De-anonymization of Mobility Trajectories: Dissecting the Gaps between Theory and Practice

Huandong Wang*, Chen Gao*, Yong Li*, Gang Wang†, Depeng Jin*, and Jingbo Sun‡
*Department of Electronic Engineering, Tsinghua University
†Department of Computer Science, Virginia Tech
‡China Telecom Beijing Research Institute
{whd14,gc16}@mails.tsinghua.edu.cn, {liyong07,jindp}@tsinghua.edu.cn,
gangwang@vt.edu, sunjb@ctbri.com.cn

***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 (small) datasets. However, it is not clear how well the theoretical estimations are preserved in practice.**

In this paper, 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 (7 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 underestimated the impact of spatio-temporal mismatches between the data from different sources, and the high sparsity of user generated data also contributes to the underperformance. Based on these insights, we propose 4 new algorithms that are specially designed to tolerate spatial or temporal mismatches (or both) and model user behavior. Extensive evaluations show that our algorithms achieve more than 17% performance gain over the best existing algorithms, confirming our insights.

I. Introduction

Anonymized user mobility traces are increasingly collected by Internet Service Providers (ISP) to assist various applications, ranging from network optimization [42] to user population estimation and urban planning [11]. 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 [40]. Recently, the US congress has moved towards repealing the Internet Privacy Rules and legalizing ISPs to share (or monetize on) user data [14]. 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 [9], [40], which shows that trajectory traces are surprisingly easy to de-anonymize. With 4 spatio-temporal points or top 3 most visited locations, results in [9], [40] show that 80%–95% of the user scan 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) [8], [10], [15], [16], [23], [27]–[29], [31]–[33], [35]. However, due to the lack of large empirical *ground-truth* datasets, researchers have to settle on small datasets (*e.g.*, 125 users in [35], 1717 users in [31]) or simulating attacks on synthetically generated data (*e.g.*, using parts of the same dataset as the victim dataset and the external information source) [23], [32], [33]. 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 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 7 major de-anonymization algorithms against our dataset, we find the existing algorithms largely fail the de-anonymization task using practical data. Their performance is far from the privacy bound [9], [40], and massive errors occur, $i.e.$, the hit-precision is less than 20%. 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

Network and Distributed Systems Security (NDSS) Symposium 2018
18-21 February 2018, San Diego, CA, USA
ISBN 1-891562-49-5
http://dx.doi.org/10.14722/ndss.2018.23211
www.ndss-symposium.org

[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.

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 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.

Overall, our work makes four key contributions:

- First, we collect the first large-scale trajectory dataset (with ground-truth) to evaluate de-anonymization attacks. The dataset contains 2,161,500 ISP trajectories and 56,683 external trajectories, which helps to overcome the limitations of theoretical analysis and small-scale validations.
- Second, we build an empirical evaluation framework by categorizing and implementing existing de-anonymization algorithms (7 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 user-level errors. Optional components such as user historical trajectories can also be added to our framework to improve the performance.
- Finally, extensive performance evaluation shows that our algorithms achieve over 17% performance gain in terms of the hit-precision. In addition, our algorithms are robust against parameter settings, *i.e.*, even without ground-truth data, by using the empirical parameters, our proposed algorithms still outperform existing ones. This results confirm the usefulness of our insights.

Our work is a first attempt to bridge the gaps between the theory bound and the practice attacks for the location trajectory de-anonymization problem. We show that failing to consider the practical factors undercuts the performance of the de-anonymization algorithms. Future work will consider building more accurate privacy metrics to quantify privacy loss given imperfect data, and develop privacy protection techniques on top of anonymized trajectory datasets.

In the following, we first categorize existing approaches to evaluating the privacy leakage in anonymized mobility datasets (§II), followed by our de-anonymization framework (§III). In §IV, we describe the large ground-truth dataset, using which we analyze the theoretical privacy bound and the performance of existing algorithms (§V). After analyzing the main reasons of the under-performance of existing approaches (§VI), we build and evaluate our own algorithms to validate our insights (§VII–VIII).

II. Related Work

De-anonymization Methods: Overview. In Table I, we summarize the key 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) [34]. 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 [8]–[10], [23], [27], [28], [31]–[33], [40]. 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 ISP dataset and external traces), existing works either focus on theoretical privacy bound [9], [40] or simulating de-anonymization attacks on a small dataset [9], [23], [32], [33], [40]. 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.

In Table I, we further categorize these algorithms based on their design principles. For example, some algorithms are designed to tolerate mistakes in the adversary's knowledge such as temporal mismatching [28] and spatial mismatching [23]. Other algorithms [27], [32], [33] implement de-anonymization attacks based on *individual user's* mobility patterns [27], [33]. Finally, researchers also develop de-anonymization algorithms based on "encountering" events [8], [31]. By considering the location context (*e.g.*, user population density), it achieves a better performance [31]. As shown in Table I, none of these algorithms checks all boxes. In particular, no algorithm simultaneously tolerates 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 [19], [20], [29], [35] and *profile* datasets [15], [16], [26] for completeness. Mudhakar *et al.* [35] and Ji *et al.* [19], [20] focused on de-anonymization based on users' graph/network structures. These algorithms can be adapted to deanonymizing location trajectories by constructing a "contact graph" to model users encountering with each other. However, these algorithms require using social network graphs as the

TABLE I. COMPARISON OF DE-ANONYMIZATION ALGORITHMS, √=TRUE, ×=FALSE, −=N/A.

