Design and Application of Privacy Protection Case Studies for Trajectory Traffic Data Publishing

Luyang Zhao, Zixi Zang

***Abstract*— With the rapid development of location-based smart services and vehicular network technologies, the collection and application of traffic trajectory data have demonstrated significant potential in fields such as intelligent driving, traffic management, and data analysis. The extensive utilization of user trajectory information within vehicular networks has highlighted data security and privacy protection as critical issues needing urgent resolution. Particularly, against the backdrop of integrated development of vehicle positioning technologies and self-organizing networks, the research focus has shifted to how traffic trajectory information privacy can be effectively safeguarded without compromising the accuracy of location services. This paper delves into how personalized privacy protection for traffic trajectories can be achieved within location-based service platforms, ensuring data usability. It provides fundamental theories, technical approaches, and evaluation methods to quantify the privacy needs of traffic trajectory data, thereby safeguarding personalized trajectory privacy. Using the PORTO-TAXI dataset as a case study, this research explores the design and application methods for trajectory privacy protection, offering innovative solutions to the field of traffic data privacy protection.**

***Index Terms*— Traffic trajectory; Differential privacy; Personalized privacy protection; Privacy budget**

I. INTRODUCTION

In the era of smart cities and digital transformation, traffic big data has become indispensable for enhancing urban management and transportation system efficiency. Sourced from diverse origins, including vehicular sensors, GPS, traffic monitoring systems, and user-generated location data, it forms an extensive network for the comprehensive analysis and monitoring of urban traffic conditions. The rapid growth of information technology, especially in intelligent transportation systems and Location-Based Services (LBS), has led to the generation and application of vast amounts of traffic trajectory data in areas like traffic management, smart navigation, and urban planning, raising significant concerns about user privacy. To address these issues, differential privacy mechanisms, developed by researchers like Dwork[1], have become pivotal in safeguarding privacy, leading to groundbreaking studies in the field.

In the digital age, the explosion of intelligent services like vehicle Internet and Location-Based Services (LBS) has generated vast amounts of traffic trajectory data, which pose significant processing and protection challenges. This case study introduces key differential privacy techniques that safeguard location privacy while ensuring data usability. Initially, Sarathy et al.[2] developed a method using the Gaussian mechanism to add noise to location data, effectively preventing data breaches by distorting the original geographic information. Building on this foundation, Chen et al.[3] enhanced privacy protection by simplifying location data publication through noise analysis, ensuring data accuracy is maintained. Andres and his team [4] introduced geo-indistinguishability to prevent malicious actors from inferring real locations using existing knowledge, providing robust data protection. Concurrently, Wang et al. [5] integrated differential privacy principles with trajectory publishing algorithms, enhancing both data security and usability.

However, traditional methods often overlook personalized privacy needs at various locations. To address this, Mahdavifar and Abadi [6][7][8], proposed a personalized differential privacy approach, adjusting privacy settings based on user-specific trajectory data to tailor privacy protection more closely to real-world applications. These advancements offer robust support for secure data processing in smart cities and intelligent transportation systems, ensuring personal privacy is protected while leveraging big data for urban and traffic management. This case study underscores the significant role of differential privacy in modern information society, balancing technological progress with privacy protection.

II. APPLICATION DISCUSSION OF THE CASE

This case study is structured into several fundamental steps: It first outlines the data sources and application context for the traffic trajectory privacy protection scenario. It then details the conceptual design, overall framework, and specific algorithmic description of the proposed solution. Finally, the study conducts experimental designs to demonstrate the performance and advantages of the solution. This case has elicited positive responses in discussions of its application within the context of traffic trajectory analysis.

III. DESIGN PHILOSOPHY OF THE CASE

The primary design concept of this case study is to quantify the privacy demands of trajectory data to manage privacy budgets at user location levels effectively, while applying differential privacy—a standard technique for data privacy—and Markov chain models to generate high-quality synthetic trajectories. The system inputs the original trajectory dataset and outputs a protected trajectory dataset ready for publication, with Figure 1 illustrating the overall process of this method.

