Anonymization and De-Anonymization of Mobility Trajectories: Dissecting the Gaps Between Theory and Practice

Huandong Wang, *Student Member, IEEE*, Yong Li, *Senior Member, IEEE*, Chen Gao, Gang Wang, Xiaoming Tao, *Member, IEEE*, and Depeng Jin, *Member, IEEE*

Abstract—Human mobility trajectories are increasingly collected by ISPs to assist academic research and commercial applications. Meanwhile, there is a growing concern that individual trajectories can be de-anonymized when the data is shared, using information from external sources (e.g., online social networks). To understand this risk, prior works either estimate the theoretical privacy bound or simulate de-anonymization attacks on synthetically created datasets. However, it is not clear how well the theoretical estimations are preserved in practice. In this article, we collected a large-scale ground-truth trajectory dataset from 2,161,500 users of a cellular network, and two matched external trajectory datasets from a large social network (56,683 users) and a check-in/review service (45,790 users) on the same user population. The two sets of large ground-truth data provide a rare opportunity to extensively evaluate a variety of de-anonymization algorithms (nine in total). We find that their performance in the real-world dataset is far from the theoretical bound. Further analysis shows that most algorithms have under-estimated the impact of spatio-temporal mismatches between the data from different sources, and the high sparsity of user generated data also contributes to the under-performance. Based on these insights, we propose four new algorithms that are specially designed to tolerate spatial or temporal mismatches (or both) and model location contexts and time contexts. Extensive evaluations show that our algorithms achieve more than 17 percent performance gain over the best existing algorithms, confirming our insights. Further, we propose two new location-privacy preserving mechanisms utilizing the spatio-temporal mismatches to better protect users' privacy against the de-anonymization attack. Evaluation results show that our proposed mechanisms can reduce the performance of de-anonymization attacks by over 8.0 percent, demonstrating the effectiveness of our insights.

Index Terms—Privacy, anonymization and de-anonymization, ISP, spatio-temporal trajectory

1 Introduction

Anonymized user mobility traces are increasingly collected by Internet Service Providers (ISP) to assist various applications, ranging from network optimization [52] to user population estimation and urban planning [13]. Meanwhile, detailed location traces contain sensitive information about individual users (e.g., home and work location, personal habits). Even after the data is anonymized, there is a growing concern that users can still be re-identified through external information [49]. Recently, the US congress has moved towards repealing the Internet Privacy Rules and legalizing ISPs to share (or monetize on) user data [15]. The key question is till yet to be answered: how much of user privacy is leaked if the ISP shares anonymized trajectory datasets?

To answer this question, early research estimates the theoretical privacy bound by assessing the "uniqueness" of the trajectories [11], [49], which shows that trajectory traces are surprisingly easy to de-anonymize. With 4 spatio-temporal points or top 3 most visited locations, results in [11], [49] show that 80–95 percent of the users can be uniquely re-identified in a metropolitan city.

Recently, researchers start to evaluate more practical attacks by de-anonymizing ISP trajectories using external information (e.g., location check-ins from social networks) [9], [12], [16], [17], [26], [31], [32], [33], [37], [38], [39], [44]. However, due to the lack of large empirical *ground-truth* datasets, researchers have to settle on synthetically generated data (e.g., using parts of the same dataset as the victim dataset and the external information source) [26], [38]. To date, it is still not clear how easy (or difficult) attackers can massively de-anonymize user trajectories in practice.

In this work, we spent significant efforts to collect two large-scale *ground-truth* datasets to close the gaps between theory and practice. By collaborating with a major ISP and two large location-based online services in China, we obtain 2,161,500 ISP trajectories (as the target dataset), 56,683 users' GPS/check-in traces from a large social network (external information) and 45,790 users' GPS traces from a large online review service (external information). The three datasets

- *H. Wang, Y. Li, C. Gao, X. Tao, and D. Jin are with the Beijing National Research Center for Information Science and Technology (BNRist), Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: whd14@tsinghua.org.cn, {liyong07, jindp}@tsinghua.edu.cn, gc16@mails.tsinghua.edu.cn, 271363488@qq.com.*
- *G. Wang is with the Department of Computer Science, University of Illinois at Urbana-Champaign (UIUC), Champaign, IL 61820 USA. E-mail: gangwang@vt.edu.*

Manuscript received 27 Sept. 2018; revised 16 Aug. 2019; accepted 27 Oct. 2019. Date of publication 11 Nov. 2019; date of current version 3 Feb. 2021. (Corresponding author: Yong Li.)
Recommended for acceptance by S. K. Das.
Digital Object Identifier no. 10.1109/TMC.2019.2952774

1536-1233 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

cover the same user population with the ground-truth mapping.[1] Using this dataset, we seek to empirically evaluate how well de-anonymization algorithms approach the privacy bound, and what practical challenges (if any) that are often neglected when designing these algorithms. Answering this question helps to provide more accurate assessment on the privacy risks of sharing the anonymized ISP traces.

By implementing and running 9 major de-anonymization algorithms against our datasets, we find the existing algorithms largely fail the de-anonymization task using practical data. Their performance is far from the privacy bound [11], [49], and massive errors occur, i.e., the hit-precision is less than 20 percent. Further analysis reveals a number of key factors that are often neglected by algorithm designers. First, there widely exist significant spatio-temporal mismatches between the ISP trajectories and the external GPS/check-in traces, caused by positioning errors and different location updating mechanisms. In addition, user trajectory datasets are highly sparse across time and users, making the de-anonymization attack very challenging in practice.

To validate our insights, we design 4 new algorithms that specially address the practical factors. More specifically, we propose a spatial matching (SM) algorithm and a temporal matching (TM) algorithm, which tolerate spatial and temporal mismatches respectively. Further, we build a Gaussian and Markov based (GM) algorithm that considers spatio-temporal mismatches simultaneously. Finally, we enhance the GM model by adding a user behavior model in terms of time context to incorporate human mobility patterns (GM-B algorithm). Extensive evaluation shows that our algorithms significantly outperform existing algorithms. More importantly, our experiments reveal new insights into the relationship between human mobility and privacy. We find that tolerating temporal mismatches is more important than tolerating spatial mismatches. An intuitively explanation is that human mobility has a strong locality, which naturally sets a bound for location mismatches. However, at the temporal dimension, since the errors are unbounded, making the algorithm aware of the temporal matches makes a bigger difference to the de-anonymization performance. Finally, the GM and GM-B algorithms achieve even better performance by considering different mismatches and human behavior models at the same time.

Having demonstrated the usefulness of the practical factors in de-anonymization attack, we further consider utilizing them to better protect users' privacy against the de-anonymization attack. Specifically, we propose 2 new location-privacy preserving mechanisms, which utilize the distribution of spatio-temporal mismatches in obfuscating the ISP trajectories. Evaluation results show that our proposed mechanisms can reduce the performance of de-anonymization attacks, demonstrating the effectiveness of our insights.

Overall, our work makes four key contributions:

- First, we collect three large-scale trajectory datasets (with ground-truth) to evaluate de-anonymization attacks. The datasets contain 2,161,500 ISP trajectories, 56,683 external trajectories, and 45,790 external trajectories respectively, which help to overcome the limitations of theoretical analysis and simulated validations.
- Second, we build an empirical evaluation framework by categorizing and implementing existing de-anonymization algorithms (9 in total) and evaluation metrics. Our evaluation on real-world datasets reveals new insights into the existing algorithms' under-performance.
- Third, we propose new algorithms by addressing practical factors such as spatio-temporal mismatches, location contexts, and time contexts. Optional components such as user historical trajectories can also be added to our framework to improve the performance. Extensive performance evaluation shows that our algorithms achieve over 17 percent performance gain in terms of hit-precision. In addition, our algorithms stay robust using parameters transferred from other external datasets.
- Finally, we propose 2 new location-privacy preserving mechanisms by utilizing the spatio-temporal mismatches to better protect users' privacy against the de-anonymization attack. Evaluation results show that our proposed mechanisms can reduce the performance of de-anonymization attacks by over 8.0 percent, demonstrating the effectiveness of our insights.

A conference version of this paper was published in [48]. Compared with the conference version, we further consider different location-privacy preserving mechanisms in our threat model. What's more, we propose two new algorithms of location-privacy preserving mechanisms which utilize the spatio-temporal mismatches to better protect users' location privacy. Experimental results show that by considering spatial and temporal mismatching, location-privacy preserving mechanisms can be more effective. In addition, additional important baselines, experimental analysis, and discussion about our proposed algorithms are supplemented in this version.

2 Threat Model

In this work, we seek to examine how much of individuals' privacy will be leaked if the ISP shares their anonymized trajectory datasets. We investigate this problem by implementing and testing a wide range of de-anonymization attack schemes against real-world trajectory datasets. To better describe the de-anonymization problem, we first formally define the threat model in this section. Our threat model mainly consists of two components, i.e., the ISP that is the data owner to publish anonymized trajectory traces, and the adversary which seeks to re-identify users in the published dataset. For the ease of reading, we summarize the key notations in Table 1.

2.1 Location Data Publishing by ISP

Before ISPs publish the trajectory dataset, usually two location-privacy preserving mechanisms (LPPMs) are implemented, i.e., anonymization and obfuscation, which are introduced as follows.

[1] Personally identifiable information (PII) has been removed before the data is handled to us. This work received the approvals from our local intuitional board, the ISP, the online social network, and the online review service.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
A List of Commonly Used Notations

Notat.	Description
$\boldsymbol{U}$	The set of true identities of all users.
$\boldsymbol{V}$	The set of pseudonyms of all users.
$\mathcal{T}$	The set of all time slots.
$\mathcal{R}$	The set of all regions.
$\mathcal{L}$	The set of anonymized ISP traces.
$\mathcal{S}$	The set of traces as external information (adversary knowledge).
$\boldsymbol{L}_v$	ISP trajectory of user with pseudonym v.
$\boldsymbol{S}_u$	External trajectory of user u.
σ	Anonymization function mapping $\boldsymbol{U}$ to $\boldsymbol{V}$.
$\boldsymbol{L}_{\sigma(u)}$	ISP trajectory of user u.
$L_v(t)$	Location in the ISP trajectory of user v at time slot t.
$S_u(t)$	Location in the external trajectory of user u at time slot t.
$D(\cdot,\cdot)$	Similarity score function between trajectories.
$R(u,D)$	The rank of the true matched trajectory of u based on similarity function D.
$\mathcal{N}(\cdot\|u_p,\Sigma_p)$	Gaussian distribution with mean vector u_p and covariance matrix Σ_p.
H_u, H_l	Maximum tolerant temporal mismatches in two time directions.
$\pi(p), \sigma(p)$	Parameters of Gaussian mixture model corresponding to temporal mismatches of p time units.
T_v	Transition matrix of user v.
E_v	Marginal distribution of user v.
$\Phi(\mathcal{S},D)$	Performance metric of de-anonymization attack.
ξ_0	Perturbation strength.
λ_h	Location hiding level.
$I(\cdot)$	Indicator function of logical expressions with $I(\mathit{true})=1$ and $I(\mathit{false})=0$.

Anonymization Mechanism. Let $\boldsymbol{U}$ represent the set of the identities of all users. Before the dataset is published, ISPs use a map function σ to anonymize it, i.e., replacing the user identity u with pseudonym $\sigma(u)$. We further define $\boldsymbol{V}$ as the set of pseudonyms of all users.

After anonymization, a spatio-temporal record in the dataset is defined as a 3-tuple (v,t,r), where $v \in \boldsymbol{V}$ is the pseudonym of the user, and r, t are the observed location and timestamp, respectively.

Obfuscation Mechanism. After anonymizing the location dataset by replacing user IDs with the pseudonyms, ISPs will further obfuscate the location records to protect user privacy, i.e., reduce the spatio-temporal information involved in each trajectory. In our work, we consider the most common obfuscation mechanisms as [39], which are summarized as follows:

- Perturbation: In perturbation mechanism, geographical coordinates of each location record are modified by adding some random noise. In this work, we consider the most common zero-mean Gaussian noise [42]. In addition, we denote the root mean square of the noise as the perturbation strength ξ_0.
- Location hiding: In this mechanism, according to the anonymization requirement, every location record is independently eliminated (i.e., its location is replaced by $\emptyset$) with probability λ_h, which is denoted as the location hiding level.

After the two LPPMs, the ISP trajectory of the user with pseudonym $v \in \boldsymbol{V}$ is represented as a T-size vector $\boldsymbol{L}_v = (L_v(1), L_v(2), \ldots, L_v(T))$, where $L_v(t)$ represents the location observed at time slot t, and T is the number of time slots. For time slots with a location record, $L_v(t)$ is the corresponding geographic coordinates. For time slots without a location record, $L_v(t)$ is $\emptyset$. We further define $\mathcal{L}$ as the set of all mobility traces in the ISP dataset, as $\mathcal{L} = \{\boldsymbol{L}_v | v \in \boldsymbol{V}\}$.

2.2 Adversary

In the de-anonymization attack, an adversary seeks to re-identify users using external information. An adversary is described by two components, i.e., utilized knowledge (external information), and attack method.

Adversary Knowledge. Adversaries can use different types of external knowledge for de-anonymization. In this paper, we mainly focus on two categories of adversaries. The first category is the company-level attacker, e.g., application and service providers who have users' sub-trajectory information uploaded by the application software installed on the users' mobile devices. The second category is the individual-level attacker, who can obtain external information by crawling the publicly available location information (online check-ins) shared by users.

For an arbitrary adversary, regardless of its category, we use a fixed-size vector $\boldsymbol{S}_u = (S_u(1), S_u(2), \ldots, S_u(T))$ to represent its external information, with $S_u(t)$ representing the location (geographic coordinates) observed at time slot t for user $u \in \boldsymbol{U}$. Similarly, we set $S(t) = \emptyset$ in time slot t without any locations. We further define $\mathcal{S} = \{\boldsymbol{S}_u | u \in \boldsymbol{U}\}$ as the set of all traces in the external information.

Attack Method. Attack method of the adversary is described by the similarity score function D defined between trajectories in ISP dataset and external information, i.e., $D : \mathcal{L} \times \mathcal{S} \rightarrow \mathbb{R}$, where $\mathbb{R}$ is the set of real numbers. Based on this similarity function, for each user u with external trajectory $\boldsymbol{S}_u$, the adversary ranks all its candidate trajectories in the ISP dataset. The goal of the adversary is to rank the ISP trajectory belonging to u, i.e., $\boldsymbol{L}_{\sigma(u)}$ as high as possible.

