Poster: Robust Fingerprint of Location Trajectories Under Differential Privacy

Yuzhou Jiang
Case Western Reserve University
yxj466@case.edu

Emre Yilmaz
University of Houston-Downtown
yilmaze@uhd.edu

Erman Ayday
Case Western Reserve University
exa208@case.edu

***Abstract*—Releasing location data without protection raises privacy and liability (e.g., due to unauthorized distribution of such datasets) concerns. To address this, we propose a framework that prevents unauthorized redistribution and protects privacy of users' location data under differential privacy at the same time. Our proposed fingerprinting scheme is robust against random flipping attacks, correlation-based flipping attacks and collusions among multiple parties, which makes it hard for the attackers to infer the fingerprinting codes and avoid accusation. Besides, the proposed fingerprinting scheme increases data utility for differentially-private datasets, which is beneficial for data analyzers in data mining.**

I. Introduction

Location-based services have become one of the most popular services in our daily lives thanks to rapid evolution in mobile technologies and internet of things. Such service providers (e.g., Google) can share such location datasets with a limited number of parties, called data analyzers. Nevertheless, malicious data analyzers, e.g., motivated by profit, may leak their copies to unauthorized parties, which brings significant privacy concerns. In order to prevent unauthorized redistribution, service providers should embed a unique fingerprint into datasets for each data analyzer to enable traceability of the potential leakage.

There are several existing fingerprinting mechanisms, e.g., Boneh-Shaw codes [1] and Tardos codes [5]. However, those traditional digital fingerprinting schemes cannot be directly applied to the location datasets because of correlations in location datasets and their particular utility requirements. In a location trajectory, i.e., an ordered sequence of location points in a location dataset, location points are highly correlated with each other, especially the adjacent location points. Thus, using publicly available correlation models (constructed from public location datasets), a malicious analyzer can identify the fingerprinted data points and avoid accusation.

On the other hand, in recent years, privacy concerns of sensitive datasets have attracted massive attention. However, existing privacy preserving approaches for location data and datasets (i) do not provide liability guarantees against dataset leakage (unauthorized redistribution); and (ii) bring excessive noise to datasets and thus sacrifice data utility.

In this work, we introduce a framework that simultaneously provides robust fingerprinting, privacy preservation, and high utility when sharing location datasets. We propose a robust correlation-based fingerprinting scheme that is robust against multiple attacks, e.g., including correlation attacks, majority collusion attacks, and probabilistic collusion attacks. To mitigate data utility degradation due to the privacy-preserving methods, we propose a smoothing scheme as a post-processing step that aims to restore most correlations between adjacent points along a trajectory and further integrate it into our fingerprinting scheme.

[1]The work was partly supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM013429 and by the National Science Foundation (NSF) under grant numbers 2141622, 2050410, 2200255, and OAC-2112606.

II. System and Threat Model

There are two parties in our setting: a service provider and several data analyzers. The service provider stores the location dataset in their data server and is willing to share them with other parties. Meanwhile, researchers and businesses, categorized as data analyzers, want to access such location datasets. To ensure users' privacy and mitigate unauthorized redistribution, the service provider applies a privacy-preserving mechanism and a fingerprinting scheme to the location dataset before sharing.

The service provider is the only entity that accesses raw data from the users. Thus, we assume the service provider is trusted (i.e., it does not distribute unauthorized copies of users' data to other unauthorized parties). The analyzers can be malicious. The malicious analyzer, denoted as the attacker, understands that the received location dataset has been perturbed under differential privacy to protect dataset participants, but it is curious about the original (non-perturbed) data values in the shared dataset. For this, it can utilize auxiliary information from public sources, e.g., correlations in the map area of interest. To avoid detection during redistribution, the attacker tries to distort the fingerprint signature. They can exploit the public correlations, collude with other analyzers, or even use both to hide their identities. They can perform four types of attacks, i.e.,

- Random flipping attacks. The attacker distorts the location position of a subset of the points that are randomly selected along the trajectory.
- Correlation-based flipping attacks. The attacker utilizes correlation data extracted from public source, analyzes pairwise correlations in the trajectory and flips the points that violate the correlation model.
- Majority collusion attacks. Multiple attackers collude by majority voting at each location point and share the copy.