	Information Used	Tolerate Spatial Mismatching	Tolerate Temporal Mismatching	Per-user Mobility Model	Considering Location Context
POIS [31]	Content	×	×	×	√
WYCI [32]	Content	×	√	√	×
HMM [33]	Content	√	×	√	×
HIST [27]	Content	×	√	√	×
ME [8]	Content	×	×	×	×
MSQ [23]	Content	√	×	×	×
NFLX [28]	Content	×	√	×	×
CG [35]	Content/Network	−	−	−	−
ODA [20]	Content/Network	−	−	−	−
SG [29]	Network	−	−	−	−
PM [16]	Profile	−	−	−	−
ULink [26]	Profile	−	−	−	−
LRCF [15]	Profile/Content	−	−	−	−

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 [15], [16], [26] (*e.g.*, age, gender, language) are not applicable to location trajectories, and thus omitted for brevity.

Privacy Protection Mechanisms. Researchers have investigated different ways to anonymize user data to preserve privacy. The most common privacy models are k-anonymity [36], l-diversity [24] and t-closeness [21]. 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.* [18] 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 $et\ al.$ [30] provided two metrics, conditional entropy and worst-case quality loss, to evaluate the privacy protection mechanisms.

Recently, researchers also explore to apply differential privacy to location trajectory datasets [3], [5], [12]. 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.

III. 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 II.

A. Location Data Publishing by ISP

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

TABLE II. 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.
$L_v(t)$	Location in the ISP trajectory of user with pseudonym v at time slot t.
$S_u(t)$	Location in the external trajectory of user u at time slot t.
σ	Anonymization function mapping $\boldsymbol{U}$ to $\boldsymbol{V}$.
D	Similarity score function between trajectories.
$R(u, D)$	The rank of the true matched trajectory of u based on similarity function D.
$T^v_{i,j}$	Transition matrix of user u.
$\Phi(\mathcal{S}, D)$	Performance metric of de-anonymization attack.
$I(\cdot)$	Indicator function of logical expressions with $I(\textit{true}) = 1$ and $I(\textit{false}) = 0$.

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.

We define the mobility trace of the user with pseudonym $v \in \boldsymbol{V}$ published by ISP as a T-size vector $\boldsymbol{L}_v = (L_v(1), L_v(2), ..., L_v(T))$ where $L_v(t)$ represents the location observed at time slot t, and T is the total number of time slots. For time slots with a location record, $L_v(t)$ is the corresponding geographic coordinate. 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 $\mathcal{L} = \{\boldsymbol{L}_v | v \in \boldsymbol{V}\}$. In this work, we mainly focus on the effectiveness of the de-anonymization attacks. We assume the ISP does not apply additional obfuscations to the data other than the common steps such as reducing the spatio-temporal resolution of the records [33]. This benefits assessing the upper-bound performance of the existing attacking methods against real-world datasets.

B. 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. Adversary 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 uniform T-size vector $\boldsymbol{S}_u = (S_u(1), S_u(2), ..., S_u(T))$ to represent its external information, with $S_u(t)$ representing the location (geographic coordinate) observed at time slot t for user $u \in \boldsymbol{U}$, In addition, we set $S(t) = \emptyset$ in time slot t without 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 $\boldsymbol{S}_u$, adversary rank of 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 ranking $R(u, D)$. For higher $R(u, D)$, $h(R(u, D))$ is larger. Then, the performance of the attack method can be expressed as follows,

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

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

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

In terms of the ranking, a well-established and widely-used evaluation metric is the hit-precision of top-k candidates, which is defined as follows,

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

For example, if the true matched 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 $h(R(u, D)) = \frac{k-2}{k}$.

IV. 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 data 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 III. Below, we describe the datasets in detail and perform a preliminary analysis.

A. 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.

B. 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 IV-E).

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 attacks. 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 collected 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 attacks. Similarly, we only access the anonymized ID, instead of the actual Weibo ID.

C. 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,

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

TABLE III. STATISTICS OF COLLECTED DATASETS.

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

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.

D. 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.

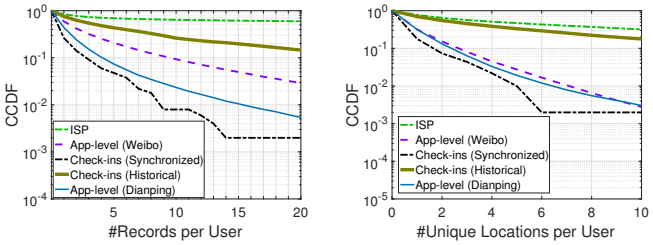
Converting Basestation ID to GPS. To construct user mobility traces from the ISP data, we first convert the ID of base stations to their geographical coordinates (longitudes and latitudes) based on the ISP offered database, and use it to represent the user location.

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 $\boldsymbol{S}_u$ and $\boldsymbol{L}_v$. 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 [32], [33]. 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.

