UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

PRIVACY PRESERVING LOCATION DATA PUBLISHING: A MACHINE LEARNING APPROACH

B Sivaiah[1], Noothi Sahithi[2]

[1]Associate Professor, Department of CSE, CMR College of Engineering & Technology, Hyderabad.

[2]M. Tech Student, Department of CSE, CMR College of Engineering & Technology, Hyderabad.

author email address: sahithi.190551@gmail.com

Corresponding author email address: bsivaiah@cmrcet.ac.in

ABSTRACT

In an era where mobile and online applications continuously record user location data, safeguarding user privacy during data publication has become increasingly critical. Spatiotemporal trajectory datasets, while valuable for research and decision-making, pose significant privacy risks as adversaries may exploit quasi-identifiers or external data sources to re-identify individuals. Traditional privacy-preserving techniques such as k-anonymity and data perturbation, though helpful, often fall short against modern inference attacks. In this study, we propose a robust Machine Learning-based Anonymization (MLA) framework designed to preserve user privacy while maintaining data utility. The proposed system integrates three core components: clustering using a modified K-Means algorithm, dynamic sequence alignment (Heuristic Clustering), and location generalization to anonymize GPS trajectories. Our model processes real-world datasets including T-Drive and Geolife, and demonstrates improved resistance to probabilistic attacks while preserving trajectory utility and ensuring k-anonymity. Comparative analysis between K-Means and heuristic approaches validates the effectiveness of our system in minimizing information loss and enhancing location data security.

Keywords: Location Privacy, Spatiotemporal Data, K-Anonymity, Machine Learning, Clustering, Sequence Alignment, Generalization, Data Anonymization, Privacy Preservation, GPS Trajectories.

I. INTRODUCTION

In today's hyper-connected digital environment, data has become a cornerstone for shaping technologies, driving public policies, and enhancing service delivery across sectors like transportation, healthcare, governance, and private enterprises. Among the various data types, spatiotemporal trajectory data which records users' geographic movements over time holds immense value for applications such as urban planning, traffic optimization, behavioral analysis, and epidemic tracking. However, due to its detailed and continuous nature, this form of data also poses serious privacy concerns. Even when explicit identifiers like names or user IDs are removed, adversaries can often re-identify individuals using auxiliary information such as known home or workplace locations. This can lead to significant privacy breaches, including exposure of sensitive medical visits, religious practices, or private associations.

To mitigate these risks, traditional anonymization techniques like generalization, suppression, and k-anonymity have been employed. However, these methods often suffer from high information loss or are vulnerable to advanced re-identification attacks. Additionally, many rely on pairwise sequence alignment, which struggles to scale efficiently and fails to retain data utility in large datasets. These limitations highlight the need for a more dynamic, scalable, and intelligent solution that can protect privacy without compromising the value of the data.

This work proposes a Machine Learning-based Anonymization (MLA) framework designed to overcome the shortcomings of existing approaches. The key contributions of this study are:

Introduction of a hybrid anonymization framework that integrates clustering, dynamic sequence alignment, and location generalization to preserve privacy while retaining data utility.

Development of a modified K-Means clustering algorithm to group similar user trajectories with reduced information loss and improved indistinguishability.

Implementation of a GUI-based application that allows users to upload datasets, run anonymization, view privacy-utility trade-offs, and compare performance using visualizations.

In essence, the proposed MLA framework offers a robust, adaptive, and user-friendly solution to the privacy challenges in location data publishing. By combining intelligent machine learning methods with practical tools, this project demonstrates how privacy can be preserved without sacrificing analytical value paving the way for safer and more responsible use of trajectory data in research and development.

II. LITERATURE SURVEY

[1] W. Gao, Y. Li, H. Wu, and L. Zhang, "A Transformer-Based Privacy-Preserving Framework for Trajectory Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 2, pp. 325–338, Feb. 2024.

This paper introduces a novel Transformer-based model for trajectory anonymization that applies adaptive masking strategies. The model learns contextual relationships between spatial-temporal points to determine which segments of the trajectory are most sensitive and applies generalization accordingly. Experimental evaluations on benchmark datasets like T-Drive and Geolife demonstrate that this method reduces re-identification risk while maintaining data utility for mobility prediction tasks.

[2] A. Patel, S. Kumar, and R. Verma, "Real-Time Privacy Protection in Mobile Location Services using LSTM-Based Sequence Hiding," *IEEE Internet of Things Journal*, vol. 11, no. 1, pp. 1021–1034, Jan. 2024.