- Probabilistic collusion attacks. Instead of using majority voting, the attackers take correlation into account while deciding each location point before sharing.

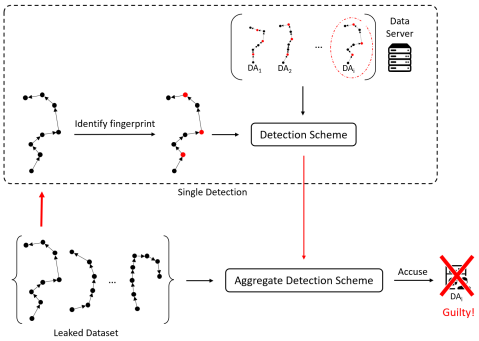

Fig. 1: Detecting the source of the unauthorized redistribution

III. Methodology

The fingerprint detection workflow (for the source of an unauthorized redistribution) is shown in Figure 1. Once a location dataset are found publicly or from other authorized sources, the service provider performs an aggregate detection scheme to identify the source of the leakage. More specifically, for each trajectory in the leaked dataset, the service provider calculates the similarity between it and its fingerprinted version stored on the server. The service provider aggregates the detection results (a set of accused analyzers) and finally accuses the analyzer of leaking the dataset by majority voting.

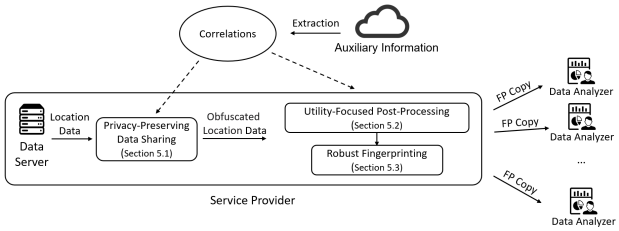

Fig. 2: The system model

The general workflow of the framework is shown in Figure 2. First, we protect the location datasets using PIM [6], i.e., a perturbation-based scheme for location data sharing under differential privacy. There are three advantages for choosing PIM. First, it publishes trajectories with timestamps, while other approaches (e.g., [3]) do not. Second, PIM considers a public correlation model, which is the same one that we use in our approach. Third, a user can choose its ideal privacy-utility tradeoff easily for such perturbation-based methods by selecting an optimal privacy budget.

Similar to other perturbation-based approaches that ensure event-level differential privacy, PIM generates high amount of noise for each location point, leading to significant utility loss in the the shared location dataset. To solve this problem, we propose a utility-focused post-processing scheme to improve the shared dataset's utility. In particular, for each point along the trajectory, we check the correlation model and if it violates the model, we select an alternative point that is highly probable (i.e., the transition probability from the previous point is larger than a threshold τ) and is the closest to the next point. By performing the post-processing scheme, spatio-temporal correlations are restored and data utility of the dataset is improved.

We propose a novel fingerprinting scheme, i.e., direction-sensitive fingerprinting scheme, for location trajectories and use the post-processing scheme while sampling the fingerprinted point. At each position along the trajectory, we check the 2-gram transition and selected the points that are highly probable as the candidates. Then, given a fingerprinting ratio p, we perform probabilistic sampling by assigning probability $1-p$ to the original point and p proportionally assigned to the rest points. If the correlation does not hold, e.g., in a differentially private trajectory, we select the point that is highly probable and is the closet to the next point along the trajectory (i.e., following the post-processing scheme), consider it as the actual point, and then perform the above-mentioned sampling.

IV. Evaluation

We implement our proposed scheme using two real-life datasets, i.e., the GeoLife dataset [7] and the Taxi dataset [4], and two synthetic datasets generated from Brinkhoff generator [2]. We compare our scheme with state-of-the-art fingerprinting approaches, i.e., Boneh-Shaw codes and Tardos codes. Our scheme achieves higher resistance against all the attacked mentioned in Section II, and evaluate the fingerprint robustness against random flipping attacks, correlation-based flipping attacks, majority collusion attacks, and probabilistic collusion attacks.

V. Conclusion

We design a system that achieves both privacy preservation and robust fingerprinting for location datasets. In future work, we plan to improve our correlation model to a higher-order model (e.g., using road structures) and analyze the performance of the scheme. In addition, a non-uniform grid in discretization can be used and different types of collusion attacks can be defined and studied. Moreover, our approach provides differential privacy and fingerprint robustness in two separate steps. Combining those two stepsis another potential future work.