E. 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.

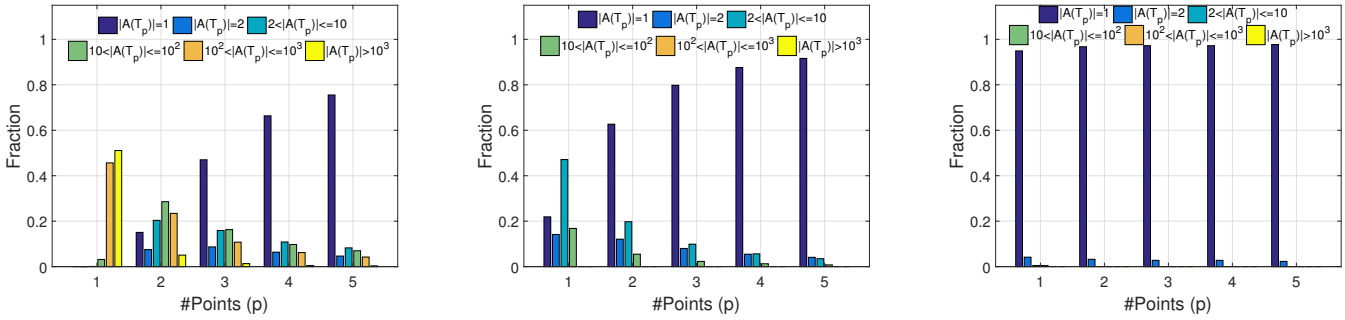


(a) Records per User (b) Distinct Locations per User

Fig. 1. Complementary cumulative distribution function (CCDF) of the number of records and number of distinct locations per user.

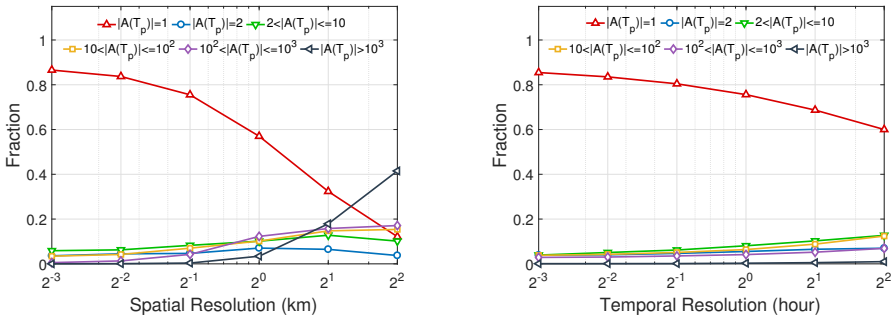
F. Preliminary Data Analysis

Fig. 1 and Table III 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% of the IPS 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 (62.1 and 9.19). The Weibo and Dianping datasets (app level) are sparse with 4.22 and 2.34 records 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). Not too surprisingly, the check-in dataset is even sparser. 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. This helps to solve the critical problem of lacking ground truth data in the existing works [9], [32].



(a) Uniqueness of ISP trajectories (b) Uniqueness of Weibo app-level trajectories (c) Uniqueness of check-in trajectories (sync.)

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.



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

Fig. 3. The impact of temporal and spatial resolutions on the privacy bound analysis of the IPS dataset. p is the number of randomly selected data points from the trajectories as the external observations.

V. 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 V-A. Then, comparing with the theoretical bound, we implement 7 existing de-anonymization algorithms in practice, and show their performance in Section V-B.

A. 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 [9], [18], [40]. In 1930, Edmond Locard showed that 12 points are sufficient to uniquely identify a fingerprint [9], [13]. 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.

The uniqueness metric 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 user's *anonymity set* 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.

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 §IV-D). For example, if two trajectories show users visiting the same locations in the same order but at different times, 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) [9].

We evaluate the uniqueness of trajectories in different datasets as the function of p. The results are shown in Fig. 2. As we can observe from Fig. 2(b) and (c), uniqueness of trajectories in Weibo app-level and check-in dataset are both very high, $e.g.$, 5 points can uniquely identify over 90% users. Results are similar for the Dianping dataset (the figure is omitted for brevity). The high uniqueness of these two types of external information guarantees their high ability to de-anonymize the ISP trajectory dataset. On the other hand, from Fig. 2(a), we can observe that the uniqueness of ISP trajectories is a bit lower. The main reason is that the number of ISP trajectories is significantly larger, $e.g.$, 38 times larger than the number of Weibo app-level trajectories. Such large quantity of the data makes individuals better hidden in the crowd. Nevertheless, the uniqueness of ISP trajectories is also high, $i.e.$, 5 points can uniquely identify over 75% users, indicating their potential high risk to be de-anonymized.

In addition, we analyze the influence of the spatio-temporal resolutions on the uniqueness. We fix the number of spatio-temporal points as 5, and the obtained results for the ISP

dataset. As shown in Fig. 3, 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%, while reducing the spatial resolution from 250m to 1km only leads to the decreasing of uniqueness by 26%. 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% users, indicating the high potential risk of individual trajectories to be de-anonymized, which exposes a big threat to users' privacy.

B. Actual Performance of Attack Methods

To examine the effectiveness of de-anonymization attacks, we implement 7 major attacking algorithms discussed in the Section II. We focus on algorithms that are designed (or can be adopted) to work on trajectory datasets.