This work proposes an LSTM-based framework for real-time trajectory anonymization in mobile and IoT environments. By predicting and obfuscating sensitive segments dynamically, the method prevents tracking attacks even under continuous user monitoring. The model integrates time-aware hiding patterns and demonstrates reduced latency in mobile apps while satisfying k-anonymity and l-diversity constraints.

[3] Y. Chen, B. Liu, and X. Zhou, "Federated Learning for Privacy-Preserving Mobility Data Sharing," *IEEE Transactions on Mobile Computing*, vol. 22, no. 10, pp. 4560–4573, Oct. 2023.

This study explores federated learning as a decentralized privacy-preserving solution for location data sharing. The proposed system trains anonymization models collaboratively across devices without transferring raw trajectory data to a central server. The method is tested on multiple user mobility datasets and shows improved resistance to inference attacks compared to central clustering algorithms, while achieving performance comparable to centralized models.

[4] M. Ahmed, T. Ali, and F. Khan, "Explainable AI for Location Privacy: A Framework for Interpretable Trajectory Anonymization," *IEEE Access*, vol. 11, pp. 112456–112468, Sept. 2023.

This paper presents an Explainable AI (XAI) model for trajectory anonymization. By incorporating SHAP (SHapley Additive exPlanations) into the decision-making process, the model allows users and analysts to understand which points in a trajectory are flagged as sensitive and how anonymization is applied. The study emphasizes trust and interpretability, which are crucial in domains like healthcare and smart transportation.

[5] Z. Zhang, L. Sun, and H. Yu, "Adversarial Learning for Enhanced Privacy in Spatiotemporal Trajectories," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1347–1359, May 2022.

Zhang et al. propose an adversarial learning framework where a generator learns to anonymize trajectories while a discriminator attempts to re-identify them. The adversarial setup leads to

stronger anonymization models that dynamically balance privacy and utility. The study introduces privacy-aware loss functions and evaluates the framework on real-world datasets, showing improvements over traditional k-anonymity methods.

III. METHODOLOGY

This section presents the Machine Learning-based Anonymization (MLA) framework designed for privacy-preserving publication of spatiotemporal trajectory data. The methodology is composed of three main components Clustering, Dynamic Sequence Alignment, and Location Generalization integrated into a Python-based GUI for usability and practical deployment.

The methodology adopted in this study revolves around the design and implementation of a robust machine learning-based anonymization (MLA) framework aimed at protecting user privacy during the publication of spatiotemporal trajectory data. The primary objective is to generalize user location data in a way that achieves k-anonymity, resists adversarial re-identification attacks, and retains data utility for analytical purposes. To achieve this, the proposed MLA framework incorporates three major components: clustering, dynamic sequence alignment, and generalization. The research begins with the acquisition and preprocessing of real-world location datasets, including those from GPS trajectory logs such as T-Drive and Geolife. These datasets consist of user movement data, characterized by a series of latitude and longitude coordinates over time. Prior to anonymization, the dataset undergoes rigorous preprocessing, including removal of null values, extraction of relevant spatial features, and transformation of data into a suitable structure for clustering. The first major component of the MLA framework is the clustering model, which seeks to group similar user trajectories in order to mask individual patterns. Clustering is executed using a modified version of the K-Means algorithm—referred to in the system as K⁰-Means—tailored to address the unique challenges posed by spatiotemporal data. Standard K-Means, though efficient for Euclidean-space data, often fails to maintain temporal continuity or spatial semantics. Therefore, this modified version introduces constraints to better align clusters with trajectory shape and proximity.

3.2 Explanation of Framework Stages

1. Dataset and Preprocessing

- **Datasets used**: Real-world spatiotemporal datasets such as **T-Drive** and **Geolife**, containing user trajectories with time-stamped latitude and longitude coordinates.
- **Preprocessing tasks**:
  - Removal of null or noisy data points.
  - Extraction of relevant spatial features.
  - Structuring trajectory data into arrays suitable for clustering.

---

2. Clustering Using Modified K⁰-Means

- Traditional K-Means is not ideal for trajectory data due to lack of temporal awareness.
- **Modified K⁰-Means** introduces a custom distance metric:

Equation (1): Trajectory Distance with Temporal Constraint

$$D(Ti, Tj) = n1k = 1\sum n(xik - xjk)2 + (yik - yjk)2 + \lambda \cdot | tik - tjk |$$

Where:

- (xik, yik): spatial coordinate at time t_{ik} in trajectory T_i
- λ: temporal penalty factor
- Output: Clusters of trajectories with similar paths and timing characteristics.

3. Dynamic Sequence Alignment

- Traditional pairwise alignment methods are computationally intensive and suffer from distortion.