More specifically, we use $R(u,D)$ to denote the rank of $\boldsymbol{L}_{\sigma(u)}$ based on similarity function D. Further, denote function h as the metric of the rank $R(u,D)$. For higher $R(u,D)$, $h(R(u,D))$ is larger. Then, the performance of the attack method can be expressed as follows,

$$\Phi(\mathcal{S},D) = \frac{1}{|U|} \sum_{S_u \in \mathcal{S}} h(R(u,D)).$$

For any adversaries, given external information $\mathcal{S}$, the target can be expressed as follows,

$$\arg\max_{D} \Phi(\mathcal{S},D).$$

In terms of the rank, a well-established and widely-used evaluation metric is the hit-precision of top-k candidates. If the rank of the true matched trajectory in the k candidates is x, the hit-precision $h(x)$ can be calculated as follows,

$$h(x) = \begin{cases} \frac{k-(x-1)}{k}, & \text{if } k \geq x \geq 1, \\ 0, & \text{if } x > k. \end{cases} \tag{1}$$

Overall, a larger hit-precision means that the true matched trajectory is ranked higher, and indicates a better de-anonymization performance. For example, if the true matched

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
Statistics of Collected Datasets

Dataset	Total #Users	Total #Records	Total #Regions (1km^2)	Mean #Recd. /User	Mean #Loc. /User
ISP	2,161,500	134,033,750	3,056	62.01	9.19
Weibo App-level	56,683	239,289	4,346	4.22	1.67
Weibo Check-in (Historical)	10,750	141,131	2,394	13.15	7.00
Weibo Check-in (Synchronized)	503	873	686	1.74	1.34
Dianping App-level	45,790	107,543	3,931	2.35	1.61

trajectory $\boldsymbol{L}_{\sigma(u)}$ has the largest similarity, i.e., $D(\boldsymbol{S}_u, \boldsymbol{L}_{\sigma(u)}) \geq D(\boldsymbol{S}_u, \boldsymbol{L}_v)$ for any $v \in \boldsymbol{V}$, then, $R(u, D) = 1$ and $h(R(u, D)) = 1$. If $\boldsymbol{L}_{\sigma(u)}$ ranks 3 in all candidate trajectories in $\mathcal{L}$, $R(u, D) = 3$ and $h(R(u, D)) = \frac{k-2}{k}$.

3 Ground-Truth Trajectory Datasets

To empirically assess the effectiveness of de-anonymization algorithms against large-scale trajectories from ISP, we collect real-world ground-truth datasets. The datasets are obtained from a major ISP, a large online social network, and a check-in/review service for *an overlapped user population*. We also have the ground-truth mapping between users across these three datasets. The datasets are obtained through our research collaborations and a summary of the datasets is shown in Table 2. Below, we describe the datasets in detail and perform a preliminary analysis.

3.1 ISP Dataset

The main dataset contains 2,161,500 ISP trajectories from a major cellular service provider in China from April 19 to April 26 in 2016 covering whole metropolitan area of Shanghai. Each trajectory is constructed based on the user's connection records to the base stations (cellular towers). Each spatial-temporal data point in the trace is characterized by an anonymized user ID, base station (BS) ID and a timestamp. This dataset will serve as the target dataset for evaluating the de-anonymization attack.

3.2 Social Network Dataset

As the external information for de-anonymizing users, we also collect datasets from Weibo, a large online social network in China with over 340 million users. The challenge is to obtain the ground-truth mapping between users in the ISP dataset and the Weibo users. This is doable from the ISP side because Weibo's mobile app uses HTTP to communicate with its servers and the Weibo ID is visible in the URL. Given the sensitivity of the data, we approached Weibo's Data and Engineering team to ask for the permission to collect the Weibo IDs *from the ISP end* for this research. After setting up a series of privacy and data protection plans, Weibo gave us the approval to use the data only for research purposes (more detailed data protection and ethical guidelines are in Section 3.5).

App-Level GPS Data. With the permission of Weibo, our collaborators in the ISP marked the Weibo sessions for users that appear in the ISP traces, within the same time window April 19 to April 26 in 2016. In this way, we construct an external GPS dataset of 56,683 matched users. In this dataset, each location trajectory is characterized by a user's Weibo ID, and a series of GPS coordinates that show up in HTTP sessions between the mobile app and Weibo server. This dataset represents location traces that users report to the Weibo server. Using this dataset as external information, we can evaluate how much Weibo service can de-anonymize a shared ISP dataset, i.e., company-level attackers. Note that the Weibo ID is only visible to the ISP collaborator. The ID has been replaced with an encrypted bitstream before the data is handled to us. A mapping between the bitstream to the anonymized ISP user ID is provided to us.

User Location Check-ins. Based on the matched Weibo IDs, our collaborator at the ISP also helped to collect a check-in dataset using Weibo's open APIs.[2] This dataset covers the same time window of previous datasets (Synchronized), as well as all the historical check-ins of the matched users (Historical). Since check-in data is publicly available to any third-parties, we use it to evaluate how much *any attackers* can de-anonymize a shared ISP dataset, i.e., individual-level attackers. Similarly, we only access the anonymized ID, instead of the actual Weibo ID.

3.3 Review Service Dataset

To make sure our analysis is not biased towards a single dataset, we collected a secondary dataset to validate our observations. The secondary dataset was collected from Dianping, the largest online review service in China. Dianping has similar features as the Yelp and Foursquare combined. It also uses HTTP for its mobile app and the user ID is visible to ISP. Following the same procedure, our ISP collaborator marked Dianping sessions in the ISP traces within the same time window April 19–26 in 2016. This produced an external GPS dataset of 45,790 matched users. Each location trajectory is characterized by a user's Dianping ID, and a series of GPS coordinates with timestamps.

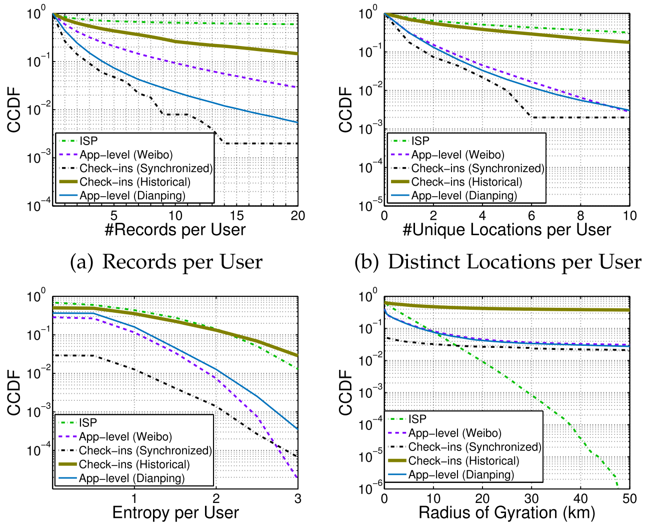
Similarly, the Dianping ID is only visible to the ISP collaborator. The ID has been replaced by an encrypted bitstream in our dataset. A mapping between the bitstream and the anonymized ISP user ID is provided to us. We have also notified Dianping Inc. about our research plan and received their consent.

3.4 Data Processing

The collected datasets have different formats and precision in terms of the time and location. We seek to format the data in a consistent manner before our evaluation.

[2] http://open.weibo. com

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.



(a) Records per User

(b) Distinct Locations per User

(c) Trajectory Entropy per User

(d) Radius of Gyration per User

Fig. 1. Complementary cumulative distribution function (CCDF) of the number of records, number of distinct locations, trajectory entropy, and radius of gyration per user.

Converting Basestation ID to GPS. To construct user mobility traces from the ISP data, we first convert the ID of base stations to their geographical coordinates (longitudes and latitudes) based on the ISP offered database, and use it to represent the user location.

Building Trajectories. Since the timestamps have different resolutions in different datasets, we build the trajectory based on discrete time intervals. More specifically, we divide the time span of a user's trace into many fixed sized time bins. Then, we add one location data point to each time bin to build the vector S_u and L_v. To systematically match GPS locations across datasets, we also map the GPS coordinates into regions with a certain spatial resolution. More specifically, we use a similar method from [38], [39]. The idea is dividing the whole city into grids, where each grid represents a "region". Different regions do not overlap with each other. In this way, we use a tuple of a time bin and a location region to consistently represent a location record. After the data processing, we define $\mathcal{T}$ and $\mathcal{R}$ as the set of all the time bins and the set of all the spatial regions, respectively. These above steps introduce two key parameters to adjust the temporal and spatial resolutions of the dataset. By default, we set the time bin as 1 hour, and the spatial resolution as 1 km. In the later analysis, we will also test different temporal and spatial resolutions to assess the influence to our results and conclusions.

3.5 Ethics

We have taken active steps to preserve the privacy of involved users in our datasets. First, all the data collected for this study was kept within a safe data warehouse server (behind a company firewall). We have never taken any fragment of the dataset away from the server. Second, the ISP employee (our collaborator) anonymized all the user identifiers, including the unique identifiers of cellular network users, and the actual IDs of Weibo and Dianping users. Specific steps (e.g., crawling Weibo check-ins) that require unencrypted Weibo/Dianping IDs were performed by the ISP employee. After obtaining the target trajectory datasets, the ISP employee removed the actual IDs from the datasets, and associated each entry with an encrypted bitstream. The mapping between the bitstream and the anonymized cellular user identifier is provided to us. The real user IDs are never made available to, or utilized by us. All our data processing was fully governed by the ISP employee to ensure compliance with the commitments of privacy stated in the Term-of-Use statements. Third, we obtained the approval for using the Weibo data and Dianping data from the Data and Engineering team of Weibo and Dianping, under the condition that the data is processed strictly following the above steps and can only be used for research. Finally, our research plan has been approved by our local institutional board.

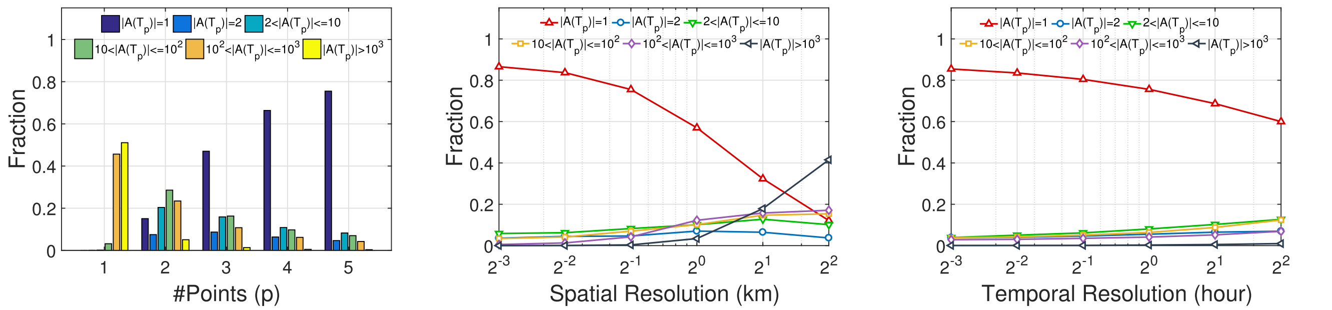
We believe through our work, we can provide more comprehensive understandings on the privacy risks of users when anonymized ISP trajectory data is shared. The results will help the stakeholders to make more informed decisions on designing privacy policies to protect user privacy in the long run.

3.6 Preliminary Data Analysis

Table 2 shows the basic statistics of the three datasets. The ISP dataset is the largest one with 2,161,500 users. The Weibo dataset (app level), as the external information source, has 56,683 users, which is about 3 percent of the ISP user population. This indicates that using this external information, the adversary still faces non-trivial noises to re-identify the target users. Compared to other datasets, the ISP dataset covers a bigger portion of a user's mobility trace with a higher average number of records and distinct locations per user (62.01 and 9.19). The Weibo and Dianping datasets (app level) have 4.22 and 2.35 records on average per user respectively. The Weibo check-in datasets cover both the same time-window as other datasets (Synchronized) as well as the historical check-ins of the users (Historical), with 1.74 and 13.15 records on average per user respectively. Not too surprisingly, the check-in dataset is sparser than the app-level datasets of Weibo and Dianping. Overall, the 4 external trajectory datasets from 2 different online services provide a diverse and large collection of user trajectories with a ground truth mapping to the ISP dataset. In addition, as shown in Table 2, the number of covered geographic regions of 1 km × 1 km in these collected datasets ranges from 686 to 4,346. The collected datasets cover both sparse trajectories (check-in dataset) and dense trajectories (app-level datasets of Weibo and Dianping) as the adversary knowledge. This helps to solve the critical problem of lacking ground truth data in the existing works [11], [38].

Further, we show the complementary cumulative distribution functions (CCDF) of the number of records, number of distinct locations, trajectory entropy, and radius of gyration in Fig. 1. Specifically, we set the maximum limits of horizontal axises as 20 and 10 in Figs. 1a and 1b respectively, because we find there is not any synchronized check-in trajectories with more than 20 records and 10 distinct visited locations. In addition, for each user u, its trajectory entropy can be calculated by $Entropy = -\sum_{r\in\mathcal{R}} P_r(u)\log P_r(u)$, in which $P_r(u)$ is the probability of visiting region r by u. It describes the regularity of traces in spatial dimension [10]. In addition, radius of gyration [18] is defined as the mean

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.



(a) Uniqueness of ISP trajectories

(b) Uniqueness of the ISP trajectories under different spatial resolutions ($p = 5$)

(c) Uniqueness of the ISP trajectories under different temporal resolutions ($p = 5$)

Fig. 2. Theoretical analysis of the privacy bound, where p is the number of randomly selected data points from the trajectories as the external observations.

square root of the distance of each point in the trajectory to its center of mass, and can be formally defined as $r_g = \sqrt{\Sigma_{t=1}^{T}(L(t) - L_{cm})^2/T}$, where $L_{cm} = 1/T\sum_{t=1}^{T} L(t)$ is the center of mass of the trajectory. It reflects the range of a user' activity area. The results coincides with Table 2. That is, ISP trajectories have largest number of records, distinct locations, and entropy, while check-in trajectories exhibit the strongest sparsity. However, as we can observe from Figs. 1c and 1d, ISP trajectories have smaller radius of gyration and larger entropy, indicating that users with larger activity area are more likely to be captured in the external datasets.

4 De-Anonymization in Practice

Based on the above three large-scale datasets, we investigate the potential privacy leakage of the ISP trajectory dataset. In order to show the theoretical bound of privacy leakage, we first investigate the uniqueness of trajectories in Section 4.1. Then, comparing with the theoretical bound, we implement 9 existing de-anonymization algorithms in practice, and show their performance in Section 4.2.