HMM: Shokri *et al.* [33] 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 $f : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ 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:

$$D_{\text{HMM}} = P(\boldsymbol{S}_u|T^v) = \sum_{\boldsymbol{Z}} \prod_{t\in\mathcal{T}} f(Z(t), S(t)) T^v_{Z(t-1),Z(t)},$$

where $\boldsymbol{Z}$ is the hidden variable representing users' true locations.

HIST: Naini *et al.* [27] 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:

$$D_{\text{HIST}} = -D_{\text{KL}}(\Gamma_u|\frac{\Gamma_u+\Gamma_v}{2}) - D_{\text{KL}}(\Gamma_v|\frac{\Gamma_u+\Gamma_v}{2}),$$

where D_{KL} the Kullback-Leibler divergence function [37].

WYCI: Rossi *et al.* [32] 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 $P(r|\boldsymbol{L}_v) = \frac{n^v_r+\alpha}{\sum_{r\in\mathcal{R}} n^v_r+\alpha|\mathcal{R}|}$, where n^v_r is the times of visit of user v to 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 [38], we set $\alpha = 0.1$. Then, their similarity score is defined as follow:

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

ME: Cecaj *et al.* [8] 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:

$$D_{\text{ME}} = \sum_{t\in\mathcal{T}} I(S(t) = L(t)).$$

POIS: Riederer *et al.* [31] 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,

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

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

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

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

NFLX: Narayanan *et al.* [28] 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 [31], which is defined as follows:

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

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

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

In addition, n^v_r is the times of visit of user v to location r. Temporal mismatches are considered in this algorithm. However, it cannot tolerate spatial mismatches.

MSQ: Ma *et al.* [23] find the matched traces by minimizing the expected square between them. That is, their similarity score can be expressed as follows:

$$D_{\text{MSQ}} = -\sum_{t\in\mathcal{T}} |L(t) - S(t)|^2.$$

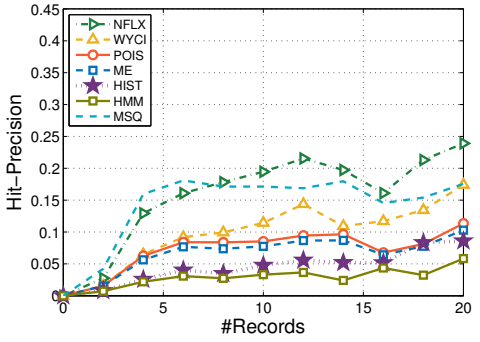
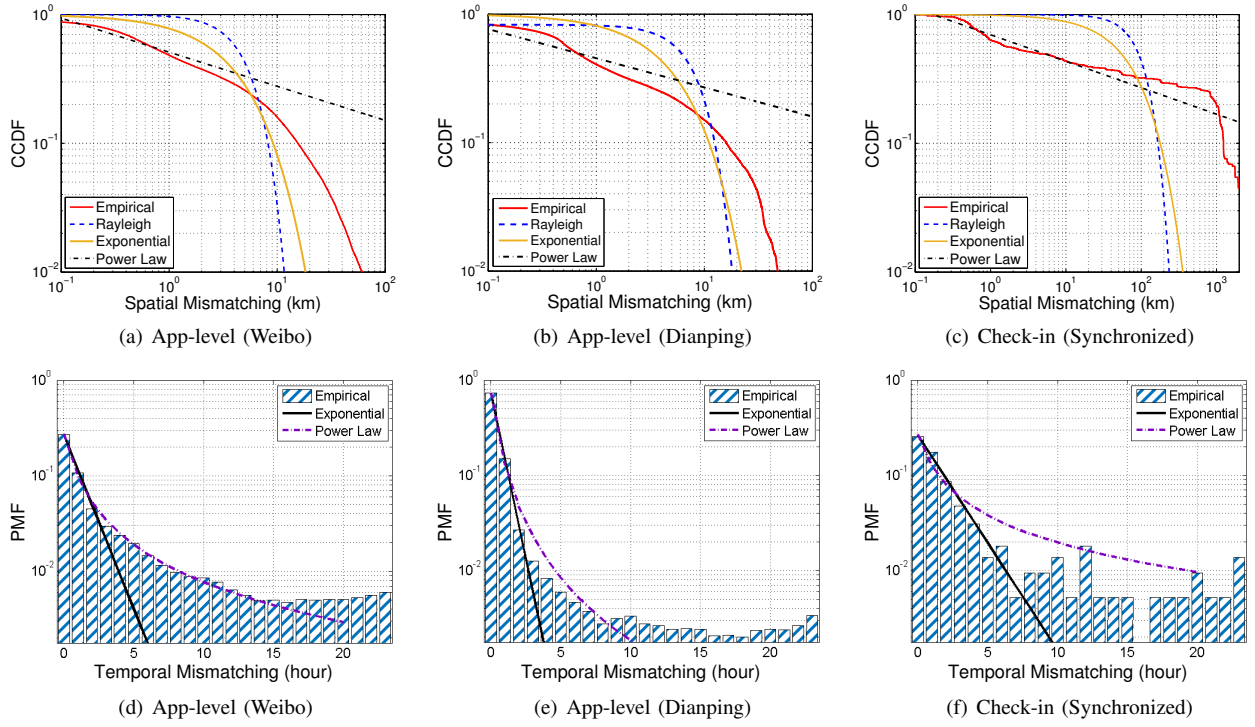


Fig. 4. Performance of different algorithms as a function of the number of records in Weibo's app-level trajectories.