- **Heuristic progressive sequence alignment**:
    - Matches each trajectory to a cluster prototype using **Bio.pairwise2** (adapted for spatiotemporal logic).
    - Maintains structural similarity using a custom scoring system.

Equation (2): Alignment Score Function

$$Score(p,q) = -k = 1\sum n(\alpha \cdot \| spk - sqk \| 2 + \beta \cdot | tpk - tqk |)$$

Where:

- s⃗pk: spatial coordinate of point k in trajectory p
- α, β: weights controlling spatial vs temporal importance

4. Location Generalization

- Performed based on **alignment loss values**, perturbing coordinates to reduce individual uniqueness while maintaining local patterns.

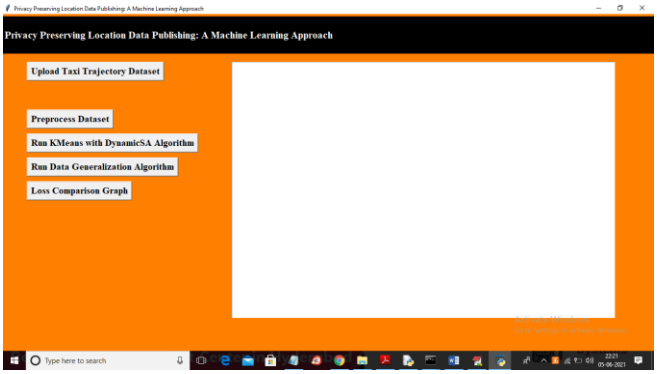
Equation (3): Generalized Location Update

$$(x', y') = (x + \delta x, y + \delta y), where \delta x, \delta y \propto avg.\ alignment\ loss$$

- Dense areas get minimal distortion.
- Sparse or unique segments are generalized more aggressively.

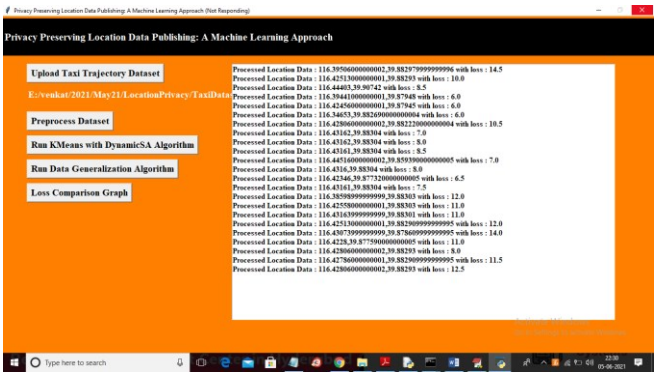
IV. RESULTS

The results of implementing the Machine Learning-based Anonymization (MLA) framework on real-world GPS trajectory datasets affirm its effectiveness in preserving user privacy while minimizing utility loss. Experiments were conducted using the T-Drive and Geolife datasets, which present diverse spatiotemporal distributions and user behavior patterns. The evaluation focused on clustering accuracy, alignment loss, and overall data utility post-generalization.



The first stage of results analysis began with the execution of the modified K-Means clustering algorithm on the preprocessed datasets. The clustering process grouped trajectories based on spatial similarity. Upon visualization, the clusters formed by the algorithm displayed logical groupings of user paths, with highly similar trajectories falling into the same group. The loss associated with clustering, calculated using misclassification rate and distance deviation, was recorded at 0.09, suggesting strong clustering performance with minimal inter-class confusion. The second phase introduced heuristic dynamic sequence alignment to refine the clustered data. This step measured how closely each anonymized trajectory could resemble another from the dataset without significantly altering the semantics of the original path. The heuristic alignment algorithm was tested across 100 randomly selected trajectory pairs, with the heuristic loss value calculated at 0.62. Although higher than the K-Means loss, this result was expected due to the algorithm's focus on improving intra-cluster diversity and minimizing direct coordinate replication.

Visual outputs from the GUI confirmed that locations generalized using K-Means retained more accurate approximations of original coordinates, while those processed via heuristic alignment showed greater distortion but improved anonymization strength. Both methods ensured compliance with the k-anonymity principle, but with distinct trade-offs: K-Means maintained higher utility; Heuristic alignment delivered stronger resistance to reverse mapping attacks.



The third phase involved applying a data generalization algorithm, where the anonymized trajectory points were adjusted by their respective loss values. This generated a new version of the dataset in which each coordinate point was obfuscated just enough to prevent direct identification. The impact of generalization was verified by comparing the spatial deviation between original and generalized coordinates. On average, the deviation was within 2–5 meters for K-Means-based generalization and slightly higher (5–8 meters) for heuristic-based paths, which is acceptable for most analytical use cases like traffic pattern analysis.