4.1 Theoretical Privacy Bound

Uniqueness of trajectory in an anonymity mobility dataset is a well-recognized metric to measure the privacy bound and the de-anonymization risks [8], [11], [19], [41], [49]. In 1930, Edmond Locard showed that 12 points are sufficient to uniquely identify a fingerprint [11]. Similarly, the analysis of the uniqueness of trajectories is to estimate the number of points necessary to uniquely identify the mobility trace of an individual.

Uniqueness metric is based on the concept of k-anonymity model [45], which is computed as follows. Let T_p denote a sub-trajectory of a user with p randomly selected spatio-temporal points. Then we search for other trajectories in the dataset that match or contain the p points of T_p. We define the matched trajectories as the *anonymity set* of T_p denoted as $A(T_p)$. Then the user's uniqueness is characterized by $|A(T_p)|$, i.e., the number of matched trajectories in the anonymity set. Intuitively, the uniqueness metric estimates how likely a user can be re-identified if an external adversary observed a random p points in her trace. If $|A(T_p)| = 1$, its anonymity set only contains one trace, i.e., trajectory of its true owner. This means the p points can uniquely re-identify the user. For example, the hit-precision with $k = 1$ cannot exceed the fraction of $|A(T_p)| = 1$ when we only have p external location records. Thus, it characterizes the upper bound of de-anonymization performance with p points.

As for other privacy metrics such as l-diversity [24] and t-closeness [27], actually they are stricter privacy metrics than k-anonymity. For example, trajectories contained in 2-anonymity set, i.e., $|A(T_p)| = 2$, are very likely to not satisfy requirement of l-diversity and t-closeness. Thus, potential risks measured based on l-diversity and t-closeness are higher than that of k-anonymity. As mentioned above, our utilized uniqueness metric has characterized the upper bound of de-anonymization performance with p points. Thus, there is no necessity to consider these stricter privacy metrics.

Note that the above trajectory matching is based on both location and time. We consider two data points match if they fall into the same location region and time bin (we defined the location region and time bin in Section 3.4). For example, if two trajectories show users visiting the same locations in the same order but at different time slots, they are not the same. The uniqueness metric is the very basic metric to quantify the de-anonymization risk. More sophisticated metric can further consider the location context (e.g., user density in a given area) and the time context (e.g., day and night patterns) [11].

We focus on the uniqueness of the ISP trajectories to show their de-anonymization risks. Specifically, we randomly sample 10 sub-trajectories with p points for each ISP trajectory and compute the average fraction of anonymity sets with different size to characterize the uniqueness of trajectories. We first show the uniqueness of ISP trajectories as the function of p in Fig. 2a. We can observe that the uniqueness of ISP trajectories is high, i.e., 5 points can uniquely identify over 75 percent users, indicating their potential high risk to be de-anonymized. In addition, we analyze the influence of the spatio-temporal resolutions on the uniqueness to show the potential privacy gain of LPPMs. For external datasets, ISP cannot implement LPPMs on them. Thus, they are ignored in this analysis. We fix the number of spatio-temporal points as 5, and obtain the uniqueness of the ISP dataset. As shown in Figs. 2b and 2c, the uniqueness measure is not very sensitive to the spatio-temporal resolution (log scale x-axis). Reducing the temporal resolution from 30 minutes to 4 hours only leads to the decreasing of uniqueness by 20 percent, while reducing the spatial resolution from 250 meters to 1 kilometer only leads to the

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

decreasing of uniqueness by 26 percent. The resolution degradation is likely to hurt the usability of the dataset which only brings in a little privacy benefit in exchange.

In summary, the obtained user trajectories are highly unique. Even when the spatial granularity is very low, 5 points are sufficient to uniquely identify over 75 percent users, indicating the high potential risk of individual trajectories to be de-anonymized, which exposes a big threat to users' privacy.

4.2 Actual Performance of Attack Methods

To examine the effectiveness of de-anonymization attacks, we implement 9 major attacking algorithms that are designed (or can be adopted) to work on trajectory datasets.

HMM. Shokri et al. [39] focus on de-anonymizing users' trajectories based on their mobility patterns. Specifically, they train a Markov model to describe the mobility of users, which is represented by the transition matrix T^v. They also define a function $f : \mathcal{R} \times \mathcal{R} \rightarrow \mathbb{R}$ to describe the spatial mismatching between the adversary's knowledge and users' true locations. After using $\boldsymbol{L}_v$ to estimate T^v, the similarity score can be calculated by

$$\begin{aligned} D_{\mathrm{HMM}}(\boldsymbol{S}_u, \boldsymbol{L}_v) &= P(\boldsymbol{S}_u | T^v) \\ &= \sum_{\boldsymbol{Z}} \prod_{t \in \mathcal{T}} f(Z(t), S(t)) T^v_{Z(t-1), Z(t)}, \end{aligned} \tag{2}$$

where Z is the hidden variable representing users' true locations. In addition, since we mainly focus on the performance of the similarity score function in terms of the rank, this similarity does not need to be normalized.

MKV. Mulder et al. [12] also focus on de-anonymization based on Markov model. Specifically, they measure the similarity of the transition matrix and marginal distribution of trajectories in different datasets, which can be defined as

$$D_{\mathrm{MKV}}(\boldsymbol{S}_u, \boldsymbol{L}_v) = \sum_{r_1, r_2 \in \mathcal{R}} E^u(r_1) T^u_{r_1, r_2} E^v(r_1) T^v_{r_1, r_2}. \tag{3}$$

HIST. Naini et al. [31] focus on de-anonymization by matching the histograms of trajectories. Specifically, they use Γ_u to denote the histogram of user u defined as $\Gamma_u(r) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} I(S_u(t) = r)$. Based on the histograms, their similarity score can be defined as

$$D_{\mathrm{HIST}}(\boldsymbol{S}_u, \boldsymbol{L}_v) = -D_{\mathrm{KL}}(\Gamma_u | \bar{\Gamma}) - D_{\mathrm{KL}}(\Gamma_v | \bar{\Gamma}), \tag{4}$$

where $\bar{\Gamma} = (\Gamma_u + \Gamma_v)/2$, and D_{KL} the Kullback-Leibler divergence function [46].

LRCF. Goga et al. [16] further consider the popularity of different regions. Specifically, they apply the *term frequency-inverse document frequency (TF-IDF)*[43] weighting scheme to the histograms, i.e., $\Lambda_u(r) = \Gamma_u(r) / \log{(IDF(r))}$, where $\mathrm{IDF}(r) = \sum_{u \in \boldsymbol{U}} \sum_{t \in \mathcal{T}} I(S_u(t) = r)$ is the number of records in region r of the whole dataset. Then, they measure the cosine similarity between Λ_u and Λ_v as follow:

$$D_{\mathrm{LRCF}}(\boldsymbol{S}_u, \boldsymbol{L}_v) = \Lambda_u^T \Lambda_v / \|\Lambda_u\| \|\Lambda_v\|. \tag{5}$$

WYCI. Rossi et al. [38] propose a probabilistic de-anonymization algorithm. They use the frequency of user login in different locations to approximate the probability of visiting these locations by $P(r|\boldsymbol{L}_v) = \frac{n_r^v + \alpha}{\sum_{r \in \mathcal{R}} n_r^v + \alpha |\mathcal{R}|}$, where n_r^v is the number of times user v visits location r, $|\mathcal{R}|$ is the number of locations in the dataset, and $\alpha > 0$ is the smoothing parameter, which is used to eliminate zero probabilities. By following the recommended setting in [28], we set $\alpha = 0.1$. Then, their similarity score is defined as follow:

$$D_{\mathrm{WYCI}}(\boldsymbol{S}_u, \boldsymbol{L}_v) = \prod_{t \in \mathcal{T}, S_u(t) \neq \emptyset} P(S_u(t) | \boldsymbol{L}_v). \tag{6}$$

ME. Cecaj et al. [9] estimate the probability of trace-user pairs being the same person according to the number of their matching elements. Their similarity score is defined as the number of meeting events as follow:

$$D_{\mathrm{ME}}(\boldsymbol{S}_u, \boldsymbol{L}_v) = \sum_{t \in \mathcal{T}} I(S_u(t) = L_v(t)). \tag{7}$$

POIS. Riederer et al. [37] mainly consider using the "encountering" events to match the same users. They assume the number of visits of each user to a location during a time period follows Poisson distribution, and an action (e.g., login) on each service occurs independently with Bernoulli distribution. Based on this mobility model, the algorithm computes a score for every candidate pair of trajectories, which can be calculated as follows,

$$D_{\mathrm{POIS}}(\boldsymbol{S}_u, \boldsymbol{L}_v) = \sum_{t \in \mathcal{T}} \sum_{r \in \mathcal{R}} \phi_{r,t}(S_u(t), L_v(t)), \tag{8}$$

where ϕ measures the importance of an "encountering" event in location r at time slot t, and can be given as follows,

$$\phi_{r,t}(S_u(t), L_v(t)) = \frac{P(S_u(t) = r, L_v(t) = r | \sigma(u) = v)}{P(S_u(t) = r) P(L_v(t) = r)}. \tag{9}$$

It can be calculated based on their mobility model with the assumptions of Poisson visits and Bernoulli actions.

NFLX. Narayanan et al. [32] propose a de-anonymization algorithm that can tolerate some mistakes in the adversary's knowledge. In order to adapt this algorithm to the trajectory data, we use the similarity score modified by [37], which is defined as follows:

$$D_{\mathrm{NFLX}}(\boldsymbol{S}_u, \boldsymbol{L}_v) = \sum_{(r,t): r = S_u(t) = L_v(t)} w_r * f_r(\boldsymbol{S}_u, \boldsymbol{L}_v), \tag{10}$$

where $w_r = 1 / \mathrm{In}(\sum_{v,t} L_v(t) = r)$ and $f_r(\boldsymbol{S}_u, \boldsymbol{L}_v)$ is given by

$$f_r(\boldsymbol{S}_u, \boldsymbol{L}_v) = e^{\frac{n_r^v}{n_0}} + e^{-\frac{1}{n_r^v} \sum_{t: S_u(t) = r} \min_{t': L_v(t') = r} \frac{|t - t'|}{\tau_0}}. \tag{11}$$

In addition, n_r^v is the number of times user v visits location r. Temporal mismatches are considered in this algorithm. However, it does not perform well under spatial mismatches.

MSQ. Ma et al.[26] find the matched traces by minimizing the expected square between them. That is, their similarity score can be expressed as follows:

$$D_{\mathrm{MSQ}}(\boldsymbol{S}_u, \boldsymbol{L}_v) = -\sum_{t \in \mathcal{T}} |L_v(t) - S_u(t)|^2. \tag{12}$$

Spatial mismatches are considered in this algorithm. However, it does not perform well under temporal mismatches.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

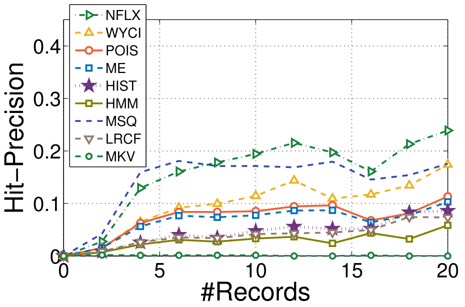


Fig. 3. Hit-precision of different algorithms as a function of the number of records in Weibo's app-level trajectories.

Note that POIS, HMM, ME, MSQ algorithms are essentially based on the "concurrent" events and do not expect temporal mismatches. For these algorithms, we define "concurrency" based on 1-hour time bins as the default setting, i.e., if timestamps of two records are within the same 1-hour time bin, we regard them as "concurrent". On the other hand, POIS, WYCI, HIST, ME, NFLX, LRCF, MKV algorithms are based on the definition of "co-located" events and do not expect spatial mismatches. For these algorithms, we define the "co-location" based on the 1 km × 1 km geographic grids, i.e., if two records are located in the same geographic grid, we regard them as "co-located". The resolution values 1 hour and 1 km are set as the default. We will further analyze the influence of the spatio-temporal resolutions to these algorithms later in Section 6.4.

Fig. 3 shows the hit-precision of all 9 algorithms for using Weibo's app-level trajectories to de-anonymize the ISP trajectories. For each external trajectory, its candidate trajectories are limited to these who have "encountered" with it among the 2,161,500 ISP trajectories, i.e., have spatio-temporal points at the same region within the same time-bin. The hit-precision is plotted as the function of the number of records in app-level trajectories, where we set k in hit-precision as 10. As shown in Fig. 3, de-anonymization algorithms based on users' mobility patterns (e.g., MKV and HIST) have the worst performance with the maximum hit-precision less than 8 percent. On the other hand, algorithms based on meeting events including ME and POIS have better performance, with the maximum hit-precision about 11 percent. Algorithms such as NFLX and MSQ achieve a better performance. Even so, their maximum hit-precision is only about 20 percent, which means even for users whose external trajectories have sufficient records, existing de-anonymization algorithms can only de-anonymize less than 20 percent of users based on the top-1 candidate trajectories. In addition, from the perspective of expectations, the true matched trajectories are ranked near the 8th position on average. The hit-precision of existing de-anonymization algorithms is far from the privacy bound obtained in Section 4.1, i.e., 5 points can uniquely identify over 75 percent users.

Note that in our experiment, datasets are already "matched"—the user population of the external dataset is already a subset of users in the target ISP dataset. This means for each trajectory in the external datasets, we know that there must be a trajectory in the ISP dataset. In practice, the attack is likely to be more difficult since the external dataset may contain users that are not in the ISP dataset (i.e., extra noise). To this end, our results are likely to represent the upper-bound performance of the de-anonymization algorithms. Next, we further investigate the reasons behind the under-performance.

5 Reasons Behind Underperformance

5.1 Spatio-Temporal Mismatch

We start by investigating the potential spatio-temporal mismatches between trajectories in different datasets. Fig. 4 shows the distribution of spatio-temporal mismatches of external datasets with respect to the ISP dataset. More specifically, for a given user, we match her trajectory in the external dataset with her ISP trajectory. We define a spatial mismatch as the geographical distance between two data records (from two trajectories) that fall into the same time slots. Similarly, we define a temporal mismatch as the minimum time interval between the external record and the ISP record at the same location region. Note that we limit the temporal mismatch within 24 hours to eliminate the influence of the second visit to the same location.

Large Spatio-Temporal Mismatches. Figs. 4a, 4b and 4c show the complementary cumulative distribution functions of spatial mismatches of different datasets. We observe that the spatial mismatches are prevalent. More than 37 percent of the records in the app-level trajectory data of Weibo have spatial mismatches over 2 km. It is similar in the other service, Dianping, of which the spatial mismatch of over 31 percent of the records are larger than 2 km. We also observe that the distribution of Weibo's app-level data and Dianping's app-level data can be approximated by the power-law distribution in the range of 0 to 10 km. After 10 km, they can be approximated better by the exponential distribution. For Weibo's check-in data, the power-law part has longer range. The large spatial mismatches can cause problems to de-anonymization algorithms that rely on exact location matching [37], [38].