References

[1] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998.

[2] Thomas Brinkhoff. A framework for generating network-based moving objects. *GeoInformatica*, 6(2):153–180, 2002.

[3] Mehmet Emre Gursoy, Ling Liu, Stacey Truex, Lei Yu, and Wenqi Wei. Utility-aware synthesis of differentially private and attack-resilient location traces. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 196–211, 2018.

[4] Luis Moreira-Matias, Joao Gama, Michel Ferreira, Joao Mendes-Moreira, and Luis Damas. Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1393–1402, 2013.

[5] Gábor Tardos. Optimal probabilistic fingerprint codes. *Journal of the ACM (JACM)*, 55(2):1–24, 2008.

[6] Yonghui Xiao and Li Xiong. Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1298–1309, 2015.

[7] Yu Zheng, Xing Xie, Wei-Ying Ma, et al. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

Poster: Robust Fingerprint of Location Trajectories Under Differential Privacy

Yuzhou Jiang[1]; Emre Yilmaz[2]; Erman Ayday[1]
yxj466@case.edu, yilmaze@uhd.edu, exa208@case.edu
[1]Case Western Reserve University, [2]University of Houston-Downtown

Abstract

Releasing location data without protection raises privacy and liability (e.g., due to unauthorized distribution of such datasets) concerns.

To address this, we propose a framework that prevents unauthorized redistribution and protects privacy of users' location data under differential privacy at the same time.

Our proposed fingerprinting scheme is robust against random flipping attacks, correlation-based flipping attacks and collusions among multiple parties, which makes it hard for the attackers to infer the fingerprinting codes and avoid accusation.

Besides, the proposed fingerprinting scheme increases data utility for differentially-private datasets, which is beneficial for data analyzers in data mining.

Motivation

Location-based services have become one of the most popular services in our daily lives, and the service providers are sharing their location datasets.

Malicious identities, e.g., motivated by profit, may leak their copies to unauthorized parties, which brings significant privacy concerns.

Existing digital fingerprinting schemes suffer from detection accuracy loss, which comes from absence of considering correlations within location data.

On the other hand, individual location privacy is important as well. Privacy- preserving schemes should be applied to offer privacy guarantee.

Therefore, we propose a robust fingerprinting scheme on top of privacy-

System Model

There are two parties in our setting: a service provider and several data analyzers.

The service provider stores the location dataset in their data server and is willing to share them with other parties.

Meanwhile, researchers and businesses, categorized as data analyzers, want to access such location datasets.

Threat Model

There are two parties in our setting: a service provider and several data analyzers.

The service provider stores the location dataset in their data server and is willing to share them with other parties.

Meanwhile, researchers and businesses, categorized as data analyzers, want to access such location datasets.

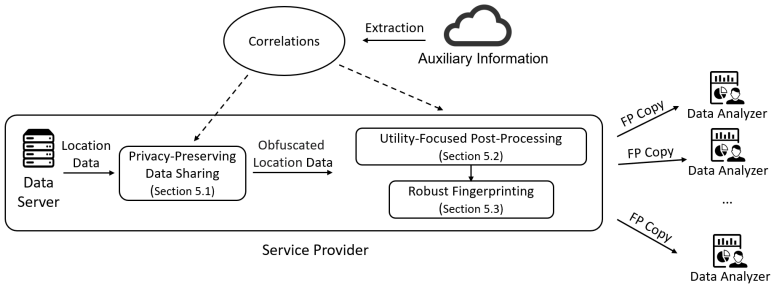

Figure 1. System overview.

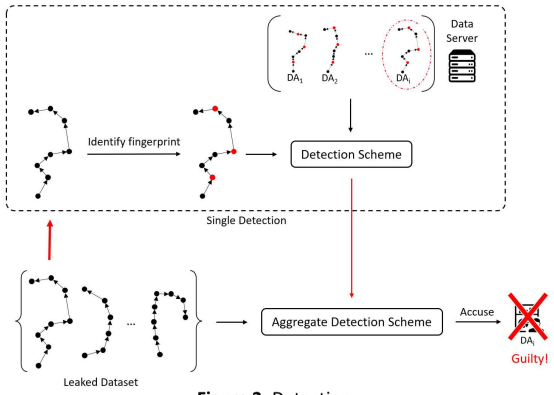

Figure 2. Detection.