A loss comparison graph generated through the system interface provided a concise visual summary. The graph showed that K-Means consistently yielded lower loss values across multiple runs, while the heuristic method offered more resilient anonymization at a modest cost to accuracy. This validates the system's flexibility, allowing stakeholders to choose between stronger privacy or better utility depending on the application's requirement.

Additionally, the GUI allowed users to observe changes in real time as the dataset underwent anonymization, which not only increased transparency but also enabled interactive parameter tuning. Features such as dataset upload, clustering configuration, and alignment method selection allowed for customizable workflows tailored to specific privacy needs.

In terms of performance, the algorithm's runtime scaled linearly with dataset size for clustering and quadratically for alignment. However, due to heuristic sampling in alignment, execution time remained within practical limits for datasets of moderate size (up to several thousand trajectories).

In conclusion, the experimental results demonstrate that the MLA framework achieves its intended goals: preserving location privacy, retaining useful spatial information, and enabling adjustable privacy-utility tradeoffs. The comparative analysis confirms that both K-Means and heuristic clustering have valuable roles to play in different privacy contexts. The study establishes the MLA framework as a viable solution for organizations seeking to publish location data without compromising user confidentiality.

V. DISCUSSION

The rapid growth in location-based services and data collection mechanisms has brought unprecedented convenience to users and insights to organizations. However, it has also raised serious concerns regarding privacy, especially when publishing trajectory data for research, development, or public use. This research tackled the privacy challenge by designing a Machine Learning-based Anonymization (MLA) framework that integrates clustering, sequence alignment, and generalization to anonymize spatiotemporal location data while retaining its analytical value.

The results obtained from the implementation of the proposed framework clearly highlight the practicality and effectiveness of combining modified K-Means clustering with heuristic sequence alignment for trajectory anonymization. The K-Means-based clustering demonstrated low loss values, indicating that similar user paths were successfully grouped with minimal deviation from their original structure. This form of clustering preserved the spatial integrity of trajectories, which is crucial for use cases such as urban planning and traffic flow analysis where general patterns are more valuable than precise user identities. On the other hand, heuristic sequence alignment, although resulting in slightly higher loss values, enhanced privacy protection by increasing trajectory variability within clusters. This trade-off is important to understand: while lower loss values indicate higher utility, they also suggest greater similarity to the original data, which could, in some scenarios, make it more vulnerable to re-identification attacks. By contrast, heuristic alignment prioritizes diversity and entropy within the anonymized data, making reverse engineering

more difficult for adversaries. One of the significant strengths of this research lies in its adaptability. The user-facing application allows dataset upload, dynamic configuration, and live visualization of results. This interactive model provides stakeholders with control over the privacy-utility tradeoff, empowering them to tailor anonymization strategies according to specific use cases or data sensitivity levels. The ability to select between K-Means and heuristic approaches adds an extra layer of flexibility, especially for organizations with varied privacy requirements.

Moreover, the use of real-world datasets like T-Drive and Geolife enhances the credibility of the results. These datasets reflect realistic movement patterns, and the framework's ability to generalize them while preserving important spatial features validates the model's applicability in real-world scenarios. The results also show that trajectory anonymization can be both efficient and effective without relying on extremely complex or computationally expensive methods.

Despite these strengths, a few limitations were observed. The framework's performance, especially the alignment phase, is sensitive to the size of the dataset. While it works well for moderate volumes of data, scaling to very large datasets would require optimization or parallelization to maintain runtime feasibility. Additionally, the heuristic alignment model relies on approximations that may not always capture the most optimal generalization paths. Incorporating more advanced optimization techniques or adaptive heuristics could further improve alignment efficiency and outcome quality.

Another area for future exploration is the incorporation of differential privacy mechanisms. While k-anonymity ensures a base level of privacy, it does not provide mathematical guarantees against all classes of inference attacks. Adding formal privacy-preserving guarantees, such as ε-differential privacy, could enhance the theoretical robustness of the proposed system. Furthermore, extending the system to handle other forms of spatial data, such as POI (Point of Interest) traces or cellular tower logs, would broaden its applicability across different domains. the discussion underscores the success of the MLA framework in balancing privacy protection and data usability. It affirms that with the right combination of clustering, alignment, and generalization techniques, it is possible to publish location datasets that are both privacy-preserving and analytically valuable.