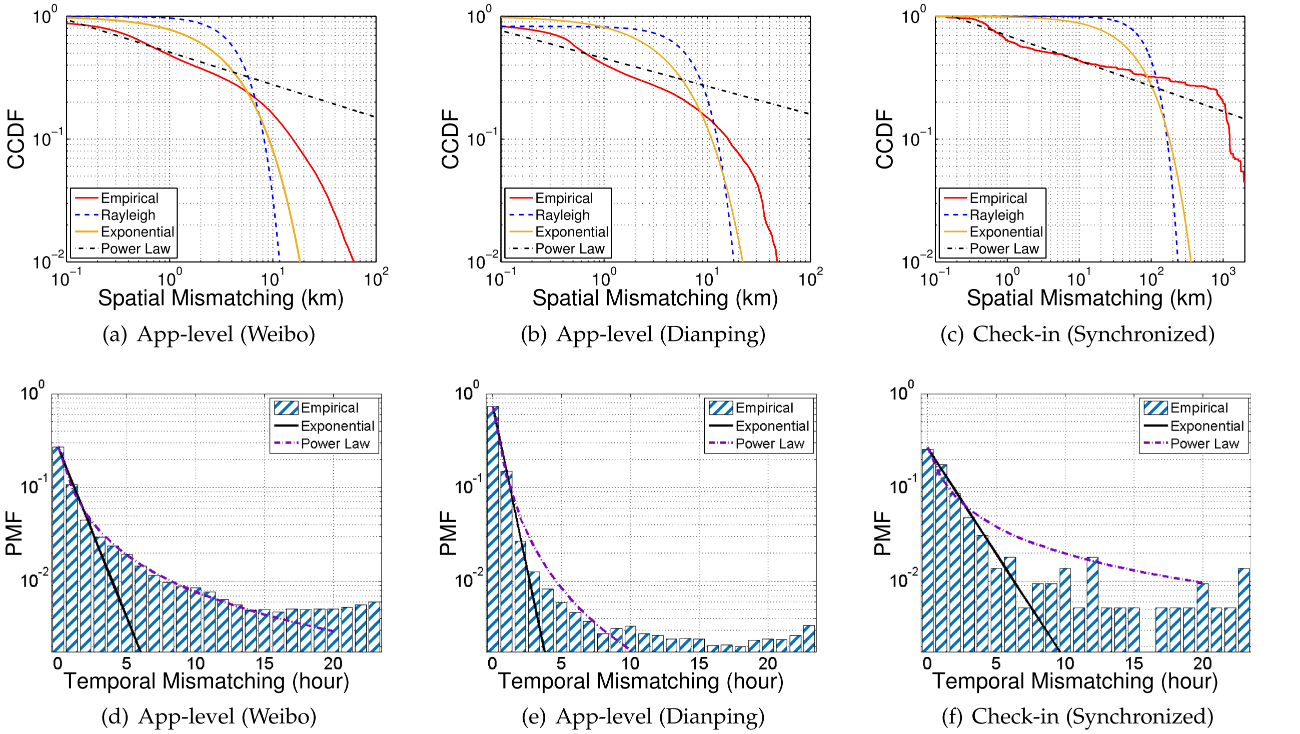
Figs. 4d, 4e and 4f show the probability mass function (PMF) of temporal mismatches. The temporal mismatches are also very prevalent. Only 30 percent of Weibo's app-level location records are in the same time slot with their corresponding ISP records. The large temporal mismatches indicate that performing exact temporal matching will introduce errors to determine the co-location of users [9], [37]. Overall, we can observe significant spatial and temporal mismatches between different datasets collected from the same set of users.

Finally, we observe that the mismatches follow different types of distributions. For example, Fig. 4c show that the spatial mismatch of Weibo's check-in data can be approximated by the power-law distribution. For Dianping, the power-law distribution fits well for the head of the empirical distribution, but does not capture the tail. To this end, modelling the spatio-temporal mismatches requires a more general framework.

Possible Reasons behind the Mismatches. There are a number of possible reasons that can cause the mismatch. We discuss some of them below.

First, *inherent GPS errors*: it is well-known that the GPS system has intrinsic source of errors [4] such as satellite errors (ephemeris and satellite clock), earth atmosphere errors (ionosphere and troposphere), and receiver errors (frequency drift, signal detection time).

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.



(a) App-level (Weibo)

(b) App-level (Dianping)

(c) Check-in (Synchronized)

(d) App-level (Weibo)

(e) App-level (Dianping)

(f) Check-in (Synchronized)

Fig. 4. Distribution of the spatial and temporal mismatching (with the ISP traces). The empirical distribution is compared with the fitting results of Rayleigh, exponential, and power-law distributions.

Second, *GPS unreachable locations*: due to the coverage of satellite signal, GPS signal is not always available in certain areas such as indoor and underground [25]. For example, when a user is on a subway going through a tunnel, the GPS reading will be interrupted leading to corrupted trajectories. Meanwhile, the user's smartphone can still connect to the nearby base station, which can lead to spatio-temporal mismatches between the ISP and the app-level trajectories.

Third, *location updating mechanisms*: to save battery life, many mobile apps do not update user GPS frequently, especially when the device is sleeping [6]. The slightly outdated GPS can still be used for non-critical services (e.g., venue recommendation), but leads to inaccurate user trajectories, especially temporal mismatches between different trajectory datasets.

Fourth, *deployment of base stations*: the base stations are placed unevenly in the city. In the ISP trajectory dataset, we use the connected BS to estimate the user's location, which may cause the spatial mismatches, especially in areas where the base stations are sparse.

Fifth, *user behavior*: for the check-in dataset, mismatches may also come from special user behavior. According to recent measurement studies [47], [51], 39.9 percent check-ins (on Foursquare) are remote check-ins with over 500 meters away from users' actual GPS location. Users often check-in at a remote location (that they are not physically visiting) to earn virtual badges or compete with their friends. Users may also check-in a few hours later after they visited a venue [47], which causes significant temporal mismatches. These factors can lead to major mismatches between the check-ins and the ISP trajectories.

Finally, *repeated user mobility*: it is possible that a user visits one location several times within the same day, but the ISP trajectory and the external trajectory capture different events among them. Based on our definition, temporal mismatch is measured by the minimum time interval between the external record and the ISP record at the same location region. Thus, we can find obvious temporal mismatches in the trajectories of this user. However, different from temporal mismatches caused by reasons such as location updating mechanisms and user behavior, these temporal mismatches actually do not come from the "errors" of the trajectories. However, a similar correlation between location records with time difference in different trajectories can be observed. Further, by elaborately modelling this phenomenon, the information contained can be utilized to improve the de-anonymization performance. Thus, we do not distinguish this factor from other "real" mismatches of trajectories.

Such spatio-temporal mismatches can lead to major errors for de-anonymization algorithms. However, many of the above factors cannot be fundamentally avoided in practice. To this end, de-anonymization algorithms should design adaptive mechanisms to tolerate these spatio-temporal mismatches.

5.2 Data Sparsity

Another possible reason for the under-performance of existing algorithms is high sparsity of the real-world mobility traces. In large-scale trajectory datasets, the vast majority of the users have very sparse location records. For example, in the ISP dataset, users on average have 62 records in a week, but 22.9 percent users have less than 1 records and 35.5 percent of the users have less than 2 records (Fig. 1). The external datasets (Weibo and Dianping) are even sparser with less than 5 records per user on average. This means that within the 1-hour time bins of the one-week period, the vast majority of the time bins are empty (with the location unknown). The high sparsity makes it difficult to accurately match trajectories across two datasets. This property is often

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

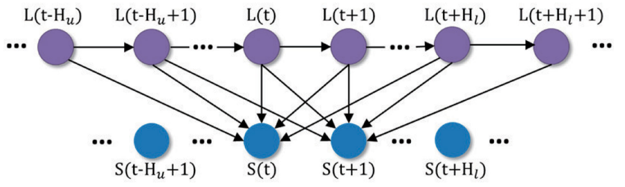


Fig. 5. Graphical model for L (ISP trajectory) and S (external trajectory).

overlooked when testing a de-anonymized algorithm on a synthetically generated dataset or a small dataset contributed by several hundreds of volunteers.

6 Our De-Anonymization Method

Inspired by the reasons of under-performance of existing algorithms, we propose new de-anonymization algorithms by addressing practical factors such as spatio-temporal mismatches and data sparsity. First, to address the spatio-temporal mismatches, we develop a Gaussian mixture model (GMM) to estimate and amend both spatial and temporal mismatches. The parameters of GMM are flexible and can be optimized according to specific datasets. Second, to address the data sparsity issue, we propose two other methods. a) We propose a *Markov-based* per-user mobility model to estimate the distribution of a given user's missing locations in the "empty" time slots of the trajectory; b) We leverage the whole dataset to aggregate global location contexts and time context features to further infer the missing location records.

Our proposed algorithms combine Gaussian mixture model and Markov model. We refer the algorithm as *GM*. Specifically in our model, each spatio-temporal point in the ISP trajectory $\boldsymbol{L}$ and external trajectory $\boldsymbol{S}$ are regarded as a random variable. Further, we show the graphical model of variables in Fig. 5. Each arrow in the graphical model indicates a dependency, which is modelled by combining Gaussian mixture model and Markov model. Finally, based on the probabilistic model, we define their similarity score function as follows,

$$D_{\mathrm{GM}(\boldsymbol{S},\boldsymbol{L})=\log p(\boldsymbol{S}|\boldsymbol{L}).} \qquad (13)$$

In this section, we will introduce how to compute this probability-based similarity score to de-anonymize location trajectories. Specifically, we first introduce how we model the spatio-temporal mismatches based on GMM in Section 6.1. Then, we introduce how we model users' mobility in Section 6.2. After that, we extend our proposed algorithm by considering time context, i.e., the information contained in the "empty" time bins of the trajectory in Section 6.3. Finally, we extensively evaluate the performance of our proposed de-anonymization algorithms in Section 6.4.

6.1 Modelling Spatio-Temporal Mismatches: Gaussian Mixture Model (GMM)

In this section, we focus on modelling the spatio-temporal mismatches between the ISP trajectory and external trajectory, which is described by arrows between ISP location records (purple nodes) and external location records (blue nodes) in Fig. 5. Due to the existence of temporal mismatches, each external location record $S(t)$ is dependent with not only the ISP location record in the same time bin $L(t)$, but also ISP location records in other time bins. Thus, in Fig. 5, there exist arrows between $S(t)$ and $L(t-H_u)$ to $L(t+H_l)$, where H_u and H_l are the maximum tolerant temporal mismatch in two time directions. Further, given a fixed temporal mismatch p, i.e., given the fact that $S(t)$ and $L(t-p)$ correspond to the same event, there also exists spatial mismatch between them. Thus, we model the difference between $S(t)$ and $L(t-p)$ as a Gaussian distribution $\mathcal{N}(\cdot|u_p,\Sigma_p)$. Overall, by considering temporal mismatches and spatial mismatches simultaneously, the probability distribution of $S(t)$ conditioned on $\boldsymbol{L}$ can be just described as a Gaussian mixture model [7], which will be introduced in detail in the following part of this section.

By definition, GMM is a finite linear superposition of Gaussian densities, which can be expressed as

$$p(x)=\sum_{k=1}^{K}\pi(k)\mathcal{N}(x|u_k,\Sigma_k),$$

where x is the random variable following GMM distribution. Each Gaussian density $\mathcal{N}(x|u_k,\Sigma_k)$ is called a component and has its own mean u_k and covariance Σ_k [7].

As for the conditional dependence structure between location records in the ISP trajectory and external trajectory of the same user shown in Fig. 5, we use component $\mathcal{N}(x|u_p,\Sigma_p)$ to model the probability density of external records with temporal mismatching of p time units. Then, let $\boldsymbol{L_C}$ represent the complete ISP trajectory, i.e., $\forall t\in\mathcal{T}$, $L_C(t)\neq\emptyset$. Conditioned on it, the probability density function (PDF) of an external record $S(t)$ can be calculated as,

$$p(S(t)|\boldsymbol{L})=\sum_{p=-H_l}^{H_u}\pi(p)\cdot\mathcal{N}(S(t)|L(t-p),\sigma^2(p)I_2), \qquad (14)$$

where $\pi(p)$ is the probability of the temporal mismatch to be p time units, and $\sigma(p)$ is the root mean square of the spatial distance between locations of trajectory $\boldsymbol{S}$ and $\boldsymbol{L}$ conditioned on the temporal mismatch of p time units. Since $S(t)$ and $L(t)$ are represented by geographical longitudes and latitudes, which are 2-dimensional vectors, I_2 is a 2×2 identity matrix. Note that the GPS coordinates have been mapped into discrete regions, and we use continuous distributions as an approximation to their probability mass function. Because we mainly focus on the performance of the similarity score function in terms of the rank, this approximation is reasonable.

In addition, H_u and H_l are the maximum tolerant temporal mismatch in two time directions. Specifically, in our problem, we only consider time delay in adversary's knowledge. Thus, we set H_l to be zero and denote H_u as H for simplicity. Parameters $\pi(p)$ and $\sigma(p)$ in (14) of our proposed model can be chosen by the empirical values shown in Fig. 4. On the other hand, they can also be estimated by EM algorithm [7] based on the ground-truth location records. Specifically, given M external records $\{S_1,\ldots,S_M\}$ with their corresponding H_u+H_l+1 ISP records in neighboring time slots, e.g., for S_n, its neighboring ISP records are $(L_{n,-H_l},\ldots,L_{n,H_u})$. In addition, we define z_{nk} as the latent variable to indicate whether S_n is generated by L_{nk} (corresponding

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

temporal mismatch is k time units). Thus, we have $\sum_{h=-H_l}^{H_u} z_{nk} = 1$. Then, in the E step of EM algorithm, we calculate the distribution of z_{nk} conditioned on the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$, which can be expressed as follows,

$$\gamma(z_{nk}) := P(z_{nk} = 1) = \frac{\pi(k)\mathcal{N}(S_n|L_{nk}, \sigma^2(k)I_2)}{\sum_{j=-H_l}^{H_u} \pi(j)\mathcal{N}(S_n|L_{nj}, \sigma^2(j)I_2)}.$$

In the M step, we re-estimate the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ using the distribution of z_{nk}, which can be expressed as follows,

$$\begin{cases} \pi(k) = \frac{1}{M}\sum_{n=1}^{M} \gamma(z_{nk}), \quad k = -H_l, \ldots, H_u, \\ \sigma^2(k) = \frac{1}{2M}\sum_{n=1}^{M} \gamma(z_{nk})|S_n - L_{nk}|^2, \quad k = -H_l, \ldots, H_u. \end{cases}$$

Then, by a finite number of repeating E and M step, we obtain the value of $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$.