(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. 5. Complementary cumulative distribution functions (CCDF) and probability mass function (PMF) of the spatial and temporal mismatching (with the ISP traces). The empirical distribution is compared with the fitting results of Rayleigh, exponential, power-law distributions.

Spatial mismatches are considered in this algorithm. However, it cannot tolerate temporal mismatches.

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 and NFLX 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 1km×1km 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 VIII.

Fig. 4 shows the performance of all 7 algorithms for using Weibo's app-level trajectories to de-anonymize the ISP trajectories. The hit-precision is plotted as the function of the number of records in app-level trajectories. As shown in Fig. 4, de-anonymization algorithms based on users' mobility patterns (*e.g.*, HIST and HMM) have the worst performance with a maximum hit-precision about 8%. On the other hand, algorithms based on meeting events including ME and POIS have better performance, with a maximum hit-precision about 11%. Algorithms such as NFLX and MSQ achieve a better performance. Even so, their maximum hit-precision is only about 20%, which is far from the privacy bound obtained in Section V-A, *i.e.*, 5 points can identify over 75% users.

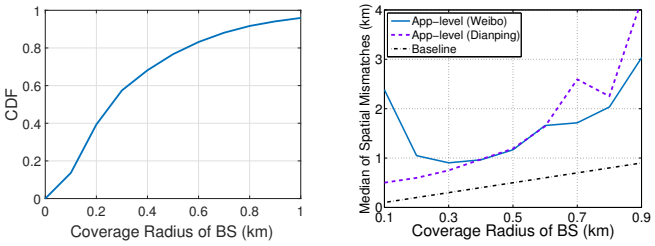
Note that in our experiment, datasets are already "matched" — the user population of the external dataset is already a subset of users in the target ISP dataset. This means for each trajectory in the external datasets, we know that there must be a trajectory in the ISP dataset. In practice, the attack 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.

VI. Reasons behind Underperformance

A. Spatio-Temporal Mismatch

We start by investigating the potential spatio-temporal mismatches between trajectories in different datasets. Fig. 5 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. Fig. 5(a), (b) and (c) show the complementary cumulative distribution functions (CCDF) of spatial mismatches of different datasets. We observe that the spatial mismatches are prevalent. More than 37% of the records in the app-level trajectory data of Weibo have spatial mismatches over 2km. It is similar in the other application, Dianping, of which the spatial mismatch of over



(a) Base Station Coverage (b) Spatial Mismatches

Fig. 6. The coverage radius of base stations, and its relationship with the spatial mismatches.

31% of the records are larger than 2km. 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 10km. After 10km, 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 [31], [32].

Fig. 5(d), (e) and (f) show the probability mass function (PMF) of temporal mismatches. The temporal mismatches are also very prevalent. Only 30% 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 collocation of users [8], [31]. 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. 5(b) and (c) 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 did 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 had 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).

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 [22]. 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.

Fourth, *deployment of base stations*: the base stations (BS) are placed unevenly in the city. In the ISP trajectory dataset, we use the connected BS to estimate the user's location, which may caused the spatial mismatches, especially in areas where the base stations are sparse. To investigate this intuition, we plot Fig. 6. We consider Weibo's and Dianping's app-level trajectory data for Fig. 6(b), and use $y = x$ as a baseline. A larger radius indicates a *sparser* placement of base stations. Not too surprisingly, a larger coverage radius (sparser BS placement) leads to bigger spatial mismatches. In addition, spatial mismatches (y axis) are significantly larger than the coverage radius (x axis), indicating that the BS placement is not the only reason for spatial mismatches.

Finally, *user behavior*: for the check-in dataset, mismatches may also come from special user behavior. According to recent measurement studies [39], [41], 39.9% 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 [39]. These factors can lead to major mismatches between the check-ins and the ISP 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 tolerant these spatio-temporal mismatches.

B. Data Sparsity

Another possible reason 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% users have less than 1 records and 35.5% 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 overlooked when testing a de-anonymized algorithm on a synthetically generated dataset or a small dataset contributed by several hundreds of volunteers.

VII. 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' missing locations in the "empty" time slots of the trajectory; b) We leverage the whole dataset to aggregate

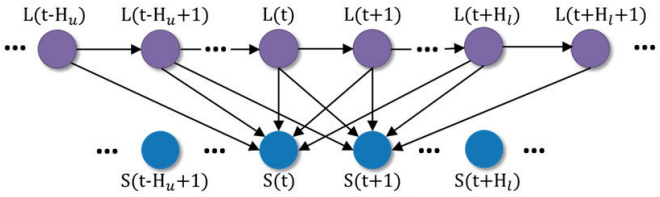


Fig. 7. Graph model for $\boldsymbol{L}$ (ISP trajectory) and $\boldsymbol{S}$ (external trajectory)

global location contexts and user behavior features to further infer the missing location records.

Our proposed algorithms combine Gaussian mixture model and Markov model. We refer the algorithm as *GM*. Fig. 7 shows the relationship of random variables in our model. Based on this probabilistic model, we define the similarity score function as follows,

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

In this section, we will introduce how to compute this probability-based similarity score to de-anonymize location trajectories.