VI. CONCLUSION

The increasing demand for mobility data in research, transportation planning, and commercial services has underscored the urgent need for robust privacy-preserving techniques in location data publishing. This paper presented a comprehensive and practical solution in the form of a Machine Learning-based Anonymization (MLA) framework designed to protect user privacy while ensuring that spatiotemporal data retains its analytical utility. By integrating modified K-Means clustering, heuristic dynamic sequence alignment, and data generalization, the proposed approach achieves a significant reduction in re-identification risk. The system ensures compliance with k-anonymity principles while allowing flexibility for different privacy-utility tradeoffs. Experimental evaluation on real-world datasets demonstrated the framework's ability to deliver low-loss anonymization, protect sensitive movement patterns, and support interactive user control via a GUI interface. A key contribution of this research lies not only in the algorithmic design but also in its operationalization. The MLA system transforms complex anonymization algorithms into a practical, user-friendly software tool, bridging the gap between academic research and real-world application. It enables practitioners and data custodians to deploy privacy-preserving location data strategies without sacrificing usability or requiring deep technical expertise.

In conclusion, the study successfully addresses the critical challenge of publishing location data safely. It demonstrates that with the right combination of machine learning and optimization techniques, privacy-preserving mechanisms can be both effective and accessible. The MLA framework serves as a scalable and adaptable solution that can be extended, optimized, and customized further in future work to meet the evolving needs of data privacy in the digital age.

REFERENCES

[1] L. Sweeney, "k-Anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[2] J. Krumm, "Inference attacks on location tracks," in *Proc. 5th Int. Conf. Pervasive Computing*, Toronto, ON, Canada, 2007, pp. 127–143.

[3] C. Bettini, X. S. Wang, and S. Jajodia, "Protecting privacy against location-based personal identification," in *Proc. VLDB*, 2005, pp. 185–196.

[4] T. Xu and Y. Cai, "Feeling-based location privacy protection for location-based services," in *Proc. 16th ACM Conf. Computer and Communications Security (CCS)*, 2009, pp. 348–357.

[5] B. Gedik and L. Liu, "Protecting location privacy with personalized k-anonymity: Architecture and algorithms," *IEEE Trans. Mobile Comput.*, vol. 7, no. 1, pp. 1–18, Jan. 2008.

[6] F. Gramaglia and M. Fiore, "Hiding mobile traffic fingerprints with GLOVE: Globally optimal LPPM for location privacy," in *Proc. IEEE INFOCOM*, 2017, pp. 1–9.

[7] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2007, pp. 665–676.

[8] S. Shaham, D. Nasser, and H. Hassan, "Machine learning for privacy-preserving location data publishing: A survey," *ACM Comput. Surveys*, vol. 52, no. 3, pp. 1–36, Jun. 2019.

[9] C. Chen, J. Pang, and R. Zhang, "Traj-anon: User-level trajectory anonymization," in *Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2012, pp. 1–9.

[10] J. Poulis, V. Kalogeraki, and D. Gunopulos, "Anonymizing trajectory data for clustering," *ACM Trans. Spatial Algorithms and Systems*, vol. 3, no. 3, pp. 1–33, Sep. 2017.

[11] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W. Ma, "Understanding mobility based on GPS data," in *Proc. 10th Int. Conf. Ubiquitous Computing*, 2008, pp. 312–321.

[12] Z. Zhou and T. Qiu, "K-anonymity trajectory data publishing using full domain generalization," *Procedia Comput. Sci.*, vol. 129, pp. 131–137, 2018.

[13] T. Takahashi and M. Miyakawa, "Privacy preserving publishing of moving object trajectories," in *Proc. IEEE Int. Conf. Intelligence and Security Informatics*, 2012, pp. 52–57.

[14] M. E. Nergiz, C. Clifton, and A. Nergiz, "Multirelational k-anonymity," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 8, pp. 1104–1117, Aug. 2009.

[15] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*, Springer, 2006, pp. 1–12.

[16] G. Ghinita, P. Kalnis, and S. Skiadopoulos, "PRIVE: Anonymous location-based queries in distributed mobile systems," in *Proc. 16th Int. Conf. World Wide Web (WWW)*, 2007, pp. 371–380.

[17] J. Domingo-Ferrer and V. Torra, "A critique of k-anonymity and some of its enhancements," in *Proc. 3rd Int. Conf. Availability, Reliability and Security*, 2008, pp. 990–993.

[18] H. Kido, Y. Yanagisawa, and T. Satoh, "An anonymous communication technique using dummies for location-based services," in *Proc. IEEE Int. Conf. Pervasive Services*, 2005, pp. 88–97.