6.2 Modelling User Mobility: Markov Model

Based on the graphical model shown in Fig. 5, we can observe that conditioned on a completely observed ISP trajectory $\boldsymbol{L}$, $S(t)$ for different t is independent with each other. Then probability density function of a full trajectory in external dataset can be calculated as follows,

$$p(\boldsymbol{S}|\boldsymbol{L}) = \prod_{S(t)\neq\emptyset} p(S(t)|\boldsymbol{L}). \tag{15}$$

However, from the analysis in Section 3.6, we can observe that users' locations in many time slots are missing, i.e., $L(t) = \emptyset$ for many $t \in \mathcal{T}$. In the case, (14) cannot be applied directly. In addition, $S(t)$ for different t also becomes dependent with each other. Thus, (15) cannot be applied. To solve it, we enumerate all possible complete trajectories of $\boldsymbol{L}$, and apply the formula of total probability with respect to them. Specifically, denote $\mathcal{C}(\boldsymbol{L})$ as the set of all possible complete trajectories of $\boldsymbol{L}$. Then the PDF of $S(t)$ conditioned on $\boldsymbol{L}$ can be calculated as follow:

$$p(\boldsymbol{S}|\boldsymbol{L}) = \sum_{\boldsymbol{L_C}\in\mathcal{C}(\boldsymbol{L})} p(\boldsymbol{L_C}|\boldsymbol{L}) \prod_{S(t)\neq\emptyset} p(S(t)|\boldsymbol{L_C}). \tag{16}$$

where external records $S(t)$ of different t are assumed to be statistically independent conditioned on each complete trajectory $\boldsymbol{L_C} \in \mathcal{C}$. The independence can be derived from the graphical model shown in Fig. 5. On the other hand, as we can observe from Section 3, the trajectories in external dataset are obviously sparser than those in the ISP dataset. It indicates that in real external trajectories, for each pair of adjacent non-empty records $S(t_1)$ and $S(t_2)$, we usually have $|t_1 - t_2| \gg H$. Thus, we can assume that external records are independent regardless of whether their dependent ISP records are observed.

As for the probability $p(\boldsymbol{L_C}|\boldsymbol{L})$, we calculate it by using a Markov model. Specifically, we use two different orders, i.e., 0-order and 1-order, Markov models as follows.

0-Order Markov Model. In the 0-order Markov model, location of each time slot is assumed to be independent with each other. Denote $E(r)$ as the marginal distribution of the user's ISP trajectory, which can be calculated as follows,

$$E(r) := p(L(t) = r) = \frac{\sum_{t\in\mathcal{T}} I(L(t) = r) + \alpha_r}{\sum_{t\in\mathcal{T}} I(L(t) \neq \emptyset) + \sum_{r\in\mathcal{R}} \alpha_r},$$

where $I(\cdot)$ is defined to be an indicator function of the logical expression with $I(true) = 1$ and $I(false) = 0$. In addition, α_r is the parameter to eliminate zero probabilities. For example, in Laplace smoothing [28], α_r is set to be the same value for different r. In our work, we use the location context to implement the smoothing as follow,

$$\alpha_r = \alpha_0 \cdot \sum_{v\in\boldsymbol{V}} \sum_{t\in\mathcal{T}} I(L_v(T) = r), \tag{17}$$

where α_r is in proportion to the number of records at location r with α_0 as the parameter to adjust the influence of location context.

Based on these definitions, the probability of a complete trajectory $L_C \in \mathcal{C}(\boldsymbol{L})$ conditioned on $\boldsymbol{L}$ can be calculated as follows,

$$p(\boldsymbol{L_C}|\boldsymbol{L}) = \prod_{t\in\mathcal{T}, L(t)=\emptyset} E(L_C(t)). \tag{18}$$

1-Order Markov Model. In the 1-order Markov model, location of each time slot is assumed to be dependent on the location in the last time slot. Denote $T(r_1, r_2)$ as the transition matrix of the user, which can be calculated as follows,

$$\begin{aligned} T(r_1, r_2) &:= p(L(t+1) = r_2 | L(t) = r_1), \\ &= \frac{\sum_{t\in\mathcal{T}} I(L(t) = r_1)I(L(t+1) = r_2) + \beta_{r_1r_2}}{\sum_{t\in\mathcal{T}} I(L(t) = r_1)I(L(t+1) \neq \emptyset) + \sum_{l_2\in\mathcal{R}} \beta_{r_1l_2}}. \end{aligned}$$

Similarly, $\beta_{r_1r_2}$ is the parameter to eliminate zero transition probabilities. We also use the aggregate transition statistics of users to help modelling users with sparse data, which can be represented as follows,

$$\beta_{r_1r_2} = \beta_0 \cdot \sum_{v\in\boldsymbol{V}} \sum_{t\in\mathcal{T}} I(L_v(t) = r_1) \cdot I(L_v(t+1) = r_2), \tag{19}$$

Then, we have

$$p(\boldsymbol{L_C}|\boldsymbol{L}) = \frac{1}{P(\boldsymbol{L})} \prod_{t\in\mathcal{T}} T(L_C(t), L_C(t+1)),$$

where $P(\boldsymbol{L})$ is a normalization constant to make the total probability equal to one, i.e., $\sum_{\boldsymbol{L_C}\in\mathcal{C}(\boldsymbol{L})} p(\boldsymbol{L_C}|\boldsymbol{L}) = 1$.

Further, based on our assumption of independence of external records discussed in (16), the computational complexity can be reduced by only considering the dependent sub-trajectory of L_C for each $S(t)$. Taking 0-order Markov model for example, for each $S(t)$, we only consider possible value of $L_C(t-p)$ for $p \in \{0, 1, \ldots, H\}$. Thus, we have

$$p(S(t)|\boldsymbol{L}) = \sum_{p=0}^{H} \sum_{r\in\mathcal{R}} p(L_C(t-p) = r|\boldsymbol{L}) \cdot \pi(p)\mathcal{N}(S(t)|r, \sigma^2(p)I_2),$$

where $p(L_C(t-p) = r|\boldsymbol{L})$ is the probability of a record at location r in time slot $t-p$ in the user's complete trajectory, which can be represented as follows,

$$p(L_C(t-p) = r|\boldsymbol{L}) = \begin{cases} E(r), & L(t-p) = \emptyset, \\ 1, & L(t-p) = r, \\ 0, & \text{otherwise}. \end{cases}$$

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

By this way, the complexity can be reduced from $O(T \cdot R^H)$ to $O(T \cdot R \cdot H)$, which is similar for 1-order Markov model. In addition, we will analyze and discuss the influence of ignoring dependency of external records in Section 9.

6.3 Modelling Time Context

In previously proposed methods, we calculate the probability $p(\boldsymbol{S}|\boldsymbol{L})$ by only considering the observed records in $\boldsymbol{S}$ such that $S(t) \neq \emptyset$ as shown in (15), and ignoring the unobserved time slots t with $S(t) = \emptyset$. However, this equation holds only when records in $\boldsymbol{S}$ and $\boldsymbol{L}$ are generated independently, which is not true in practice. For example, when a person is using cellular phone, the location will be requested by some applications with a larger probability. Similarly, when a user shares a check-in, it is more likely to access Internet in the near time (e.g., navigation services, location-based services). The consequence here is that spatio-temporal records in different datasets are not generated independently. Thus, in order to calculate the conditional probability $p(\boldsymbol{S}|\boldsymbol{L})$ more accurately, we need to consider the similarity score in terms of correlation of record generation in different datasets.

Specifically, we focus on whether there exists a record at time slot t in $\boldsymbol{S}$ and $\boldsymbol{L}$ while ignoring their concrete value. Thus, we define the 0-1 variable I_x to indicate whether x equals to $\emptyset$, i.e., if $x = \emptyset$ then $I_x = 0$; otherwise $I_x = 1$. Then, the similarity score can be expressed as

$$\begin{aligned} D_{\mathrm{B}}(\boldsymbol{S}, \boldsymbol{L}) :=& \log \prod_{t \in \mathcal{T}} P(I_{S(t)} | I_{L(t)}) \\ =& \sum_{\eta, \chi \in \{0,1\}} (1 - |I_{S(t)} - \eta|)(1 - |I_{L(t)} - \chi|) \log P_{\eta|\chi}, \end{aligned}$$

where the correlation is characterized by four parameters $P_{1|1}$, $P_{1|0}$, $P_{0|1}$, and $P_{0|0}$. For example, $P_{0|1}$ represents the probability of $S(t)$ to be $\emptyset$ under the condition of $L(t) \neq \emptyset$. Then, the combined similarity score can be calculated as

$$D_{\mathrm{GM-B}} = D_{\mathrm{GM}} + D_{\mathrm{B}}.$$

We refer to this upgrade version of GM algorithm as the GM-B algorithm. However, different with $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ in GMM, which can be set to be empirical value, parameters of $P_{x|x}$ highly depend on the ground truth data. For the same reason, the GM-B algorithm can only be used when there is a thorough understanding of the dataset (e.g., sufficient ground truth data to train the parameters). Thus, GM-B algorithm shows the best performance that can be achieved based on our proposed method, while GM algorithm shows the performance when we do not have sufficient ground truth data.

6.4 Performance Evaluation

In this section, we evaluate our algorithms compared with baseline methods on different trajectory datasets. In addition, we vary key parameters and experiment settings to examine the robustness of the proposed algorithms.

6.4.1 Baseline Algorithm

For baseline comparisons, except for the 9 major attacking algorithms, we also propose two simplified versions of our proposed algorithms which only consider spatial mismatches and temporal mismatches, respectively. We refer to them as spatial matching algorithm and temporal matching algorithm.

Spatial Matching Algorithm (SM). The SM algorithm ignores the mismatch in temporal dimension, and only matches records at the same time slot with Gaussian distribution. Then, its similarity score can be defined as

$$D_{\mathrm{SM}}(\boldsymbol{S}, \boldsymbol{L}) = \log \prod_{S(t) \neq \emptyset} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(S(t) - L(t))^2}{2\sigma^2} \right).$$

Similarly with GM algorithm, when $L(t)$ is $\emptyset$, the marginal distribution is used to estimate the PDF of $S(t)$.

Temporal Matching Algorithm (TM). On the contrary, the temporal matching algorithm only matches locations by regions, and it sums the weighted minimum time interval to obtain the similarity score as follows,

$$D_{\mathrm{TM}}(\boldsymbol{S}, \boldsymbol{L}) = \sum_{S(t) \neq \emptyset} \pi(\arg\min\nolimits_{p \in \mathcal{T}, S(t) = L(p)} |t - p|).$$

Specifically, we use empirical temporal mismatch distribution shown in Fig. 4 as $\pi(t)$.

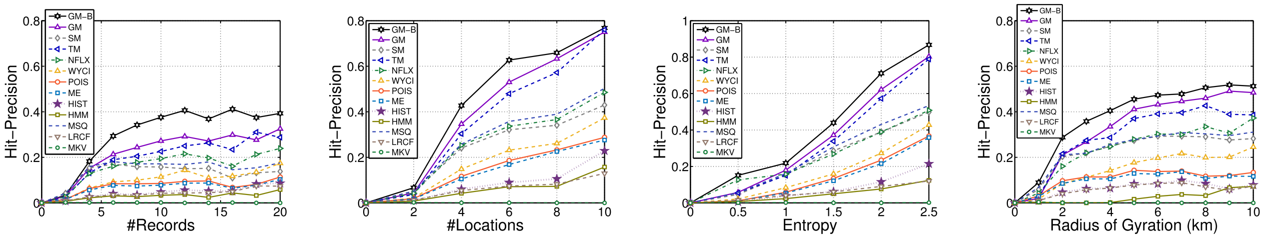
6.4.2 Experimental Settings

Our proposed algorithms require value of parameter $\pi(p)$ and $\sigma(p)$ describing the spatio-temporal mismatches between trajectory datasets to be matched. For de-anonymization based on Weibo's and Dianping's app-level trajectories, we randomly select 5,000 external location records with their time-adjacent ISP location records to estimate these parameters based on EM algorithm. As for de-anonymization based on Weibo's check-in trajectory dataset, since the number of ground truth user is very limited (503 users for Weibo's synchronized check-in dataset), we set $\pi(p)$ based on the distribution of temporal mismatches shown in Fig. 4f, and set $\sigma(p)$ to be 0.5 kilometer for each p. As for the parameters of location context, we set $\alpha_0 = |\mathcal{R}|/N$ and $\beta_0 = |\mathcal{R}|^2/N$, where N is the total number of records in the ISP dataset. Then, we calculate these parameters based on (17) and (19). As for parameters of time context, they are estimated based on location records with randomly sampled time-bins, of which the number is the same with that used in estimating $\pi(p)$ and $\sigma(p)$. For other state-of-the-art algorithms based on "concurrent" or "co-located" events, we set the spatial and temporal resolution as 1 km and 1 hour, respectively. In addition, we set k as 10 in hit-precision by default.

6.4.3 Experimental Results

De-anonymization using Weibo's App-level Trajectories. As a primary experiment, we evaluate the performance of different algorithms by using Weibo's app-level trajectories as the external information to de-anonymize the ISP dataset. This experiment corresponds to the attack of company-level adversaries. Specifically, we implement de-anonymization algorithms by using all the 56,683 Weibo's app-level trajectories as external information. Then, the hit-precision is calculated as functions of different metrics of external trajectories shown in Fig. 6, including number of records, number of

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.



(a) # of records in app-level trajectories (b) # of locations in app-level trajectories (c) Entropy of app-level trajectories (d) Radius of gyration of app-level trajectories

Fig. 6. Hit-precision of different de-anonymization algorithms using Weibo's app-level trajectories as the external information (company-level attacker).

distinct locations, trajectory entropy, and radius of gyration of the external trajectories.

Fig. 6a shows that SM algorithm does not perform better than existing algorithms, especially compared with those tolerating spatio-temporal mismatches, e.g., NFLX and MSQ. On the other hand, TM algorithm shows a higher hit-precision than SM algorithm, indicating tolerating temporal mismatches is more important than tolerating spatial mismatches in de-anonymization attacks. The intuition is that spatial mismatches are bounded by the strong locality of human movements, while temporal mismatches are not physically bounded.

