ISSN: 2366-1313

A NOVEL MACHINE LEARNING APPROACH FOR PRIVACY PRESERVING LOCATION DATA PUBLISHING

R.SRINIVAS[1], P. VENKATA SAI KUMAR[2], S. SIVA SAI KRISHNA[3], SK. ALLAUDHIN BASHA[4], T. GANESH[5]

[1]Assistant Professor, CSE,Chalapathi Institute of Technology,Guntur, India
[2]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[3]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[4]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[5]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

ABSTRACT: Now-a-days due to mobile all online applications are recording user locations and then storing those in their apps and these location details can be used to track users. Sometimes some malicious users can track the user location to know where user is travelling to like bank, hospital or any other locations. To overcome this problem and to provide security to user location data many data anonymization techniques such as K-Anonymity and data perturbation are introduced, where Data perturbation will add noise to user data . And K-Anonymity will adjust user data into groups. But above techniques are not reliable because there is a chance of identifying noise data added user locations. To overcome this, there were three important techniques in Machine Learning named Clustering model, Dynamic Sequence Alignment and Data Generalization. Where these models will provide more security and generalize the data which cannot be easily understood to track.

1. INTRODUCTION

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 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 travelled by the users. Such side information enables them to reveal sensitive information that can cause physical, financial, and 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.

ISSN: 2366-1313

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 adopted the notion of k-anonymity for spatiotemporal datasets and proposed an anonymization algorithm based on generalization. Xu et al. investigated the effects of factors such as spatiotemporal resolution and the number of users released on the anonymization process. Dong et al. focused on improving the existing clustering approaches.

They proposed an anonymization scheme based on achieving k-anonymity by grouping similar trajectories and removing the highly dissimilar ones. More recently, the authors in developed an algorithm called k-merge to anonymize the trajectory datasets while preserving the privacy of users from probabilistic attacks. Local suppression and splitting techniques were also considered to protect privacy in However, there are three major problems with the above-mentioned approaches.

- 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.
- There is no unified metric to evaluate and compare the existing anonymization methods.

2. LITERATURE SURVEY

Machine Learning Aided Anonymization of Spatiotemporal Trajectory Datasets

The big data era requires a growing number of companies to publish their data publicly. Preserving the privacy of users while publishing these data has become a critical problem. One of the most sensitive sources of data is spatiotemporal trajectory datasets. Such datasets are extremely sensitive as users' personal information such as home address, workplace and shopping habits can be inferred from them. In this paper, we propose an approach for anonymization of spatiotemporal trajectory datasets. The proposed approach is based on generalization entailing alignment and clustering of trajectories. We propose to apply k'-means algorithm for clustering trajectories by developing a technique that makes it possible. We also significantly reduce the information loss during the alignment by incorporating multiple sequence alignment instead of pair wise sequence alignment used in the literature. We analyze the performance of our proposed approach by applying it to Geolife dataset, which includes GPS logs of over 180 users in Beijing, China. Our experiments indicate the robustness of our framework compared to prior works.

New Australian government data sharing and release legislation

Whether it is claiming a Medicare rebate, having Av passport checked before an overseas trip, lodging attacks return or simply looking at the weather forecast, every day millions of Australians rely on services delivered by the Australian Government. Australians expect government services to

ISSN: 2366-1313

be seamless, easy and fast just like their normal experience of shopping and banking. This fuels the need for the government to keep pace with the private sector—and aspire to be a market leader when it comes to delivering services for Australian people and businesses. The Morrison Government is committed to making it easier and faster for Australians to access the services they need by ensuring people and businesses are at the very heart of service design and delivery. As the Minister responsible for the National Disability Insurance Scheme and Government Services, I have been talking with people right across Australia about how even the smallest improvement to government services can have a big impact on people's lives. Improvements such as fewer questions on an aged care form, making it easier to report income online, or even a single, clear point of access such as an app on your phone can all have an immediate and lasting positive impact.

3. EXISTING SYSTEM

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.

