information, we developed a variant of the k-means method to ensure k-anonymity. Our suggested methodology

outperforms prior efforts in terms of spatial usefulness, according to experimental findings using real-world GPS datasets.

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.