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

In order to model the strong mismatches in the adversary's knowledge in terms of both spatial dimension and temporal dimension, we adopt the Gaussian mixture model (GMM). By definition, GMM is a linear superposition of finite Gaussian densities, which can be expressed as:

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

where 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 shown in Fig. 7, we use component $\mathcal{N}(x|u_p, \Sigma_p)$ to represent the probability density of external records with temporal mismatching of p time units. Then, let 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 $S(t)$ belonging to the same user can be calculated as,

$$p(S(t)|\boldsymbol{L}) = \sum_{p=-H_l}^{H_u} \pi(p) \cdot \mathcal{N}(S(t)|L(t-p), \sigma^2(p)I_2), \quad (1)$$

where $\pi(p)$ is the probability of the temporal mismatch to be p time units, and $\sigma(p)$ is the mean square root of the spatial mismatch conditioned on the temporal mismatch of p time units. In addition, 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.

Parameters $\pi(p)$ and $\sigma(p)$ can be chosen by the empirical values shown in Fig. 5. On the other hand, they can also be estimated by EM algorithm [7]. Specifically, given M external records $\{S_1, ..., S_M\}$ with their corresponding $|H_u + H_l|$ ISP records in neighboring time slots, *e.g.*, for S_n, its neighboring ISP records are $(L_{n,-H_l}, ..., L_{n,H_u})$. In addition, we define z_{nk} as the latent variable to indicate whether S_n are generated by L_{nk} (corresponding 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 π and σ, which can be expressed as follows,

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

In the M step, we re-estimate the parameters π and σ using the distribution of z_{nk}, which can be expressed as follows,

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

Then, by a finite number of repeating E and M step, we obtain the value of π and σ. Specifically in our problem, we only consider time delay in adversary's knowledge. Thus, we set H_l to be zero. By defining $G_{\pi,\sigma}(p, r_1, r_2) = \pi(p) \cdot \mathcal{N}(r_1|r_2, \sigma^2(p)I_2)$, (1) can be simplified as:

$$p(S(t)|\boldsymbol{L}) = \sum_{p=0}^{H} G_{\pi,\sigma}(p, S(t), L(t-p)), \quad (2)$$

where u of H_u is ignored for simplicity.

B. Modelling User Mobility: Markov Model

Based on the graph model shown in Fig. 7, 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 (PDF) of a full trajectory in external dataset can be calculated as follows,

$$p(\boldsymbol{S}|\boldsymbol{L}) = \prod_{S(t)\neq\emptyset} p(S(t)|\boldsymbol{L}). \quad (3)$$

However, from the analysis in Section IV-F, we can observe that users' locations in many time slots are missing, *i.e.*, $\exists t \in \mathcal{T}$ such that $L(t) = \emptyset$. In the case, (2) cannot be applied directly. In addition, $S(t)$ for different t also becomes dependent with each other. Thus, (3) 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 L. Then the PDF of $S(t)$ conditioned on L can be calculated as follow:

$$p(\boldsymbol{S}|\boldsymbol{L}) = \sum_{\boldsymbol{L_C}\in\mathcal{C}(\boldsymbol{L})} p(\boldsymbol{L_C}|\boldsymbol{L}) \prod_{S(t)\neq\emptyset} p(S(t)|\boldsymbol{L_C}). \quad (4)$$

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. Let $E(r)$ to be the margin distribution of the user, 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,

$\alpha(r)$ is the parameter to eliminate zero probabilities. For example, in Laplace smoothing [25], $\alpha(r)$ is set to be the same value for different r. In our work, we use the location context to implement the smoothing as follow,

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

where $\alpha(r)$ 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 $\boldsymbol{L}_C \in \mathcal{C}(\boldsymbol{L})$ conditioned on $\boldsymbol{L}$ can be calculated as follows,

$$p(\boldsymbol{L_C}|\boldsymbol{L}) = \prod_{t\in\mathcal{T}, L(t)=\emptyset} E(L_C(t)). \tag{5}$$

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 $T_{r_1r_2}$ as the transition probability matrix of the user, which can be calculated as follows,

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

Similarly, $\beta(r_1, r_2)$ 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,

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

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 $P(\boldsymbol{L})$ can be calculated by using n-order transition matrix.

On the other hand, as we can observe from Section IV, the trajectories in external information are obviously sparser than those in the anonymized dataset. It indicates that in real external trajectory, for each pair of adjacent non-empty $S(t_1)$ and $S(t_2)$, we generally 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. In this way, the computational complexity can be significantly reduced. Taking 0-order Markov model for example, we have:

$$p(S(t)|\boldsymbol{L}) = \sum_{p=0}^{H} \sum_{r\in\mathcal{R}} G(p, S(t), r)p(L_C(t-p) = r|\boldsymbol{L}),$$

where π and σ in $G_{\pi,\sigma}$ are omitted for simplicity. In addition, $p(L_C(t-p) = r|\boldsymbol{L})$ is the probability of a record at location r in time slot $t-p$, which can be represented as follows,

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

By this way, the complexity can be reduced from $O(T \cdot R^H)$ to $O(T \cdot R \cdot H)$, which is also similar for 1-order Markov model. The influence of ignoring dependency of external records will also be analyzed in Section VIII.