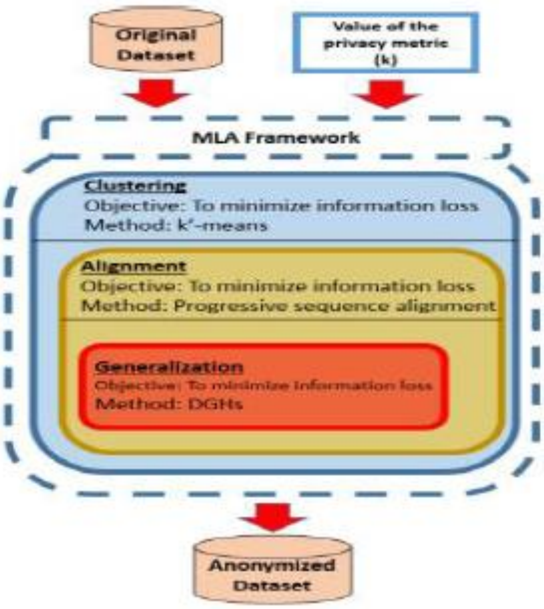
Disadvantages:

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.

4. PROPOSED SYSTEM:

Author has introduce Machine Learning based data privacy preserving technique which consists of three models and these three models will provide more security and anonymize or generalized which cannot be easily understand or crack.

5. ARCHITECTURE DIAGRAM


Original Dataset
Value of the privacy metric (k)
MLA Framework
Clustering
Objective: To minimize information loss
Method: k'-means
Alignment
Objective: To minimize information loss
Method: Progressive sequence alignment
Generalization
Objective: To minimize information loss
Method: DGHs
Anonymized Dataset

6. MODULES:

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**:

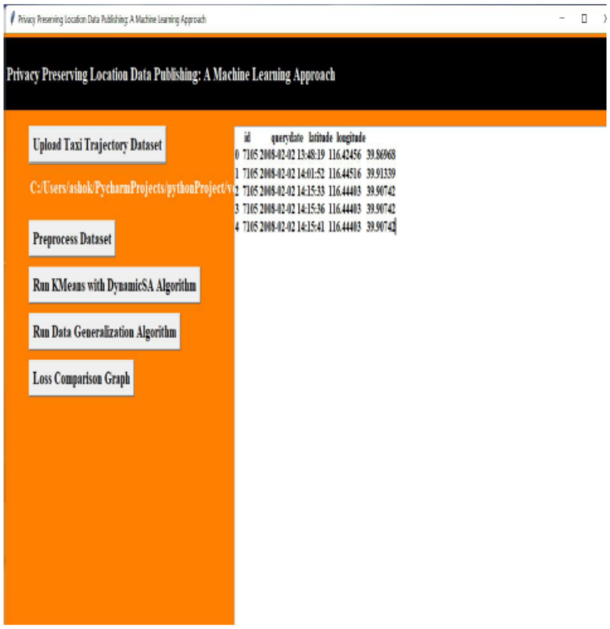
In this module or algorithm, we will take location form cluster member and then take random locations from original dataset and both these records will be aligned to get location which has minimal loss.

3) **Data Generalization:**

In this module user location will be generalized or anonymised by summing up location with loss values.

Advantages:

Loss value will be less so that algorithms work efficiently.


Privacy Preserving Location Data Publishing: A Machine Learning Approach
Privacy Preserving Location Data Publishing: A Machine Learning Approach
Upload Taxi Trajectory Dataset
Preprocess Dataset
Run KMeans with DynamicSA Algorithm
Run Data Generalization Algorithm
Loss Comparison Graph

```
C:/Users/ashok/PycharmProjects/pythonProject/v
```

```
   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
```

7.CONCLUSION & FUTURE SCOPE

we have proposed a framework to preserve the privacy of users while publishing the Taxi Trajectories datasets. The proposed approach is based on an efficient alignment technique termed as pairwise 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-means algorithm for guaranteeing the privacy of overly sensitive datasets. The experimental results on taxi trajectory datasets indicate the superior spatial utility performance of our proposed

framework compared with the previous works.

In future, this model can be used to compare various machine learning algorithm generated prediction models and the model which will give higher accuracy will be chosen as the prediction model. This project work can be extended to higher level by working with multiple datasets at a time with multiple attributes.

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.