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.pp139-146*

PRESERVING PRIVACY IN THE ERA OF BIG DATA A ML BASED ANONYMIZATION FRAMEWORK

Aarthi Kasthuri[1], A. Shruthilaya[2], B. Jayavarshitha[2], C. Kavya [2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science Engineering
[1,2]Malla Reddy Engineering College for Women, Maisammaguda, Dhulapally, Kompally, Secunderabad-500100, Telangana, India

To Cite this Article

Aarthi Kasthuri, A. Shruthilaya, B. Jayavarshitha, C. Kavya **, "PRESERVING PRIVACY IN THE ERA OF BIG DATA A ML BASED ANONYMIZATION FRAMEWORK"** *Journal of Science and Technology, Vol. 08, Issue 12 - Dec 2023, pp139-146*

Article Info

Received: 17-11-2023 Revised: 27 -11-2023 Accepted: 07-12-2023 Published: 17-12-2023

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: Anonymization Frame work, Machine Learning, Big Data.

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.

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-

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.pp139-146*

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

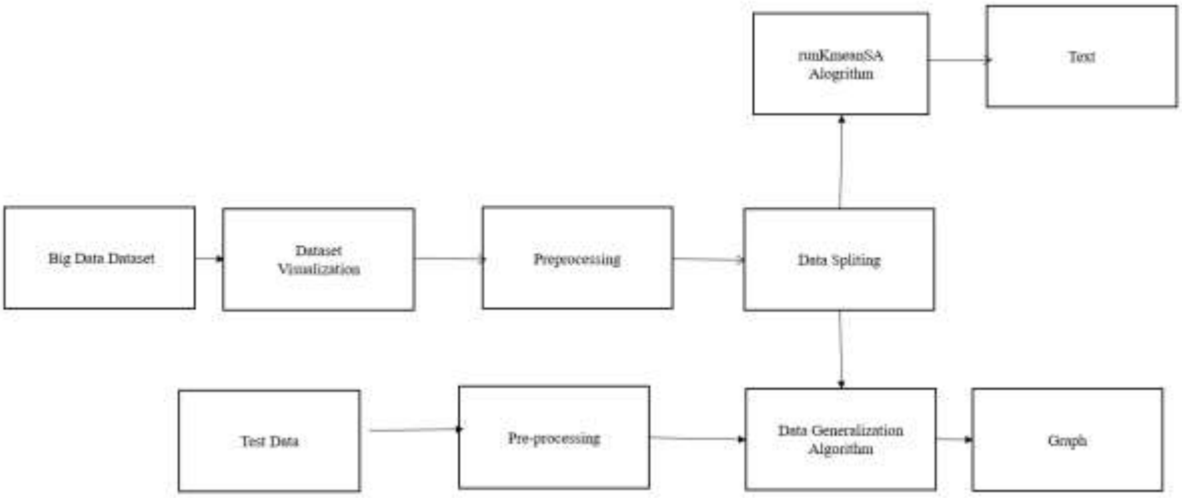


Figure.1: Block diagram of proposed diagram.

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 this 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.

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.

ISSN: 2456-5660 Volume 8, Issue 12 (Dec -2023)
www.jst.org.in DOI:https://doi.org/10.46243/jst.2023.v8.i12.pp139-146

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

Advantages of proposed system

K-Means clustering has several advantages, which make it a popular choice for various data clustering tasks:

- Simplicity and Speed: K-Means is relatively easy to understand and implement. It's computationally efficient and can handle large datasets with ease, making it suitable for real-time or batch processing.
- Scalability: K-Means scales well with the number of data points and can handle high-dimensional data efficiently. This scalability is particularly valuable when dealing with big data.
- Versatility: K-Means can be applied to a wide range of data types and is not limited to any specific domain. It is commonly used in areas such as customer segmentation, image compression, document clustering, and more.
- Interpretability: The results of K-Means are easy to interpret. Each cluster represents a group of similar data points, allowing for meaningful insights and straightforward visualizations.
- Deterministic Results: Given the same initial conditions, K-Means will produce the same results. This determinism is useful for reproducibility and consistency in data analysis.
- Efficient for Large Datasets: K-Means doesn't require storing the entire dataset in memory during processing, making it memory-efficient and suitable for datasets that don't fit into memory.
- Robustness: K-Means can handle noisy data and outliers to some extent. However, preprocessing steps like outlier removal may be necessary for better results.
- Parallelization: K-Means is parallelizable, and various libraries and tools offer parallel implementations, which can significantly speed up the clustering process on multi-core processors or distributed computing environments.
- Initialization Methods: Advanced initialization methods like K-Means++ help improve convergence and reduce the chances of getting stuck in suboptimal solutions.
- Consistent Clusters: In most cases, K-Means produces relatively stable clusters over multiple runs, especially when using K-Means++ initialization.
- Quantization and Compression: K-Means can be used for image compression and data quantization, reducing data storage requirements while preserving essential information.

4. RESULTS AND DISCUSSION

Dataset Description:

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.pp139-146*

This dataset contain information related to geographical locations and time. Here's a brief description of each column:

- id: This column represent a unique identifier for each record in the dataset. It is a sequential number identification.
- querydate: This column represents the date and time when a geographical location was queried or recorded. It include both date and time information.
- latitude: This column contains numerical values representing the latitude of a geographical location. Latitude is a measure of how far north or south a location is from the equator.
- longitude: This column contains numerical values representing the longitude of a geographical location. Longitude is a measure of how far east or west a location is from the prime meridian.

Results description:

- The figure 2 depicts the main interface of the application, providing an overview of the tool for anonymization framework. It include various features and options for users to interact with the application.
- The figure 3 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.
- The figure 4 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
- The figure 5 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 2: Displays the GUI of ML based anonymization framework.

