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.pp23-30*

PRIVACY PRESERVING LOCATION DATA PUBLISHING: A MACHINE LEARNING APPROACH

D Nikhil Teja[1], Patan Abdulsalam Khan[2], A Sunny[3], R Shashi Rekha[4]

[1,2,3]B. Tech Student, Department of CSE (Cyber Security), Malla Reddy College of Engineering and Technology, Hyderabad, India.
[4]Assistant Professor, Department Of CSE (Data Science), Malla Reddy College Of Engineering and Technology, Hyderabad, India.

To Cite this Article
D Nikhil Teja, Patan Abdulsalam Khan , A Sunny, R Shashi Rekha, " PRIVACY PRESERVING LOCATION DATA PUBLISHING: A MACHINE LEARNING APPROACH"
Journal of Science and Technology, Vol. 08, Issue 12,- Dec 2023, pp23-30

Article Info
Received: 12-11-2023 Revised: 22-11-2023 Accepted: 02-12-2023 Published: 12-12-2023

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: Publishing of data, Machine learning, Privacy preserving.

I INTRODUCTION

Privacy preservation plays a major role on data mining and transfers the data between different users. Publishing the data or information can hide the user's id, latitude, longitude, time, date and can share the data to the third party. There are large amount data
includes person's private details like id, gender, location etc. The admin can generate key to the third party to identify the hidden details of a person for analyzing the data. One of the most sensitive data is location trajectories. Spatiotemporal dataset is used in this framework, which include GPS trajectories for mobile users. The database

includes k- anonymity for grouping the similar trajectories. The privacy metric for the publication of Spatiotemporal datasets is k- anonymity. The proposed algorithm is based on signature generation method, which is used to generate a key for the users in a digital manner. In signature generation method we use ECC algorithm for digital signature. We improved clustering approach and propose the k-means clustering. By using k-anonymity the similar trajectories were grouped and removed the dissimilar ones. Splitting techniques are used to protect the data privacy. In this paper, the proposed method is used to enhance the MLA framework to preserve the users privacy publication of Spatiotemporal datasets. The MLA framework has three algorithms: preprocessing, clustering, signature is used for the purpose of efficiency and security. The process of anonymization is to cluster. The ECC algorithm generate a private key and public key for the public users to see the personal details of a particular person. MLA algorithms are applied on real-world GPS datasets following different times and domains. Here the information loss is inversely proportional to the security level. Privacy preserving technique utilizes high quality datasets. Methods used in this paper are distribution of data, preprocessing the datasets, data mining algorithms, data hiding, signature generation, privacy preservation. The final results show the utility of the dataset and anonymization on MLA framework.

II LITERATURE SURVEY

Data Privacy Through Optical K- Anonymization

It proposes a practical method for determining an optimal anonymization of a given dataset. The optimal anonymization perturbs the input data as little as is necessary to achieve anonymity. Several different cost metrics have been proposed, through most aim in one way or another to minimize the amount of information loss results the generalization and suppression operations that are applied to produce the transformed dataset. The ability to compute optimal anonymizations let us more definitively investigate impacts of various coding techniques and problem variation on anonymization quality. It also allows to use better quantify the effectiveness of other nonoptimal methods. Winkler has proposed using simulated annealing to attack the problem but provides no evidence of its efficacy. The more theoretical side, Meyerson and Williams have recently proposed an approximation algorithm of optimal anonymization.

Mondrian Multidimensional K- Anonymity

Attacks can be reduced by using k- anonymity. The objective of k-anonymization technique is to protect the privacy of the every individual. The subject to this constrains, it is important that the released the data remain as “useful” as possible. This paper is a new multidimensional recoding model and a greedy algorithm for k-anonymization, an approach with several important advantages: the greedy algorithm has more efficient that proposed optimal kanonymization algorithms for single dimensional models. The greedy algorithm has the time complexity of O(nlogn), were the optimal algorithms are in worst cases. Higher quality results were produced while using greedy multidimensional algorithm than by using optimal single dimensional algorithm.

Machanavajjhala Measured Anonymity by the l-diversity This paper proposed that uncertainty of linking QID with some particular sensitive values. Wang proposed to bond the confidence of inferring a particular sensitive value using one or more privacy templets specified the data provider. Wong proposed some generalization methods to simultaneously achieve k-anonymity and bond confidence. Xiao and Tao

limited the breach probability, which is similar to the motion the confidence, and allowed a flexible threshold for each individual. K-anonymization for data owned by multiple parties for considered.