We utilize the PORTO-TAXI dataset[9] for analyzing traffic trajectory data and studying personalized differential privacy

Manuscript received April 23, 2024

Luyang Zhao (School of Computer Science and Technology, TianGong University, TianJin, China)
Zixi Zang (School of Software, TianGong University, TianJin, China)

protection. This dataset provides a wealth of experimental data for researching urban traffic flow, optimizing traffic management, and personalizing privacy protection. In implementing personalized differential privacy protection for traffic trajectory data, especially when handling the trajectory data of taxi drivers in the PORTO-TAXI dataset, assessing privacy needs is the initial step. The main goal is to identify the privacy sensitivity of individual users at different trajectory points. The privacy sensitivity of trajectory locations relates to multiple factors; next, we describe in detail the process of modeling users' privacy needs based on trajectory location characteristics and the algorithmic steps for implementing personalized privacy protection using the example of assessing privacy demands based on location features.

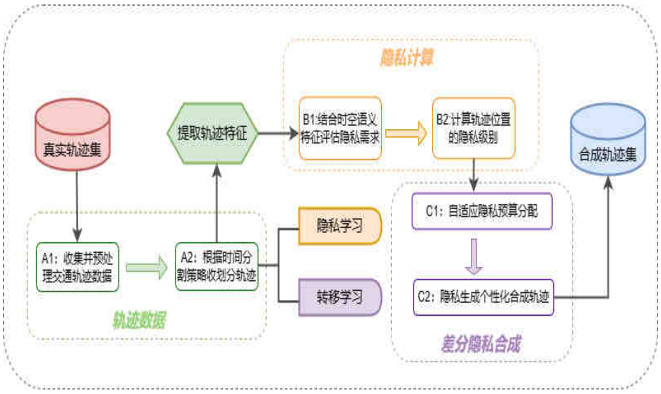
隐私计算
真实轨迹集
提取轨迹特征
B1:结合时空语义特征评估隐私需求
B2:计算轨迹位置的隐私级别
合成轨迹集
C1：自适应隐私预算分配
隐私学习
A1：收集并预处理交通轨迹数据
A2：根据时间分割策略划分轨迹
转移学习
C2：隐私生成个性化合成轨迹
轨迹数据
差分隐私合成

Fig. 1 Overall framework of the proposed method

IV. Algorithm Description of the Case

A. Privacy Needs Assessment

In our case study assessing the privacy needs of trajectory data within the PORTO-TAXI based on location characteristics, the core idea of our method revolves around the understanding that the trajectories formed by taxi drivers while carrying out their daily tasks not only reveal their travel paths but also contain significant private information. For instance, drivers might frequently stop at specific locations such as airports, train stations, or residential areas, which could be directly associated with their work routines and personal life. Moreover, frequently traveled routes might disclose the drivers' regular commuting patterns and preferences. Drawing on these observations, we identify stops and frequently passed points as critical location features for assessing privacy needs.

Upon the foundation of privacy needs assessment, we formalized the concept of privacy levels. We first defined two critical parameters: the stay duration threshold ($T_{threshold}$) and the frequent passage count threshold ($F_{threshold}$). $T_{threshold}$ is the minimum duration for which a taxi's stay at a location is considered purposeful, thus more privacy-sensitive. $F_{threshold}$ is the minimum number of times a taxi passes a location within a certain period, with counts above this threshold indicating routine pathways and preferences.For each location i its privacy level (P_i) is calculated as:

$$P_i = w_t \cdot T_i + w_f \cdot F_i$$

where T_i is the ratio of the stay duration at location i to $T_{threshold}$ (T_i=0 if the threshold is not exceeded), and F_i is the ratio of pass counts at location i to $F_{threshold}$ (F_i=0 if the threshold is not exceeded). Here, w_t and w_f are the weight coefficients for stay duration and pass counts in the privacy level computation, reflecting their importance in assessing privacy sensitivity.

B. Privacy Budget Allocation

With this formula, each location in the PORTO-TAXI dataset can be assigned a specific privacy level, indicating its privacy sensitivity. Locations with higher privacy levels require stricter privacy measures in data processing and publishing. The privacy budget allocation strategy assigns a privacy budget to each location point of a trajectory based on its privacy level, establishing a mapping relationship between privacy level and budget. The privacy budget (B_i) for each location i is calculated as:

$$B_i = \frac{\epsilon \cdot P_i}{\sum_{j=1}^{n} P_j}$$

where ϵ represents the total privacy budget, a crucial parameter in differential privacy controlling the overall risk of privacy leakage of the dataset, and n is the total number of locations in the dataset requiring privacy protection. This method ensures that the privacy budget is allocated proportionally to the privacy level, directing more privacy resources to locations with higher sensitivity and achieving dynamic and personalized privacy management across the dataset.