Methodology

First, we protect the location datasets using PIM, i.e., a perturbation-based scheme for location data sharing under differential privacy.

To address the problem of data utility degradation, we propose a utility-focused post-processing scheme to improve the shared dataset's utility.

Then we apply our proposed fingerprinting scheme, i.e., direction-sensitive fingerprinting scheme, for location trajectories and use the post-processing scheme while sampling the fingerprinted point.

- At each position along the trajectory, we check the 2-gram transition and selected the points that are highly probable as the candidates.
- Given a fingerprinting ratio p, we perform probabilistic sampling by assigning probability 1-p to the original point and p proportionally assigned to the rest points.
- If the correlation does not hold, e.g., in a differentially private trajectory, we select the point that is highly probable and is the closet to the next point along the trajectory (i.e., following the post-processing scheme), consider it as the actual point, and then perform the above-mentioned sampling.

Datasets

- **GeoLife**: 17,621 trajectories generated by 182 users in Beijing
- **Taxi**: 1,710,670 taxi trajectories in Porto, Portugal
- **Brinkhoff**: 5,000 trajectories in Oldenburg and San Joaquin, respectively, from the Brinkhoff generator for moving objects.

Results

The proposed fingerprinting scheme outperform two existing schemes, i.e., the Boneh-Shaw codes and the Tardos codes in identifying the malicious analyzer against all four attacks.

The proposed scheme achieves higher data utility in most cases.

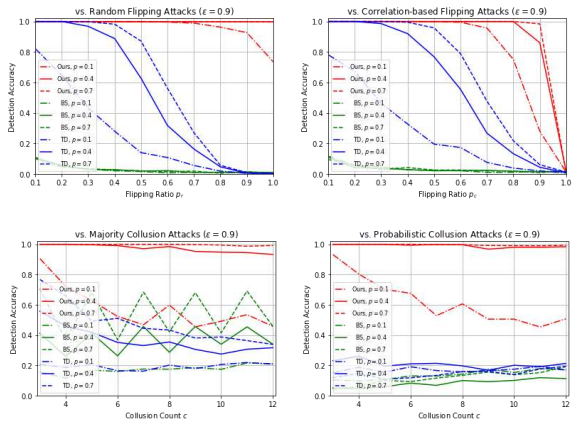

Figure 3. Detection accuracy on differentially private location datasets.

		$\epsilon = 0.9$			$\epsilon = 1.7$			$\epsilon = 2.5$		
		DSFS	BS [5]	Tardos [31]	DSFS	BS [5]	Tardos [31]	DSFS	BS [5]	Tardos [31]
GeoLife [37]	QA Area AvRE	**10.291**	17.000	17.268	**2.339**	4.490	3.758	**0.686**	1.471	1.815
	QA Pattern AvRE	**2.572**	4.636	4.854	**1.193**	2.008	1.496	**0.464**	1.151	1.318
	Popularity KT [16]	**0.609**	0.556	0.560	**0.752**	0.660	0.667	**0.827**	0.763	0.790
	Trip Error [16]	**0.758**	0.812	0.808	**0.667**	0.789	0.793	**0.548**	0.701	0.709
	Diameter Error [16]	**0.138**	0.310	0.308	**0.121**	0.236	0.237	**0.114**	0.208	0.204
	DTW Distance	**297.118**	396.844	399.360	**139.651**	181.780	182.917	**79.682**	97.826	100.114

Table 1. Utility evaluation.

Conclusion

We design a system that achieves both privacy preservation and robust fingerprinting for location datasets.

In future work, we plan to improve our correlation model to a higher-order model (e.g., using road structures) and analyze the performance of the scheme.

In addition, a non-uniform grid in discretization can be used and different types of collusion attacks can be defined and studied.

The work was partly supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM013429 and by the National Science Foundation (NSF) under grant numbers 2141622, 2050410, 2200255, and OAC-2112606.

Full version: https://arxiv.org/abs/2204.04792