IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 6.252

Volume 2, Issue 6, June 2022

Privacy Preservation of Location Data Publishing

Dr. M. L. Valarmathi[1], C. Devipriya[2], Saranya P[3], Banumathi S[4], Nathira Begum S[5]

Professor, Department of Computer Science and Engineering1

Assistant Professor, Department of Computer Science and Engineering2

UG Scholar, Department of Computer Science and Engineering3,4,5

Dr. Mahalingam College of Engineering and Technology, Pollachi, Tamil Nadu, India

Abstract: *Machine learning is important for future development and access for large detailed datasets. The privacy preserving machine learning enables maintaining of the data privacy and confidentiality. Machine learning enables new services in using sensitive data. This paper uses location trajectories and the application of this framework is the privacy preservation of location based data. Researchers had verified that publishing trajectories data would cause risk of user's privacy and also capable of identifying their locations, personal details and so on. Therefore, we have applied anonymization techniques and developed the data to preserve the privacy for the users. We propose a framework for Spatiotemporal datasets termed ML based anonymization (MLA). We use machine learning algorithms for clustering the dataset. To propose the trajectories we use k-means algorithm. The k-means is a type of clustering algorithm used in many real time applications, especially for analysis of data. Moreover, we improve alignment method for progressive sequence alignment of MLA. In this paper, we generate signature key for the public user and generate a digital signature for public users. Signature generation method use elliptic curve cryptography (ECC) algorithm. As a result on Spatiotemporal trajectory datasets indicate a high utility performance of ouranonymization based on MLA framework.*

Keywords: Machine learning, Location trajectories, Spatiotemporal dataset, k-means algorithm, Signature generation, Elliptic curve cryptography

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.

DOI: 10.48175/568

II. LITERATURE SURVEY

2.1 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.

2.2 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 k-anonymization 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.

2.3 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.

2.4 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. DATASET

In this paper we have used Spatiotemporal dataset which is based on trajectories. Thisdataset consist of person's user id, date & time, latitude, longitude. Spatiotemporal dataset is used for data analysis and the dataare collected in space and time. Forexample, moving objects. It indicates the location and frequency.

USERID	DATE &TIME	LATITUDE	LONGITUDE
1	2008-02-02 13:33:52	116.36422	139.88781
2	2008-02-03 12:22:46	116.48245	39.90531
3	2008-02-04 11:57:26	116.33809	39.89797

Table 1: Spatiotemporal dataset

IV. RESEARCH METHOLOGIES

4.1 Existing System

The existing system just swaps the user id when they reached the intersection. Adversaries were prevented by doing algorithms, so the particular user's identification was secured.

It made a distinction between sensitive andinsensitive location nodes of trajectories. Their algorithm only groups the paths around the sensitive nodes and exploits generalization to create super nodes.

"Protecting privacy in trajectories by using user-centric approach", it shifted the burden of privacy preservation in data publishing to the user side. This attempted to anonymize the data on the mobile phones before storage on the database as they would have more control over their privacy.

Disadvantages

Existing system is not sufficient for preserving the privacy of users. Adversariescan re-identify individuals in datasets basedon 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 reputational harms to people.

In the existing systems, the information lossis high because the use of pairwise sequence alignment.

V. IMPLEMENTATION

The proposed method is S-MLA, that enhanced anonymization (S-MLA) topreserve the privacy for the users in the publication of spatiotemporal trajectory dataset.

S-MLA consist of three interworking algorithms: digital signature generation, clustering and alignment.

5.1 Proposed System

- **Alignment:** By formulating theanonymization process is an optimization problem and finding an alternative representation of the system.
- **Clustering**: We are able to apply machine clustering algorithms for clustering trajectories. For this purpose, k-means clustering algorithm is applied.
- **Digital Signature Generation:** ECC(Elliptic Curve Cryptography) is used to generate the digital signature for users, to protect from network based attacks.

Advantages

It enhance the performance of sequencealignment clusters by considering the multiple sequence alignment instead of pairwise sequence alignment.

Networks based attacks is handled by usingdigital signature algorithms.

K-means algorithm for guaranteeing the k-anonymity in overly sensitive datasets.

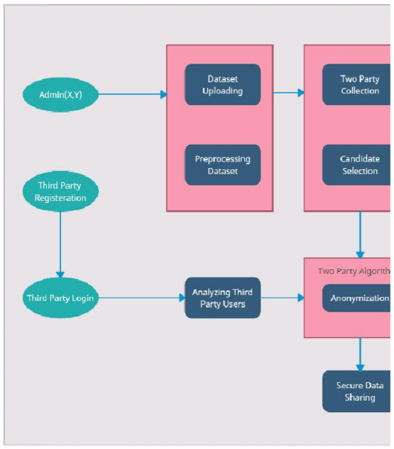


Figure 5.1.1 Proposed Model

5.2 Modules Admin Login

Admin is the main user of our system. Admin can only view all the data in the dataset. Admin should login to the system to ensure authentication. Admin can able to upload the dataset, preprocessing the dataset and do all other operations for the privacy preserving spatial data publishing.