V. Experimental Analysis of the Case

A. Experimental Setup and Dataset

The sample algorithm presented in this case study is implemented in Python and executed on an AMD EPYC 7371 server equipped with 28GB of memory and running Ubuntu 20.04 LTS. Each experimental configuration was repeated five times, with the average results reported. For the experiments, 2,000 trajectory records from the PORTO-TAXI dataset were extracted, which included the coordinates (longitude and latitude), timestamps, and identifiers of each taxi. The collected traffic trajectory data underwent a two-step preprocessing procedure. First, noise and outlier values were removed to address anomalies caused by GPS signal loss or errors. Second, the data format was standardized to ensure that all trajectory data adhered to a uniform temporal and spatial resolution. This preprocessing improves the quality of the traffic data, facilitating more accurate analysis and enhanced protection of user privacy.

B. Evaluation Methods and Metrics

Privacy technologies for trajectory data primarily assess effectiveness through two dimensions: privacy protection and data utility. To effectively evaluate the design and performance of privacy protection algorithms, appropriate evaluation methods and metrics should be chosen based on the evaluation goals. It is also crucial to consider factors such as execution efficiency, spatiotemporal costs, application requirements, and service quality to conduct a comprehensive analysis and evaluation of trajectory privacy protection algorithms. The degree of privacy protection is usually measured by the ability to resist attack models or by quantifying privacy loss. When evaluating the privacy protection level of an algorithm, it is essential to select suitable attack models and corresponding privacy loss metrics. Under the same level of privacy protection, higher service quality received by the user indicates better privacy protection technology. The service quality of data reflects the utility of the published dataset, typically measured by the discrepancy between the published data and the original data.

Evaluation Perspective	Evaluation Significance	Evaluation Indicators
Privacy Protection	Degree of Privacy Disclosure	Location Privacy Measurement
		Query Privacy Measurement
Service Quality	Quality of Service Results	Data Quality Loss
		Data Utility Measurement
Overhead	Cost of Privacy Protection	Storage Cost
		Computational Cost

Tab. 1 The Evaluation Analysis of Trajectory Privacy Protection Techniques

Taking the PORTO-TAXI dataset as an example, three metrics can be used to measure the similarity between the original dataset D_{raw} and the synthesized dataset D_{syn}: visit frequency, transition pattern, and dwell time. The visit frequency reveals the frequently traversed hotspots or routes by taxis, reflecting the mobility and congestion patterns of the city; the transition pattern provides traffic flow information from one area to another, crucial for urban planning and traffic management; dwell time indicates the average waiting time of taxis at certain peak activity areas or congested spots. These metrics not only aid in assessing the resemblance of the synthetic data to the actual taxi movements but also facilitate a deeper understanding of the urban traffic network dynamics. Through these three metrics, we can comprehensively evaluate the similarity between D_{raw} and D_{syn} , thereby quantifying the utility of the synthetic data.

C. Experimental Results and Analysis

To explore the efficacy of the personalized privacy protection method for traffic trajectory data proposed in the case study, an experimental design and analysis were conducted based on the example algorithm outlined in Section 4. This algorithm primarily uses location characteristics to determine the privacy budget. The experiments employed metrics discussed in Section 5.2 to quantitatively compare the synthetic trajectory sets generated under three different privacy allocation strategies: NP (No Privacy, where no noise is added during the synthesis process), FPD (Fixed Privacy Budget, where a constant privacy budget is used at each step of trajectory synthesis), and PDP (Privacy Budget by Degree, where a graded privacy budget is applied at each synthesis step). The experiments demonstrated the utility of the synthesized trajectory data across varying base privacy budgets, ranging from $\varepsilon = 0.2$ to $\varepsilon = 2.0$, in increments of 0.2.