T-closeness Privacy beyond K-anonymity and I-diversity While k-anonymity protect against the identity disclosure, it won't provide sufficient protection against attribute disclosure. The notion of I-diversity attempts to solve this problem by requiringthe equivalence class that has at least 1 well represented values for each sensitive attribute. We use the earth mover distance measure for our-closeness requirement; thishas advantage of taking into consideration the sematic closeness of attribute values.

III WORKING METHODOLOGY

But 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.

Generalization is currently one of the mainstream approaches for the anonymization of spatiotemporal trajectory datasets. The generalization technique is predicated on two interrelated mechanisms: clustering and alignment. Clustering aims at finding the best grouping of trajectories that minimizes a predefined cost function, and the alignment process aligns trajectories in each group. The notion of k-anonymity was adopted in [8] for anonymization of spatiotemporal datasets. The authors proved that the anonymization process is NP-hard and followed a heuristic approach to cluster the trajectories. The use of 'edit distance' metric for anonymization of spatiotemporal datasets was proposed in [9]. In this work, the authors target grouping the trajectories based on their similarity and choose a cluster head for each cluster to represent the cluster. Also, dummy trajectories were added to anonymize the datasets further. Yarovoy et al. [10] proposed to use Hilbert indexing for clustering trajectories. The authors in [5], [11] chose to avoid alignment by selecting trajectories with the highest similarity as representatives of clusters. Poulis et al. [12] investigated applying restriction on the amount of generalization that can be applied by proposing a user-defined utility metric. Takahashi et al. [13] proposed an approach termed as CMAO to anonymize the real-time publication of spatiotemporal trajectories. The proposed idea is based on generalizing each queried location point with k 1 other queried location by other users, and hence, achieving k-anonymity. The current state-of-art technique for applying generalization to spatiotemporal datasets is based on generalization hierarchy (DGH) trees. In essence, DGH can be seen as a

coding scheme to anonymize trajectories. We have categorized types of DGHs in the literature as:

Full-domain generalization: This technique emphasizes 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.

IV PRIVACY MODEL

Adversary Model In our work, we consider coordinates and the time of queries both to be quasi-identifiers, as they can be linked to other databases and compromise the privacy of users. We also assume that no uniquely identifiable information is released while publishing the dataset. However, the adversary may:

Already know about part of the released trajectory for an individual and attempt to identify the rest of the trajectory. For instance, the adversary is aware of the workplace of an individual and attempts to identify his or her home address.

Already know the whole trajectory that an individual has traveled but try to access other information released while publishing the dataset by identifying the user in the dataset. For instance, the published dataset may also include the type of services provided to users and if the adversary can identify a user by its trajectory, it can also know the services provided to that user.

To this end, our aim is to protect users against the adversary's attempt to access sensitive information that may compromise user privacy.

V RESULTS

In the above screen click on 'Upload Taxi Trajectory Dataset' button to upload dataset.

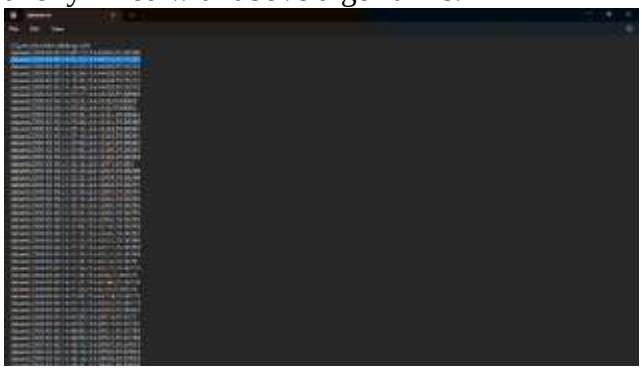
In above screen selecting and uploading taxi trajectory file and then click on 'Open' button to load data-set and to get below screen.

In above screen data set loaded and now click on 'Preprocess Dataset' button to remove empty values and then extract latitude and longitude location from above dataset.

In the above screen dataset preprocessing completed and now click on 'Run K-Means with Dynamic SA Algorithm' button to run KMEANS on datasets with Dynamic SA. This algorithm will group all similar locations into the same cluster and then perform DYNAMIC SA.

In above screen each location is processed and then calculating loss value with dynamic sequence alignment which align two locations by choosing minimal loss location.

In the above screen KMEANS loss is 0.09 and Heuristic Clustering (also known as Dynamic SA) loss is 0.62. Now click on 'Run Data Generalization Algorithm' button to generalize data with loss value. In below screen in first record, you can see real location values from dataset and in next screen same location was generalized or anonymized with above algorithms.



In below screen you can see the same location is generalized with other values.

In above screen you can all location values are generalized so no malicious users can understand correct location. Now click on 'Loss Comparison Graph' button to get below graph.

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.

VI 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.

VII FUTURE ENHANCEMENTS

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.

VIII 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.