Uploading Dataset

This module is used to upload the dataset. The admin can choose one file from the computer location and uploading in the screen, the file will be stored in one locationfor further accessing. The next step is preprocessing, after uploading the dataset.

Preprocessing

The preprocessing method is to find the missing values in the dataset. The real- world data has a lot of missing values. Missing value can be lead to data corruptionand failed to yield output. The handling of missing data is very important during the preprocessing, the dataset as many machinelearning algorithms do not support missingvalues.

Data Alignment

Anonymity is the process to secure the data,which done in this module data alignment. The purpose of alignment is to find two trajectories to minimize the overall cost of generalization and suppression. In this project, we adopt a multiple SA technique called progressive SA for anonymization ofspatiotemporal trajectories. The k- anonymity metric ensures that the users arek-anonymous, implying that they cannot beidentified from at least k-1 other users in theanonymized published dataset. To achieve k-anonymity, significant loss ofinformation incurs during the generalization process of different algorithms. However, there is no unified metric to measure how much information has actually been lost to achieve k- anonymity.

Clustering

Clustering can be search for hidden patternsthat may exist in datasets. Members of each cluster are very similar toeach other because of grouping data entries in disjointed clusters. Clustering method is applied inmany application areas, such as data analysis and pattern recognition. There are three clustering approaches area, such as data analysis and pattern recognition. Thereare three clustering approaches likeheuristic, k-means, iterative k-means. The two algorithms is our proposed approaches to significantly improve the utility of published spatiotemporal datasets. The heuristic algorithm are presented for thepurpose of comparison. Each one of these approaches works independently and can be embedded in the MLA framework to clustertrajectories.

K-Means Clustering

Grouping data based on similarity patterns based on distance is known as k-meansclustering. K-means clustering is simple and elegant approach for the dataset into k distance. It is a popular unsupervisedlearning algorithm. In this paper we use three groups of clusters: cluster 0, cluster 1,cluster 2. This algorithm takes the input dataset and divides into k-number ofclusters. K-values must be predefined in ofk center points. The goal of this algorithm is to find groups the data, with number of groups represented by the variable k. The clusters must be non-overlapping. The k- means algorithm tries to determine the k-setthat will make the square error the smallest.

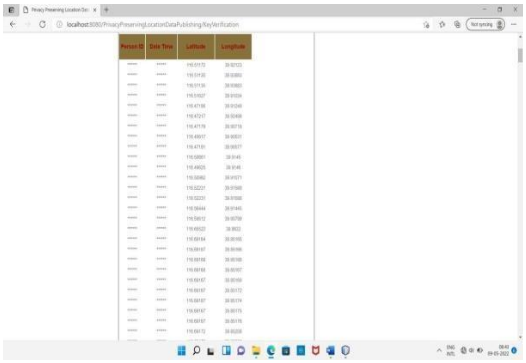
Signature Generation

In this module, a separate signature key isgenerated for each users. By this signature only, the user can access their data. ECC algorithm is used as module to generate a secret key signature. Secret key (SK) isused for generating the ECC(Elliptic Curve Cryptography) hashing algorithm. The signature generation key is to verifys the data. It uses a mathematical algorithms for validate the data and the messages. It is used in many applications for future purposes.

ECC Algorithm

Elliptic curve cryptography(ECC) is apublic key cryptography. It provide high security with small keys when compared tonon-ECC. ECC is based on algebraicstructure of elliptic curves. ECC are used for key generation, digital signatures,

pseudo-random generators etc. In ECC, the public key is equation of elliptic curve and point that lies on thatcurve. The private key is used for creating a signature digitally and also for calculating the digital signature algorithm. This typically involves taking a cryptographyhash of the data and operating on it mathematically using the private key. The public key can check the signature was created using the private key and the appropriate signature validation algorithm.
Identify information: The who owns thecertificate and which domains the certificate is valid for.
A public key: The public key pair, the site owner controls and keeps secret the associated private key.
The functionally provide same outcome for other digital signing algorithms, because ECDSA is based on the more efficient elliptic curve cryptography, ECDSA requires smaller keys to provide equivalent security and therefore it is more efficient. The ECDSA supports the algorithms basedon elliptic curves.