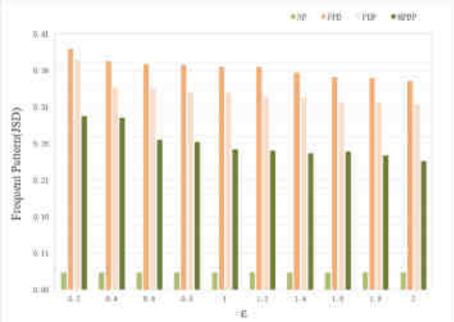

Fig. 2 VF Utility under Different Privacy Allocation Strategies

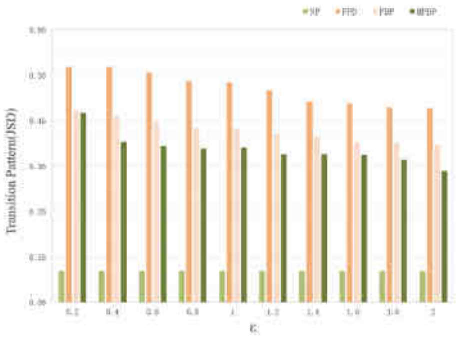

Fig. 3 TP Utility under Different Privacy Allocation Strategies

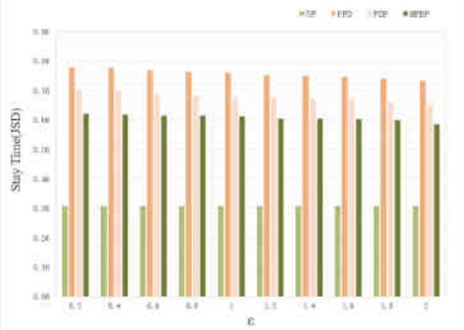

Fig. 4 DT Utility under Different Privacy Allocation Strategies

The NP algorithm serves as a baseline in this study, relying solely on a Markov model to predict and generate trajectories without incorporating any noise, thus offering the lowest level of privacy protection but setting a performance benchmark for other algorithms. The MPDP-generated synthetic trajectories show the highest resemblance to the original dataset, demonstrating superior capabilities in maintaining data utility. This indicates that, compared to the FDP algorithm, which simply applies a fixed privacy budget, or the conventional PDP algorithm, which is based on location features, MPDP enhances data utility and provides more granular protection of user data by personalizing the privacy budget adjustments. The implementation of the MPDP algorithm ensures that user trajectory data retains its original statistical characteristics and spatiotemporal patterns while receiving more precise privacy protection, thereby advancing the application and development of privacy protection technologies in intelligent transportation systems.

Conclusion

To meet the needs for personalized location privacy protection of traffic trajectory data, this paper proposes a research approach that carefully calibrates privacy budgets based on user preferences to ensure personalized privacy. It introduces the foundational theories, technical pathways, and evaluation methods for personalized traffic trajectory privacy protection within location service platforms. Using the PORTO-TAXI dataset as an example, this study provides a demonstrative case of research on personalized privacy protection for traffic trajectories and evaluates the quality of service in trajectory data publication. This effective privacy protection solution, designed for large-scale trajectory datasets, has the potential to significantly impact urban traffic analysis and flow prediction.

REFERENCES

1. Dwork C, McSherry F, Nissim K, et al. (2006). Calibrating noise to sensitivity in private data analysis. Proceedings of the Theory of Cryptography Conference. Berlin, Heidelberg, pp. 265-284.
2. Sarathy R, Muralidhar K. (2010, Sep.). Some additional insights on applying differential privacy for numeric data. In Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2010. Proceedings. Springer Berlin Heidelberg, pp. 210-219.
3. Chen R, Fung BC, Desai BC, Sossou NM. (2012, Aug.). Differentially private transit data publication: a case study on the Montreal transportation system. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 213-221.
4. Andrés ME, Bordenabe NE, Chatzikokolakis K, Palamidessi C. (2013, Nov.). Geo-indistinguishability: Differential privacy for location-based systems. In Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security. pp. 901-914.
5. Wang S, Sinnott R O. (2017, Jul.). Protecting personal trajectories of social media users through differential privacy. Computers & Security. 67: 142-163.
6. Mahdavifar S, Deldar F, Mahdikhani H. (2022, Jan.). Personalized Privacy-Preserving Publication of Trajectory Data by Generalization and Distortion of Moving Points. Journal of Network and Systems Management. 30(1): 1-42.
7. Komishani E G, Abadi M. A generalization-based approach for personalized privacy preservation in trajectory data publishing [C]//6th International Symposium on Telecommunications (IST). IEEE, 2012: 1129-1135.
8. Mahdavifar S, Deldar F, Mahdikhani H. Personalized Privacy-Preserving Publication of Trajectory Data by Generalization and Distortion of Moving Points[J]. Journal of Network and Systems Management, 2022, 30(1): 1-42.
9. Moreira-Matias L, Gama J, Ferreira M, Mendes Moreira J, Damas L. (2013, Sep.). Predicting Taxi–passenger Demand Using Streaming Data. IEEE Transactions on Intelligent Transportation Systems. 14(3): 1393-402.