C. Modelling User Behavior

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 (3), and ignoring the unobserved time slots t with $S(t) = \emptyset$. However, (3) 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 are 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:

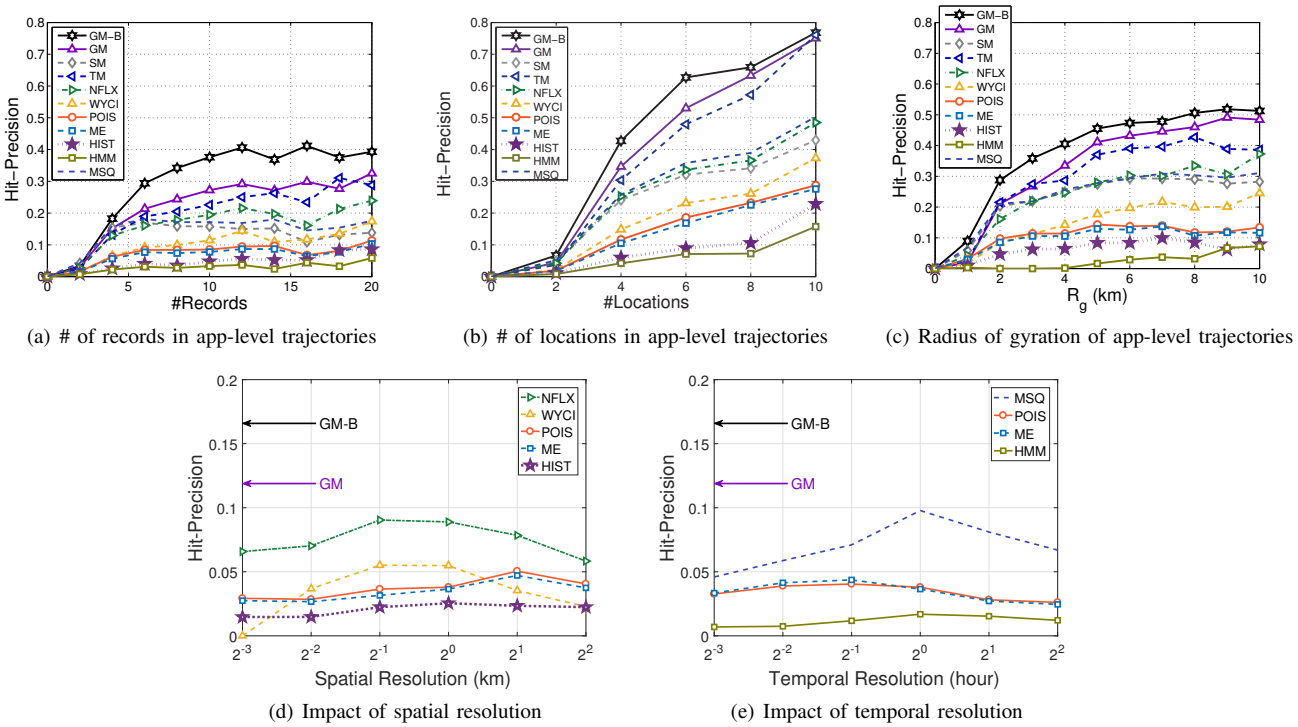
$$D_{\mathrm{GM-B}} = D_{\mathrm{GM}} + D_{\mathrm{B}}.$$

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 in practice, while GM algorithm shows the performance when we do not have sufficient ground truth data.

D. Baseline Algorithm

For baseline comparisons, we also propose two simplified versions which only consider spatial mismatches and temporal mismatches, respectively. We refer to them as spatial matching (SM) algorithm and temporal matching (TM) algorithm.

Spatial Matching Algorithm (SM). The SM algorithm ignores the mismatch in temporal dimension, and only matches



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

(d) Impact of spatial resolution (e) Impact of temporal resolution

Fig. 8. Performance of different de-anonymization algorithms using Weibo's app-level trajectories as the external information.

records at the same time slot with Gaussian distribution. Then, its similarity score can be defined as:

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

Similarly with GM algorithm, when $L(t)$ is $\emptyset$, the margin 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,

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

Specifically, we use empirical temporal mismatch distribution shown in Fig. 5 as $\pi(t)$.

VIII. Performance Evaluation

Now, we systematically evaluate the performance of our algorithms and compare them with existing and baseline methods. In the following, we apply our algorithms on different trajectory datasets to perform de-anonymization. In addition, we vary key parameters and experiment settings to examine the robustness of the proposed algorithms.