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.pp139-146*

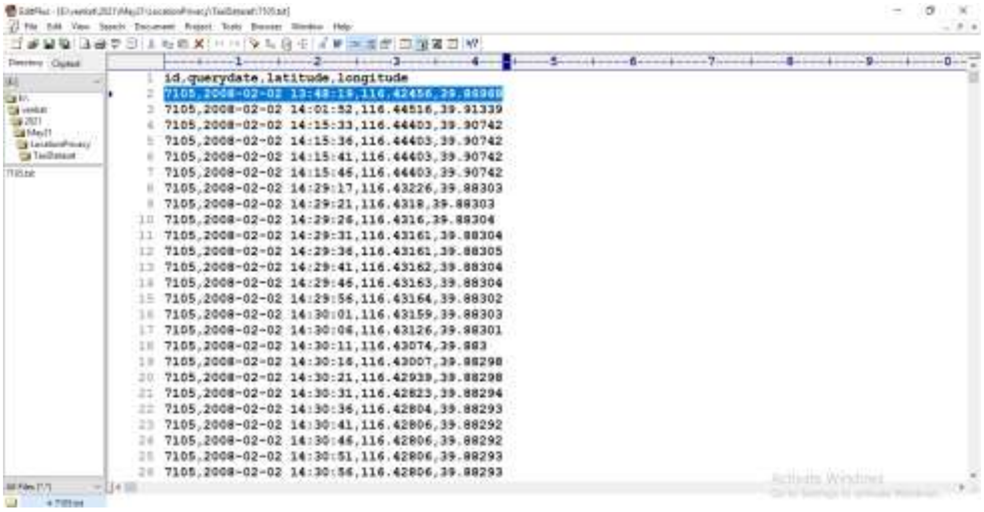


Figure 3: The screen has the real location values of dataset.

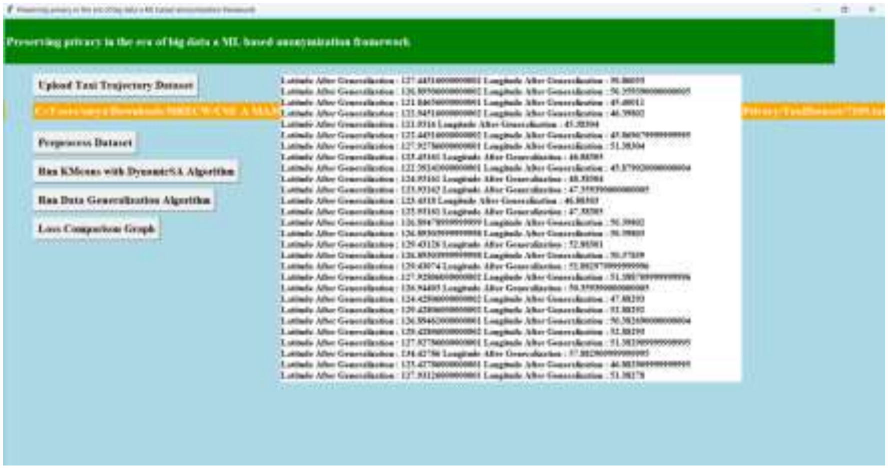
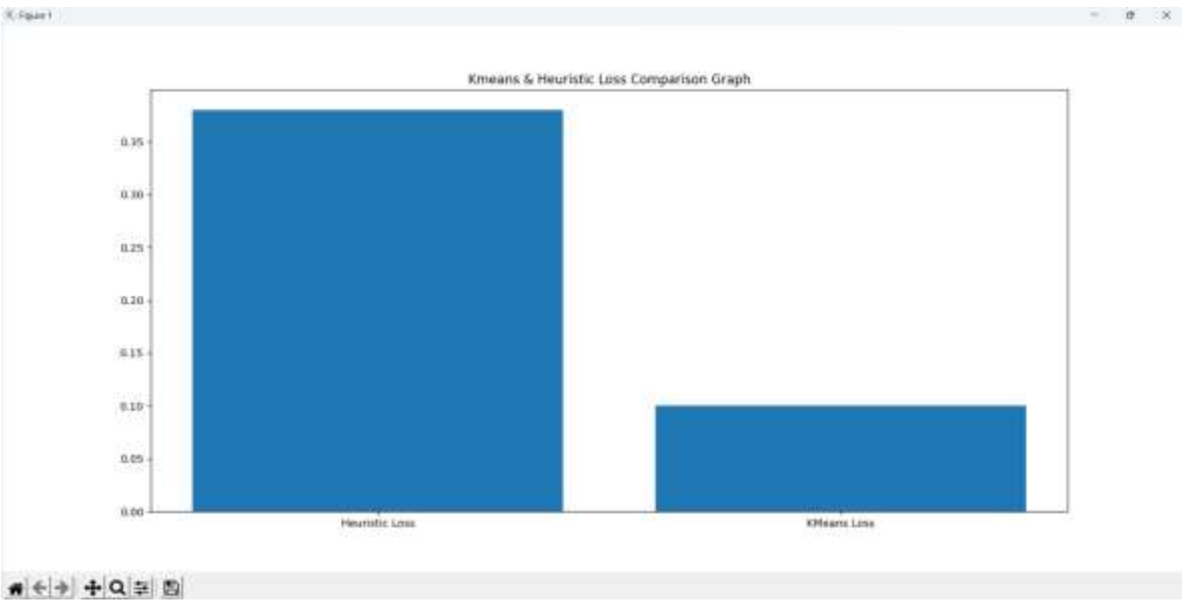


Figure 4: Displays the latitude and longitude values after running data generalization algorithm.



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.pp139-146*

Figure 5: Shows the comparison of heuristic and Kmeans loss.

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.