In addition, we find that GM algorithm (modelling both spatial and temporal mismatches) achieves much better results. The hit-precision of GM is 10 percent higher compared with existing algorithms. Finally, by comprehensively modelling users' behavior, GM-B algorithm achieves another significant performance gain (7 percent hit-precision). Overall, a large number of records help to improve the de-anonymization accuracy. The best hit-precision of our proposed algorithm achieves 41 percent for external trajectories with more than 10 records, improving over 72 percent compared with the existing algorithms.

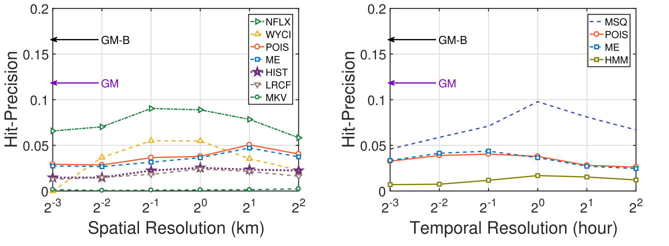
We notice that after the number of records get higher than 10, the performance gain stalls. In Fig. 6b, we directly show the relationships between the hit-precision with the number of distinct locations of external trajectories. The results show a very different trend: the hit-precision is rapidly growing with the number of distinct locations. For external trajectories with about 10 distinct locations, we can de-anonymize the corresponding ISP trajectory with the best hit-precision over 77 percent.

From Fig. 6d, we can observe the best hit-precision in terms of radius of gyration only achieve 52 percent. Compared with Figs. 6b and 6c, the result indicates that trajectory entropy is more dominating factors in the de-anonymization attack.

As mentioned in Section 4.2, POIS, WYCI, HIST, ME, NFLX, LRCF and MKV are based on "co-located" events. These algorithms are likely to be sensitive to spatial mismatches and even spatial resolutions. To be fair for these algorithms, we examine their hit-precision under different spatial resolutions (temporal resolution is set to the default value 1 hour). For comparison purposes, we also mark the hit-precision of GM and GM-B in the figures (using default 1 hour and 1 km). As shown in Fig. 7a, most algorithms, i.e., NFLX, WYCI, LRCF and HIST, achieve their highest hit-precision under our default spatial resolution of 1 km, while POIS and ME algorithms achieve their highest hit-precision under the spatial resolution of 2 km. Our proposed algorithms still outperform existing algorithms, i.e., the GM and GM-B algorithms improve the mean hit-precision by 31.6 and 83.8 percent relative to the best hit-precision of existing algorithms respectively.

Similarly, POIS, HMM, ME and MSQ are based on "concurrent" events, making them potentially sensitive to temporal resolutions. Fig. 7b shows their hit-precision of under different temporal resolution (spatial resolution is set to default 1 km). The result shows that HMM and MSQ algorithms achieve their highest hit-precision under our default temporal resolution of 1 hour, while POIS and ME achieve their highest hit-precision under the temporal resolution of 30 min. Our proposed algorithms still outperform existing algorithm, e.g., performance gap of GM and GM-B algorithms are 21.6 and 69.9 percent relative to the best existing algorithm respectively.

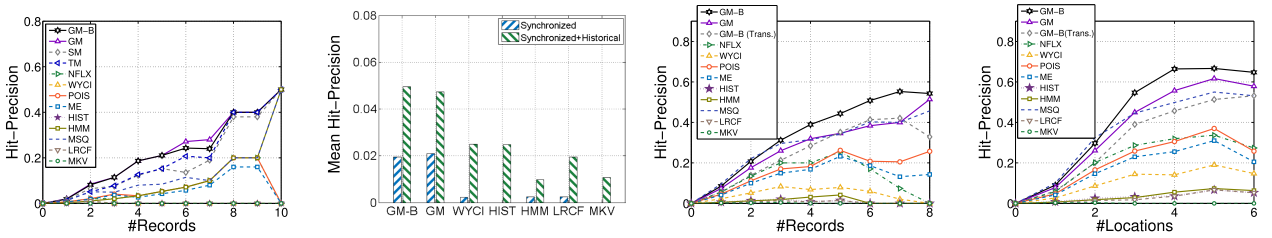
Validation using Weibo Check-in Trajectories. To validate our observations, we further evaluate the performance of our proposed algorithms using Weibo's check-in trajectories as external information. This experiment corresponds to the attack of individual-level adversaries. We first focus on the 503 check-in trajectories that have at least 1 records at the same time-window with the ISP dataset. The hit-precision is shown as the function of the number of records of check-in trajectories in Fig. 8a. Not too surprisingly, we can observe that individual-level adversaries are not as powerful as company-level adversaries, i.e., the hit-precision of de-anonymization using Weibo's check-in trajectories shown in Fig. 8a is obviously worse than that of using Weibo's app-level trajectories shown in Fig. 6a. In addition, more records in check-in trajectories help to improve the de-anonymization accuracy. Our proposed GM and GM-B algorithm outperform other algorithms. The largest performance gap between our proposed algorithms and existing



(a) Impact of spatial resolution (b) Impact of temporal resolution

Fig. 7. Impact of spatial and temporal resolution.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.



(a) Check-in (Synchronized) (b) Check-in (Historical) (c) App-level (Dianping) (d) App-level (Dianping)

Fig. 8. Hit-precision of different de-anonymization algorithms using Weibo Check-in trajectories (individual-level attacker) and Dianping (company-level attacker) as the external information.

algorithms achieves about 20 percent when there are 8 records in the check-in trajectories.

Fig. 8b shows the mean hit-precision of de-anonymization based on synchronized and historical Weibo check-ins. The mean hit-precision is very low because the synchronized check-ins are extremely sparse. For example, as shown in Fig. 1, over 80 percent users have less than 2 records. The historical check-ins have more data points but can no longer use the "encountering event" to match with the ISP data, leading to a low hit-precision. In addition, the historical check-ins can help to improve the de-anonymization accuracy for certain algorithms (e.g., WYCI, HIST, LRCF and our proposed GM, GM-B algorithms). Therefore, we only show their mean hit-precision of using historical check-ins versus not using them. Clearly, utilizing the historical check-in improves the hit-precision of all the algorithms. Intuitively, historical check-ins can greatly mitigate the sparsity issues of synchronized check-in trajectories.

Validation using Dianping Trajectories. Finally, we apply our algorithms to de-anonymize the ISP dataset using the 45,790 app-level trajectories from Dianping as the external information, which represents another company-level adversary. This experiment has two purposes. First, to use Dianping's dataset to evaluate the performance of our algorithms. Second, to simulate the scenario where ground-truth is not available to train the GM-B algorithm. Here, we assume the attacker does not have the ground-truth data from Dianping to estimate the parameters for the GM-B algorithm. Instead, we transfer the parameters estimated from the Weibo dataset to the Dianping experiment (transferred GM-B). As shown in Figs. 8c and 8d, the transferred GM-B has a competitive performance with the best existing algorithm and GM algorithm with parameters learnt from Dianping trajectory data. The result shows the robustness of our proposed algorithm.

In summary, we demonstrate that de-anonymization attack can be more effective by tolerating spatial and temporal mismatching (GM algorithm), and modeling the user behavior in terms of time context of the given service (GM-B algorithm). Specifically, the total performance gain in terms of hit-precision is more than 17 percent compared with the existing algorithms. Further, by adding historical check-ins and location context, another 30 to 150 percent relative gain can be achieved. In addition, the result suggests that even without ground-truth data to estimate parameters, our proposed algorithms will stay robust using parameters transferred from other external datasets. We also show that the proposed algorithms are robust against other parameter settings of the models, which can be found in Section 9. Overall, all these results confirm the usefulness of our insights.

7 Mismatch-Aware Location-Privacy Preserving Mechanism

In the last section, we have demonstrated that by considering the spatio-temporal mismatches between the ISP trajectories and external trajectories, we can bridge the gaps between theoretical bound and practical attacks for the location trajectory de-anonymization problem. Then, the coming question is whether we can utilize the spatio-temporal mismatches to better protect users' privacy. In this section, we focus on protecting users' location privacy by utilizing the spatio-temporal mismatches. Specifically, we develop a mismatch-aware perturbation mechanism in Section 7.1. Then, we develop a mismatch-aware location hiding mechanism in Section 7.2. Finally, we evaluate our proposed location-privacy preserving mechanisms in Section 7.3.

7.1 Mismatch-Aware Perturbation Mechanism

We aim to design a perturbation mechanism that utilizes the spatio-temporal mismatches to better protect users' privacy against de-anonymization attacks. For ISPs, obtaining the external trajectories $S_u(t)$ of all users is unrealistic in practice. Thus, we assume that ISPs only know the distribution of the spatio-temporal mismatches between users' ISP trajectories and external trajectories, which can be obtained based on the external information of a small part of users.

Specifically, we still use the Gaussian mixture model as in (14) to model the distribution of spatio-temporal mismatches, which can be characterized by parameters $(\pi(p), \sigma(p))$ for $p \in \{-H_l, \ldots, H_u\}$. In addition, users' location records are unevenly distributed in time dimension for both the ISP dataset and the external dataset. Thus, we define $p_S(t)$ as the probability that the timestamp of a random location record is t in the external dataset, which can be calculated by

$$p_S(t) = N_S(t) / \sum_{i=1}^{T} N_S(i), \tag{20}$$

where $N_S(t) = |\{u \in \boldsymbol{U} | S_u(t) \neq \emptyset\}|$ is the number of non-empty location records at time t in the external dataset. Similarly, we define $p_L(t)$ as the probability that the timestamp of a random location record is t in the ISP dataset.

Based on these definitions, we now design the mismatch-aware perturbation mechanism. A direct idea is to add larger noise to spatio-temporal points in ISP trajectories with smaller mismatches between external trajectories, while keeping the perturbation strength unchanged. Here, we add random noise to spatio-temporal points of the ISP dataset at different time t with different standard deviation ξ_t. Then, our goal is to minimize the correlation between the

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

ISP trajectory and external trajectory belonging to the same user. We use similarity score (15) of our propose GM algorithm to characterize this correlation. Then, the problem can be expressed by an optimization problem as follows:

$$\begin{aligned}&\mathbf{min}\ E(p(S_u(t)|\boldsymbol{L}_u+\boldsymbol{\epsilon}_u))\\ &\mathbf{s.t.}\ \begin{cases}E(|\epsilon_u(t)|^2)=\xi_t^2, \quad t=1,\ldots,T,\\ \sum_{t=1}^{T}p_L(t)\cdot\xi_t^2=\xi_0^2,\end{cases}\end{aligned}\tag{21}$$

where $E(\cdot)$ denotes the expectation. In addition, $\boldsymbol{\epsilon}_u=(\epsilon_u(1), \epsilon_u(2),\ldots,\epsilon_u(T))$ is the added noise, where $\epsilon_u(t)\sim\mathcal{N}(0, \xi_t^2 I_2)$. In addition, based on (14), we have

$$\begin{aligned}&p(S_u(t)|\boldsymbol{L}_u+\boldsymbol{\epsilon}_u)\\ &=\sum_{p=-H_l}^{H_u}\frac{\pi(p)}{2\pi\sigma^2(p)}\exp\left(-\frac{|S_u(t)-(L_u(t-p)+\epsilon_u(t-p))|^2}{2\sigma^2(p)}\right).\end{aligned}$$

Since solving this optimization problem relies on the value of all the external spatio-temporal points $S_u(t)$, which is not available in practice, we make some simplifications to (21) and try to eliminate the influence of $S_u(t)$.

Specifically, we use the approximation $exp(x)\approx 1+x$ based on Taylor series in the target function of (21). In addition, since $E(\epsilon_t^u)=0$, we only need to consider the term with $|\epsilon_t^u|^2$. Finally, by removing terms irrelevant to the optimization variables ϵ_t^u and ξ_t and utilizing the first constraint in (21), the target function of (21) can be simplified as follows:

$$\begin{aligned}E(p(S_u(t)|\boldsymbol{L}_u+\boldsymbol{\epsilon}_u)) =&E\left(-\sum_{p=-H_l}^{H_u}\frac{\pi(p)}{\sigma^4(p)}|\epsilon_u(t-p)|^2\right)\right),\\ &=-\sum_{t=1}^{T}\sum_{p=-H_l}^{H_u}p_S(t)p_L(t-p)\frac{\pi(p)}{2\sigma^4(p)}\xi_{t-p}^2.\end{aligned}$$

Then, we can obtain the simplified optimization problem as follows:

$$\begin{aligned}&\mathbf{max}\ \sum_{t=1}^{T}\sum_{p=-H_l}^{H_u}\frac{p_S(t)\pi(p)}{\sigma^4(p)}\xi_{t-p}^2 p_L(t-p)\\ &\mathbf{s.t.}\ \sum_{t=1}^{T}\xi_t^2\cdot p_L(t)=\xi_0^2.\end{aligned}\tag{22}$$

This problem can be easily solved. However, the solution only has one non-zero element ξ_{t_0}, where $t_0=\arg\max_t \sum_{p=-H_l}^{H_u}\frac{p_S(t+p)\pi(p)}{\sigma^4(p)}$. In practice, adding noise only to a small part of spatio-temporal points is not reasonable. Thus, we let ξ_t^2 be proportional to this coefficient, which can be represented as follows:

$$\xi_t^2\propto\left(\sum_{p=-H_l}^{H_u}\frac{p_S(t+p)\pi(p)}{\sigma^4(p)}\right)^{\gamma},\tag{23}$$

where γ is a tunable parameter, and we set γ as 1 by default.

7.2 Mismatch-Aware Location Hiding Mechanism

In order to design a mismatch-aware location hiding mechanism, we consider eliminating ISP location records with more contribution to (14) with higher probability, while keeping the total number of eliminated location records unchanged, of which the corresponding optimization problem can be represented as follows:

$$\begin{aligned}&\mathbf{min}\ E(p(S_u(t)|\hat{\boldsymbol{L}}_u))\\ &\mathbf{s.t.}\ \begin{cases}E(z_u(t))=\lambda_t, \quad t=1,\ldots,T,\\ \sum_{t=1}^{T}p_L(t)\lambda_t=\lambda_h,\\ \hat{L}_u(t)=\begin{cases}\emptyset, & z_u(t)=1,\\ L_u(t), & z_u(t)=0,\end{cases}\end{cases}\end{aligned}\tag{24}$$

where $z_u(t)$ is a binary random variable to indicate whether record $L_u(t)$ is eliminated, and $\hat{L}_u$ is the trajectory after location hiding mechanism. Based on (14), the target function of (24) can be calculated as follows:

$$E(p(S_u(t)|\boldsymbol{L}_u))\propto-\sum_{t=1}^{T}\sum_{p=-H_l}^{H_u}\lambda_{t-p}p_L(t-p)p_S(t)\pi(p)C(t,p),$$

where $C(t,p)=E(\mathcal{N}(S_u(t)|L_u(t-p),\sigma^2(p)I_2))$ is independent of optimization variable λ_t. Considering that $C(t,p)$ describes the distribution of spatial mismatches, we assume it does not change much with time t and temporal mismatch p. Thus, location records at time t will be eliminated with the probability of λ_t expressed as follows:

$$\lambda_t\propto\left(\sum_{p=-H_l}^{H_u}\pi(p)p_S(t+p)\right)^{\delta},\tag{25}$$

where δ is a tunable parameter, and we set δ as 1 by default.