localhost:8080/PrivacyPreservingLocationDataPublishing/KeyVerification

Person ID	Date Time	Latitude	Longitude
*****	*****	116.51172	39.92123
*****	*****	116.51130	39.93883
*****	*****	116.51136	39.93883
*****	*****	116.51027	39.91034
*****	*****	116.47186	39.91246
*****	*****	116.47217	39.90498
*****	*****	116.47179	39.90718
*****	*****	116.45617	39.90631
*****	*****	116.47181	39.90577
*****	*****	116.50861	39.9145
*****	*****	116.49025	39.9146
*****	*****	116.50462	39.91571
*****	*****	116.52221	39.91560
*****	*****	116.52231	39.91588
*****	*****	116.56444	39.91445
*****	*****	116.58612	39.90709
*****	*****	116.66522	39.9022
*****	*****	116.69184	39.90165
*****	*****	116.69187	39.90166
*****	*****	116.69188	39.90168
*****	*****	116.69168	39.90167
*****	*****	116.69167	39.90169
*****	*****	116.69187	39.90172
*****	*****	116.69187	39.90174
*****	*****	116.69167	39.90175
*****	*****	116.69187	39.90176
*****	*****	116.69172	39.90208

Figure 5.2.1 Result

VI. CONCLUSION

Businesses and even governments collect data through many digital platforms like social media, E-health, e-commerce, entertainment, e-government and etc. Theyuse to serve their customers/citizens. The collected data can be sensitive and those data can be stored, analyzed and in good probability, anonymized and may shared with others. Data privacy has been called "The most important issue in the next decade". So, the privacy preservation is most important while publishing data especially publishing location data is the most important. Because if a Person's location isrevealed, then the adversaries may identifytheir location and misuse. We used machineclustering algorithms for clustering trajectories. We propose to use k-meansalgorithm for this purpose. Digital signaturegeneration: ECC(Elliptic CurveCryptography) is used to generate the digital signature for users, to protect from network based attacks. By using Machine learning clustering approach, in theanonymization process incurred loss is minimized. To protect from network-basedattacks, the signature key is generated by using ECC Algorithm. K-Means algorithmis used to 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]. R. Agrawal, A. Evfimievski, and R.Srikant. Information sharing across privatedatabases. In Proceedings of the ACM International Conference on Management of Data, 2003.

[2]. B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy,accuracy, and consistency too: A holistic solutiontocontingencytablerelease. In Proceedingsofthe ACM Symposium on Principles of Database Systems (PODS),2007.

[3]. R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In Proceedings of the IEEE International Conference on Data Engineering (ICDE), 2005.

[4]. R.Bhaskar,S.Laxman,A.Smith,andA.Thakurta. Discovering frequent patterns in sensitive data. In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2010.

[5]. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactivedatabase privacy. In Proceedings of the ACM Symposium on Theory of Computing(STOC), 2008.

[6]. J. Brickell and V. Shmatikov. Privacy- preserving classifier learning. In Proceedings of the International Conference on Financial Cryptography andData Security, 2009.

[7]. P. Bunn and R. Ostrovsky. Secure two-party k-means clustering. In Proceedings ofthe ACM Conference on Computer and Communications Security (CCS), 2007.

[8]. K. Chaudhuri, C. Monteleoni, and A. Sarwate. Differentially private empirical risk minimization. Journal of Machine Learning Research (JMLR), 12:1069–1109, July 2011.

[9]. K. Chaudhuri, A. D. Sarwate, and K. Sinha. Near-optimal differentially private principal components. In Proceedings ofthe Conference on Neural Information Processing Systems, 2012.

[10]. C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu. Tools for privacy preserving distributed data mining. ACM InternationalConferenceonKnowledgeDiscoveryandDataMining (SIGKDD) Explorations Newsletter, 4(2):28–34, December 2002.

[11]. Dinur and K. Nissim. Revealing information while preserving privacy. In Proceedings of the ACM Symposium on Principles of Database Systems (PODS),2003.

[12]. C. Dwork. A firm foundation for private data analysis. Communications of the ACM, 54(1):86–95, 2011.

[13]. C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In Proceedings of the 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT), 2006.

[14]. C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In Proceedings of the Theory of CryptographyConference (TCC), 2006.

[15]. Frank and A. Asuncion. UCI machine learning repository, 2010.

[16]. Friedman and A. Schuster. Data mining with differential privacy. InProceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2010.

[17]. B. C. M. Fung, K. Wang, R. Chen, and P.S.Yu. Privacypreserving data publishing: A survey of recent developments. ACMComputing Surveys, 42(4):1–53, June 2010.

[18]. B. C. M. Fung, K. Wang, and P. S. Yu. Anonymizing classification data for privacy preservation. IEEE Transaction on Knowledge and Data Engineering (TKDE),19(5):711–725, May 2007.

[19]. S. R. Ganta, S. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2008 Learning Research (JMLR), 12:1069–1109, July 2011.

DOI: 10.48175/568