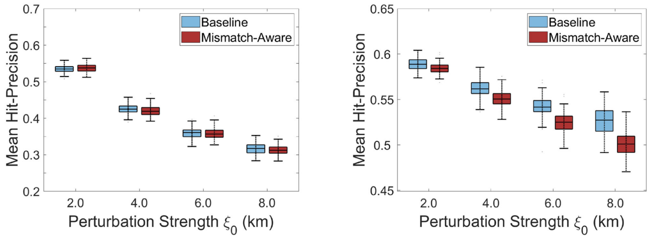
7.3 Performance Evaluation

In this section, we evaluate the performance of our proposed mismatch-aware perturbation mechanisms against de-anonymization attacks.

7.3.1 Baseline Methods

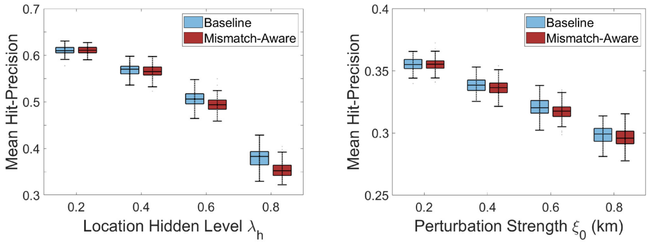
We compare our proposed mismatch-aware location hiding mechanism with the baseline location hiding mechanism that eliminates each location record with the same probability, where $\lambda_t=\lambda_h$ for all $t\in\mathcal{T}$. We first implement these mechanisms on the ISP trajectories, and then calculate the mean hit-precision of previous de-anonymization algorithms. A lower mean hit-precision indicates better performance of the location hiding mechanisms. Specifically, we consider two de-anonymization algorithms with the best performance including GM and NFLX algorithms. Similarly, as for our proposed mismatch-aware perturbation mechanism, we compare it with the baseline perturbation mechanism that adds noise with equal standard deviation to each location record, where $\xi_t=\xi_0$ for all $t\in\mathcal{T}$. Different with experiments of location hiding mechanisms, NFLX algorithm is based "co-located" events, and the performance of perturbation mechanisms against this algorithm is greatly influenced by the spatial granularity. Thus, we evaluate perturbation mechanisms against MSQ algorithm instead. Further, we define relative improvement of our proposed LPPMs as $\gamma=(\hat{h}_{\rm B}-\hat{h}_{\rm MA})/\hat{h}_{\rm B}$, where $\hat{h}_{MA}$ is the median of hit-precision of the given de-anonymization algorithm with

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.



(a) GM algorithm (b) MSQ algorithm

Fig. 9. Performance of our proposed mismatch-aware perturbation mechanism.



(a) GM algorithm (b) NFLX algorithm

Fig. 10. Performance of our proposed mismatch-aware location hiding mechanism.

our proposed mismatch-aware LPPMs, and $\hat{h}_{\mathrm{B}}$ is that of the baseline.

7.3.2 Experimental Settings

We compare the performance of different location-privacy preserving mechanisms on the ISP trajectory dataset with Weibo's app-level trajectories as external information. Different with previous experiments of de-anonymization attacks, of which the results are deterministic, the performance of location-privacy preserving mechanisms is stochastic. For example, it is influenced by the randomly drawn Gaussian noise. Thus, we repeat our experiments for 100 times with different random noise, and show the statistical results. On the other hand, since one external trajectory might have "encountered" with thousands of ISP trajectories, in order to reduce the computational time, for each external trajectory, we random select 50 negative ISP trajectories from those who have "encountered" with it as the candidate trajectory set, which is different from experiments in Section 6.4.

7.3.3 Experiment Results

Boxplots of GM algorithm and MSQ algorithm under the two perturbation mechanisms with different perturbation strength ξ_0 are shown in Fig. 9, where the box plots quartiles, and the band inside the box is the median. Larger reduction of de-anonymization performance (hit-precision) indicates better performance of the location-privacy preserving mechanism. From the results, we can observe that a large perturbation strength helps to protect users' privacy, i.e., reducing the hit-precision of de-anonymization attack. In addition, the mismatch-aware perturbation mechanism outperforms the baseline in most situations. Specifically, it reduces the hit-precision of GM algorithm by relative improvement $\gamma = 1.6\%$ with the perturbation strength of 8.0 km, demonstrating the correctness of the above theory. In addition, the proposed mismatch-aware perturbation mechanism also works on MSQ algorithm. It reduces the hit-precision of MSQ algorithm by relative improvement $\gamma = 5.0\%$ with the perturbation strength of 8.0 km.

In addition, from Fig. 9, we can observe that GM has worse performance than MSQ. The reason is that adding noise to the ISP trajectories by the perturbation mechanisms changes the distribution of spatio-temporal mismatches between the ISP trajectories and external trajectories, which is more destructive to GM algorithm. This performance degradation can be reduced by re-estimating parameters $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ based on trajectories after the perturbation mechanisms. However, we mainly focus on the upper bound of the performance of the perturbation mechanisms. Thus, we keep parameters in GM algorithm unchanged in the experiments. Overall, results demonstrate the effectiveness of the utilizing spatio-temporal mismatches in location-privacy preserving mechanisms.

Then, boxplots of the obtained hit-precision of GM algorithm and NFLX algorithm under two different location hiding mechanisms with different location hiding level λ_h are shown in Fig. 10. As we can observe, a large location hiding level also helps to reduce the hit-precision of de-anonymization attack. The mismatch-aware location hiding mechanism outperforms the baseline in most situations. The performance gain of the proposed mechanism is larger for higher location hidden level. Specifically, when location hiding level is 0.8, the relative improvement γ given GM algorithm and NLFX algorithm is 8.0 and 1.1 percent, respectively, indicating the effectiveness of the proposed location hiding mechanism.

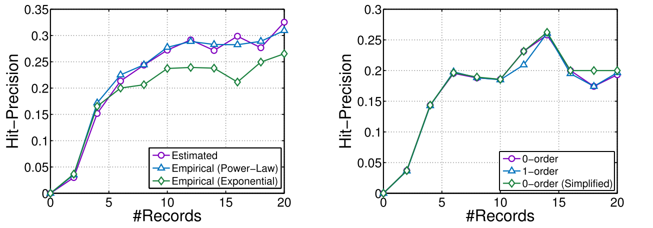
In summary, we demonstrate that by considering spatial and temporal mismatches, location-privacy preserving mechanisms can be more effective. Specifically, our proposed mismatch-aware perturbation mechanism and location hiding mechanism can reduce the performance of de-anonymization attacks by over 8.0 percent, demonstrating the usefulness of our insights. Though the relative improvement of 8.0 percent dose not solve the problem of protecting users' privacy essentially, it provides a new idea for designing location-privacy preserving mechanisms. With equal de-anonymization risk, based on our insights, we can add smaller perturbation or hide less location to the trajectory datasets to keep more utility of them.

8 Related Work

De-Anonymization Methods: Overview. There have been a number of de-anonymization algorithms proposed in recent years. These algorithms seek to re-identify users from anonymized datasets leveraging external information (not all the algorithms are applicable to location traces). We classify them into three main categories based on the utilized user data: *content* (user activities such as timestamps, location), *profile* (user attributes such as username, gender, age), and *network* (relationship and connections between users) [40]. Location trajectory data belongs to the "content" category.

De-anonymization of Location Trajectories. Focusing on the user *content*, a number of de-anonymization algorithms have been proposed [9], [11], [12], [26], [31], [32], [37], [38],

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.



(a) Mismatching distribution (b) Dependence between records

Fig. 11. Impact of assumptions to the performance of our proposed de-anonymization algorithms.

[39], [49]. Most of these algorithms can be directly applied or easily adapted to trajectory datasets. However, due to the lack of large scale ground-truth datasets (matched the ISP dataset and the external traces), existing works either focus on theoretical privacy bound [11], [49] or simulating de-anonymization attacks on synthetically generated datasets [11], [26], [38], [49]. Our work seeks to use a large scale ground-truth dataset to explore their empirical performance and identify practical factors (if any) that are often neglected by algorithm designers.

Specifically, some algorithms are designed to tolerate mistakes in the adversary's knowledge such as temporal mismatching [32] and spatial mismatching [26]. Other algorithms [12], [16], [31], [38], [39] implement de-anonymization attacks based on *individual user's* mobility patterns [31], [39]. Finally, researchers also develop de-anonymization algorithms based on "encountering" events [9], [37]. By considering the location context (e.g., user population density), it achieves a better performance [37]. However, no algorithm performs well under spatial and temporal mismatches simultaneously. In particular, no algorithm simultaneously considers both spatial and temporal mismatches.

De-Anonymization of Network/Profile Data. Since we focus on the de-anonymization of *location trajectory datasets*, we only briefly introduce the algorithms designed for *network* datasets [20], [22], [33], [36], [44] and *profile* datasets [17], [30] for completeness. Mudhakar et al. [44] and Ji et al. [20], [22] focused on de-anonymization based on users' graph/network structures. Zhang et al. [50] de-anonymized multiple social networks simultaneously by minimizing the friendship inconsistency of users. Nilizadeh et al. [34] proposed an enhanced de-anonymization algorithm by utilizing the community structure of social network. Qian et al. [36] focused on the theoretical de-anonymization gain with different background knowledge for social network de-anonymization. These algorithms can be adapted to de-anonymizing location trajectories by constructing a "contact graph" to model users' encountering with each other. However, these algorithms require using social network graphs as the external information, which are not available in our scenario. Thus, their approaches cannot be applied to solving our problem. On the other hand, algorithms designed for *profile* datasets [17], [30] (e.g., age, gender, language) are also not applicable to location trajectories.

Privacy Protection Mechanisms. Researchers have investigated different ways to anonymize user data to preserve privacy. The most common privacy models are k-anonymity [45], l-diversity [27] and t-closeness [24]. Related to these three models, a number of specific techniques have been proposed to anonymize location trajectory data. Osman et al. [2] proposed a technique to protect privacy by shifting trajectory points in space that are close to each other in time. Marco et al. [19] proposed an algorithm named GLOVE to grant k-anonymity of trajectories through specialized spatio-temporal generalization. Another work from Osman [1] developed a time-tolerant method. Simon et al. [35] provided two metrics, conditional entropy and worst-case quality loss, to evaluate the privacy protection mechanisms. Ji et al. [21], [23] systematically investigated existing graph data anonymization algorithms, utility metrics, and de-anonymization attacks. Based on these techniques, they further proposed a uniform and open-source secure graph data publishing system. Meyerowitz et al. [29] developed a system to anonymize location data in real time by sending the predicted future locations of multiple users simultaneously to location-based services. Recently, researchers also explore to apply differential privacy to location trajectory datasets [3], [5], [14]. For example, Andrés et al. [5] introduced geo-indistinguishability, which used criteria of differential privacy to make sure the user's exact location is unknown while keeping enough utility for certain desired service. Gergely et al. [3] studied an anonymization scheme to release spatio-temporal density data based on differential privacy. In our work, the definition of privacy is based on the uniqueness of user trajectories, whose privacy model is based on k-anonymity.

9 Discussion

Validation of Assumptions. We first examine the impact of parameter $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ in the GMM. In Figs. 8c and 8d, we have shown that our proposed algorithms stay robust using parameters transferred from other external datasets. The reason is that a number of intrinsic factors leading to spatio-temporal mismatches are similar with different external trajectory datasets, e.g., GPS errors, GPS unreachable locations, or even repeated user mobility. Thus, parameters that well capture these common factors are useful when transferred between different external datasets. In addition, instead of using parameters produced by the EM algorithm, we also try to apply different parameters from the empirical distribution fitting: $\sigma(p)$ is set to be 0.5 km for all p, and $\pi(p)$ is set to be the power-law or exponential distribution. Then, we compare their hit-precision in Fig. 11a. From the results, we find that GM algorithm using power-law empirical parameters outperforms the one using exponential empirical parameters. The result is consistent with our prior observation that Weibo's mismatches follow a power-law distribution. In addition, the hit-precision of using power-law empirical parameters is very close to that of the ground-truth parameters estimated by the EM algorithm. This indicates that our algorithm is robust—the performance does not depend on an accurate parameter estimation as long as the suitable distribution model is selected.

Next, we examine the impact of the order of Markov model and the dependence between external records, which is ignored in Section 6. We show the hit-precision of using 0-order Markov, 1-order Markov, and 0-order Markov with ignored dependency between external records in Fig. 11b. Specifically, we use 0-order (simplified) to represent the

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

GM algorithm with 0-order Markov mobility model and ignored dependency between external records. Parameters of π and σ are all set to be the value estimated by EM algorithm. As shown in Fig. 11b, very small difference of hit-precision can be observed between different settings, indicating that the order of Markov and dependency between external records have small impact on the hit-precision. The main reason is that external trajectories are very sparse so that we can ignore the dependence of different records of each external trajectory.

Finally, we discuss the impact of quantification to spatio-temporal mismatches. In the process of data collecting and publishing, the quantified effect is an important cause of spatial mismatches. As we mentioned in Section 5.1, when using the connected base stations to estimate the user's location, the sparser deployment of base stations will cause larger spatial mismatches. However, in the process of data processing for de-anonymization, the quantified effect in turn reduces the spatial mismatches. With larger spatial granularity, more users' location records of different trajectory datasets at the same time will be mapped into the same spatial region. Thus, more spatial mismatches are destructed. However, this process also reduces the information contained in each location record. Thus, as we can observe from Fig. 7a, when the spatial granularity is small, spatial mismatches are the main bottleneck of de-anonymization, and the hit-precision of all algorithms increases with the spatial granularity. However, when the spatial granularity is large (larger than 1 km in Fig. 7a), spatial mismatches are not the main bottleneck any more. In addition, larger spatial granularity will reduce the information contained in each location record. Thus, the hit-precision decreases with the spatial granularity. It also exhibits similar trend for temporal mismatches shown in Fig. 7b. Overall, mismatch ubiquitously exists in user generated data. Even for other types of external data, e.g., fake age and gender in user profile can be regarded as another type of mismatches. Thus, designer of de-anonymization method should keep vigilant about the impact of mismatches when dealing with practical problems.

Implications for Future Work. The main reason of neglecting the spatio-temporal of existing algorithms is the lack of such large-scale real-world ground truth dataset. Without it, the distribution of spatio-temporal mismatches cannot be characterized correctly. In addition, noise and mismatch ubiquitously exists in user generated data. The existence of spatio-temporal mismatches makes the de-anonymization attack harder, and existing de-anonymization algorithms that neglected spatio-temporal mismatches actually suffer from under-performance based on our analysis. However, our study also demonstrates the damage of spatio-temporal mismatches to the de-anonymization performance is not irreversible. By elaborately modelling the spatio-temporal mismatches, the de-anonymization performance can be significantly improved. Our work has key implications to de-anonymization algorithm designers by highlighting the key factors that matter in practice. For example, we show that temporal mismatches are more damaging than spatial mismatches. The intuition is that spatial mismatches are naturally bounded by the strong locality of human movements. To this end, having the algorithm tolerating temporal mismatches (or both) is the key. Overall, further work of de-anonymization method should keep vigilant about the impact of spatio-temporal mismatches when dealing with practical problems.

On the other hand, in order to provide better location privacy protections, practical factors should also be considered. Our result shows that user mobility patterns, location context, and time context all have helped the de-anonymization. This means it might be no longer sufficient to use simple mechanisms to manipulate the time and location points in the original trajectories. Privacy protection algorithms should consider the user, location, and time context to provide stronger privacy guarantees. We also show that the distribution of spatio-temporal mismatches can be utilized to better protect users' privacy against the de-anonymization attack. Future work of location-privacy preserving mechanisms can utilize the spatio-temporal mismatches to better preserve users' location privacy and keep more utility of the trajectory dataset at the same time.

10 Conclusion

In this work, we use two sets of large-scale *ground truth* mobile trajectory datasets to extensively evaluate commonly used de-anonymization methods. We identify a significant gap between the algorithms' empirical performance and the theoretical privacy bound. Further analysis then reveals the main reasons behind the gap: the algorithm designers often under-estimate the spatio-temporal mismatches in the data collected from different sources and the significant noises in user-generated data. Our proposed new algorithms that are designed to cope with these practical factors in both de-anonymization attack and location-privacy preserving mechanisms have shown promising performance, which confirms our insights. In future work, we plan to investigate de-anonymization attacks by considering other types of external information, e.g., social graphs [20], [22], [33], [44] or users' profile [17], [30], and the impact of other types of mismatches (e.g., fake user profile) on them.

Acknowledgments

This work was supported in part by The National Key Research and Development Program of China under Grant 2018YFB1800804, the National Nature Science Foundation of China under Grants U1936217, 61971267, 61972223, 61941117, and 61861136003, Beijing Natural Science Foundation under Grant L182038, Beijing National Research Center for Information Science and Technology under Grant 20031887521, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] O. Abul, F. Bonchi, and M. Nanni, "Anonymization of moving objects databases by clustering and perturbation," *Inf. Syst.*, vol. 35, no. 8, pp. 884–910, 2010.

[2] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *Proc. IEEE 24th Int. Conf. Data Eng.*, 2008, pp. 376–385.

[3] G. Acs and C. Castelluccia, "A case study: Privacy preserving release of spatio-temporal density in paris," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1679–1688.

[4] E. Akim and D. Tuchin, "GPS errors statistical analysis for ground receiver measurements," in *Proc. Int. Symp. Space Flight Dyn.*, 2003, pp. 16–20.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.

[5] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, “Geo-indistinguishability: Differential privacy for location-based systems,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2013, pp. 901–914.

[6] N. Banerjee, A. Rahmati, M. Corner, S. Rollins, and L. Zhong, “Users and batteries: Interactions and adaptive energy management in mobile systems,” *Proc. Int. Conf. Ubiquitous Comput.*, 2007, pp. 217–234.

[7] C. M. Bishop, “Pattern recognition,” *Mach. Learn.*, vol. 128, pp. 1–58, 2006.

[8] A. Boutet, S. B. Mokhtar, and V. Primault, “Uniqueness assessment of human mobility on multi-sensor datasets,” Res. Rep., LIRIS UMR CNRS 5205, 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01381986/document

[9] A. Cecaj, M. Mamei, and F. Zambonelli, “Re-identification and information fusion between anonymized CDR and social network data,” *J. Ambient Intell. Humanized Comput.*, vol. 7, no. 1, pp. 83–96, 2016.

[10] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, “Bridging the gap between physical location and online social networks,” in *Proc. 12th ACM Int. Conf. Ubiquitous Comput.*, 2010, pp. 119–128.

[11] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, “Unique in the crowd: The privacy bounds of human mobility,” *Scientific Reports*, vol. 3, 2013, Art. no. 1376.

[12] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel, “Identification via location-profiling in GSM networks,” in *Proc. 7th ACM Workshop Privacy Electron. Soc.*, 2008, pp. 23–32.

[13] P. Deville *et al.*, “Dynamic population mapping using mobile phone data,” *Proc. Nat. Acad. Sci. United States America*, vol. 111, no. 45, pp. 15 888–15 893, 2014.

[14] C. Dwork, “Differential privacy: A survey of results,” in *Proc. Int. Conf. Theory Appl. Models Comput.*, 2008, pp. 1–19.

[15] T. Fox-Brewster, *Now Those Privacy Rules Are Gone, This Is How ISPs Will Actually Sell Your Personal Data*, USA: Forbes, 2017.

[16] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, “Exploiting innocuous activity for correlating users across sites,” in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 447–458.

[17] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi, “On the reliability of profile matching across large online social networks,” in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1799–1808.

[18] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, “Understanding individual human mobility patterns,” *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[19] M. Gramaglia and M. Fiore, “Hiding mobile traffic fingerprints with GLOVE,” in *Proc. 11th ACM Conf. Emerg. Netw. Experiments Technol.*, 2015, Art. no. 26.

[20] S. Ji, W. Li, N. Z. Gong, P. Mittal, and R. A. Beyah, “On your social network de-anonymizablity: Quantification and large scale evaluation with seed knowledge,” in *Proc. Netw. Distrib. Syst. Security Symp.*, 2015.

[21] S. Ji, W. Li, P. Mittal, X. Hu, and R. A. Beyah, “SecGraph: A uniform and open-source evaluation system for graph data anonymization and de-anonymization,” in *Proc. 24th USENIX Conf. Security Symp.*, 2015, pp. 303–318.

[22] S. Ji, W. Li, M. Srivatsa, and R. Beyah, “Structural data de-anonymization: Quantification, practice, and implications,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2014, pp. 1040–1053.

[23] S. Ji, P. Mittal, and R. Beyah, “Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey,” *IEEE Commun. Surveys Tut.*, vol. 19, no. 2, pp. 1305–1326, Second Quarter, 2017.

[24] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in *Proc. IEEE 23rd Int. Conf. Data Eng.*, 2007, pp. 106–115.

[25] H. Liu, H. Darabi, P. Banerjee, and J. Liu, “Survey of wireless indoor positioning techniques and systems,” *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev*, vol. 37, no. 6, pp. 1067–1080, Nov. 2007.

[26] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao, “Privacy vulnerability of published anonymous mobility traces,” *IEEE/ACM Trans. Netw.*, vol. 21, no. 3, pp. 720–733, Jun. 2013.

[27] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “L-diversity: Privacy beyond k-anonymity,” *ACM Trans. Knowl. Discovery Data*, vol. 1, no. 1, 2007, Art. no. 3.

[28] C. D. Manning *et al.*, *Introduction to Information Retrieval*, vol. 1. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[29] J. Meyerowitz and R. Roy Choudhury, “Hiding stars with fireworks: Location privacy through camouflage,” in *Proc. 15th Annu. Int. Conf. Mobile Comput. Netw.*, 2009, pp. 345–356.

[30] X. Mu, F. Zhu, E. P. Lim, J. Xiao, J. Wang, and Z. H. Zhou, “User identity linkage by latent user space modelling,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1775–1784.

[31] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, “Where you are is who you are: User identification by matching statistics,” *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 2, pp. 358–372, Feb. 2016.

[32] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in *Proc. IEEE Symp. Security Privacy*, 2008, pp. 111–125.

[33] A. Narayanan and V. Shmatikov, “De-anonymizing social networks,” in *Proc. 30th IEEE Symp. Security Privacy*, 2009, pp. 173–187.

[34] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn, “Community-enhanced de-anonymization of online social networks,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2014, pp. 537–548.

[35] S. Oya, C. Troncoso, and F. Pérez-González, “Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms,” in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2017, pp. 1959–1972.

[36] J. Qian, X.-Y. Li, Y. Wang, S. Tang, T. Jung, and Y. Fan, “Social network de-anonymization: More adversarial knowledge, more users re-identified?” *ACM Trans. Int. Technol.*, vol. 19, no. 3, pp. 1–22, 2019.

[37] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, “Linking users across domains with location data: Theory and validation,” in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 707–719.

[38] L. Rossi and M. Musolesi, “It’s the way you check-in: Identifying users in location-based social networks,” in *Proc. 2nd ACM Conf. Online Social Netw.*, 2014, pp. 215–226.

[39] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, “Quantifying location privacy,” in *Proc. IEEE Symp. Security Privacy*, 2011, pp. 247–262.

[40] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, “User identity linkage across online social networks: A review,” *SIGKDD Explorations Newslett.*, vol. 18, no. 2, pp. 5–17, 2017.

[41] R. Singh, G. Theodorakopoulos, M. K. Marina, and M. Arapinis, “On choosing between privacy preservation mechanisms for mobile trajectory data sharing,” in *Proc. IEEE Conf. Commun. Netw. Security*, 2018, pp. 1–9.

[42] A. Solanas and A. Martínez-Ballesté, “Privacy protection in location-based services through a public-key privacy homomorphism,” in *Proc. 4th Eur. Conf. Public Key Infrastructure: Theory Practice*, 2007, pp. 362–368.

[43] K. Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *J. Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[44] M. Srivatsa and M. Hicks, “Deanonymizing mobility traces: Using social network as a side-channel,” in *Proc. ACM Conf. Comput. Commun. Security*, 2012, pp. 628–637.

[45] L. Sweeney, “k-anonymity: A model for protecting privacy,” *Int. J. Uncertainty Fuzziness Knowl.-Based Syst.*, vol. 10, no. 05, pp. 557–570, 2002.

[46] J. A. Thomas and T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.

[47] G. Wang, S. Y. Schoenebeck, H. Zheng, and B. Y. Zhao, “”Will check-in for badges”: Understanding bias and misbehavior on location-based social networks,” in *Proc. Int. Conf. Weblogs Social Media*, 2016.

[48] H. Wang, C. Gao, Y. Li, G. Wang, D. Jin, and J. Sun, “De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice,” in *Proc. Netw. Distrib. Syst. Security*, 2018.

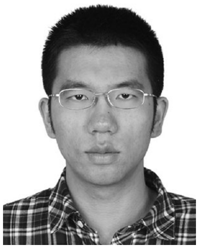
[49] H. Zang and J. Bolot, “Anonymization of location data does not work: A large-scale measurement study,” in *Proc. 17th Annu. Int. Conf. Mobile Comput. Netw.*, 2011, pp. 145–156.

[50] J. Zhang and S. Y. Philip, “Multiple anonymized social networks alignment,” in *Proc. IEEE Int. Conf. Data Mining*, 2015, pp. 599–608.

[51] Z. Zhang *et al.*, “On the validity of geosocial mobility traces,” in *Proc. 12th ACM Workshop Hot Topics Netw.*, 2013, Art. no. 11.

[52] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, “Big data-driven optimization for mobile networks toward 5G,” *IEEE Netw.*, vol. 30, no. 1, pp. 44–51, Jan./Feb. 2016.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.



Huandong Wang (Student Member, IEEE) received the BS degrees in electronic engineering and mathematical sciences from Tsinghua University, Beijing, China, in 2014 and 2015, respectively. He is currently working toward the PhD degree in the Department of Electronic Engineering, Tsinghua University. His research interests include software-defined networks, wireless ad hoc network, and mobile big data.

Yong Li (Senior Member, IEEE) received the BS degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2007 and the PhD degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. He is currently a faculty member of the Department of Electronic Engineering, Tsinghua University. He has served as General Chair, TPC Chair, SPC/TPC Member for several international workshops and conferences, and he is on the editorial board of two IEEE journals. His papers have total citations more than 8300. Among them, 10 are ESI Highly Cited Papers in Computer Science, and four receive conference best paper (run-up) awards. He received IEEE 2016 ComSoc Asia-Pacific Outstanding Young Researchers, Young Talent Program of China Association for Science and Technology, and the National Youth Talent Support Program.

Chen Gao received the BS degree in electronic engineering from Tsinghua University, Beijing, China, in 2016. He is working toward the PhD degree with the Department of Electronic Engineering, Tsinghua University, Beijing, China. His research interests include the theory and applications of mobile big data and data mining.

Wang Gang received the BE degree in electrical engineering from Tsinghua University, Beijing, China, in 2010, and the PhD degree in computer science from UC Santa Barbara, CA, in 2016. He is an assistant professor with the Department of Computer Science, University of Illinois at Urbana-Champaign (UIUC). He was the recipient of the Google Faculty Research Award (2018), Best Practical Paper Award from ACM SIGMETRICS (2013), Outstanding Dissertation Award (2016), and PhD dissertation fellowship (2015) from UC Santa Barbara. His research interests include security and privacy, online social networks, mobile networks, and crowdsourcing.

Xiaoming Tao (Member, IEEE) received the BE degree from the School of Telecommunications Engineering, Xidian University, in 2003, and the PhD degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2008. She is currently a professor with the Department of Electronic Engineering, Tsinghua University. Her research interests include wireless communications and networking, and multimedia signal processing.

Depeng Jin (Member, IEEE) received the BS and PhD degrees from Tsinghua University, Beijing, China, in 1995 and 1999, respectively, both in electronics engineering. Currently he is an associate professor with Tsinghua University and vice chair of the Department of Electronic Engineering. He was awarded the National Scientific and Technological Innovation Prize (Second Class), in 2002. His research interests include telecommunications, high-speed networks, ASIC design, and future internet architecture.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: Tsinghua University. Downloaded on July 31,2023 at 13:14:58 UTC from IEEE Xplore. Restrictions apply.