A. De-anonymization Attack

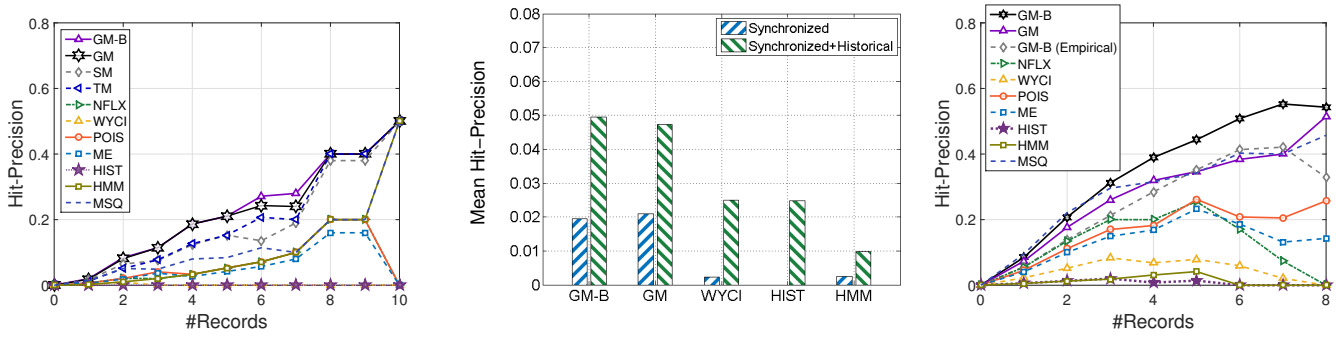
De-anonymization using Weibo's App-level Trajectories. As a primary experiment, we evaluate the performance of different algorithms by using Weibo's 56,683 app-level trajectories as the external information to de-anonymize the ISP dataset. In Fig. 8, the hit-precision is calculated as functions of different metrics of external trajectories, including number of records, number of distinct locations, and the radius of gyration [17] of the external trajectories.

Fig. 8(a) 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 better performance 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% higher compared with existing algorithms. Finally, by comprehensively modelling users' behavior, GM-B algorithm achieves another significant performance gain (7% 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% for external trajectories with more than 10 records, improving over 72% compared with the existing algorithms.

We notice that after the number of records get higher than 10, the performance gain stalls. In Fig. 8(b), 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%.

Radius of gyration reflects the range of a user' activity area. It is defined as the mean square root of the distance



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

Fig. 9. Performance of different de-anonymization algorithms using Dianping and Weibo Check-in trajectories as the external information.

of each points in the trajectory to its center of mass [17]. It can be calculated by $r_g = \sqrt{\Sigma_{t\in\mathcal{T},S(t)\neq\emptyset}(S(t)-S_{cm})^2/n}$, where n is the number of non-empty elements in $\boldsymbol{S}$, *i.e.*, $n = \sum_{t\in\mathcal{T}} I(S(t)\neq\emptyset)$. In addition, $S_{cm} = \sum_{t\in\mathcal{T},S(t)\neq\emptyset} S(t)/n$ is the center of mass of the trajectory. As we can observe, the best hit-precision in terms of radius of gyration only achieve 52%. Compared with Fig. 8(b), the result indicates that the number of distinct locations is a more dominating factor in the de-anonymization attack.

As mentioned in Section V-B, POIS, WYCI, HIST, ME and NFLX 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 performance under different spatial resolutions (temporal resolution is set to the default value 1 hour). For comparison purposes, we also mark the performance of GM and GM-B in the figures (using default 1 hour and 1km). As shown in Fig. 8(d), most algorithms, *i.e.*, NFLX, WYCI and HIST, achieve their best performance under our default spatial resolution of 1km, while POIS and ME algorithms achieve their best performance under the spatial resolution of 2km. 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% relative to the best performance of existing algorithms respectively.

Similarly, POIS, HMM, ME and MSQ are based on "concurrent" events, making them potentially sensitive to temporal resolutions. Fig. 8(e) shows their performance of under different temporal resolution (spatial resolution is set to default 1km). The result shows that HMM and MSQ algorithms achieve their best performance under our default temporal resolution of 1 hour, while POIS and ME achieve their best performance under the temporal resolution of 30min. Our proposed algorithms still outperform existing algorithm, *e.g.*, performance gap of GM and GM-B algorithms are 21.6% and 69.9% relative to the best existing algorithm respectively.

Validation using Weibo Check-in Trajectories. To validate our observations, we further evaluate the performance of our algorithms using Weibo check-ins trajectories as external information. We firstly focus on the 503 check-in trajectories that have at least 1 records at the same time-window with ISP dataset. The hit-precision is shown as the function of the number of records of check-in trajectories in Fig. 9(a). As we can observe, more records in check-in trajectories help to improve the de-anonymization accuracy. In addition, our propose GM and GM-B algorithm outperform other algorithms. The largest performance gap between our proposed algorithms and existing algorithm achieves about 20% when there are 8 records in the check-in trajectories.

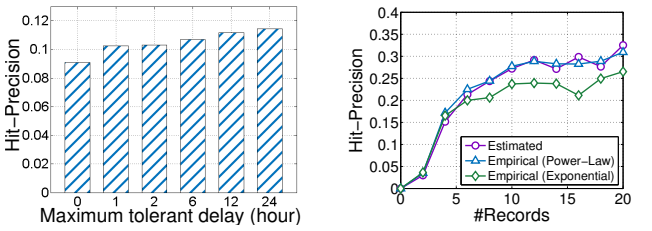
Fig. 9(b) shows the mean hit-precision of de-anonymization based on synchronized and historical Weibo check-ins. The mean hit-precision is very low because the synchronized check-ins are extremely sparse. For example, as shown in Fig. 1, over 80% users have less than 2 records. The historical check-ins 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, HMM 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 performance of all the algorithms. Intuitively, historical check-ins can greatly mitigate the sparsity issues 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. 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 directly apply the parameters estimated from the Weibo dataset to the Dianping experiment (empirical GM-B). As shown in Fig. 9(c), the empirical 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.

B. Parameter Evaluation

Finally, we examine how selected parameters in our algorithm influence the attack results.

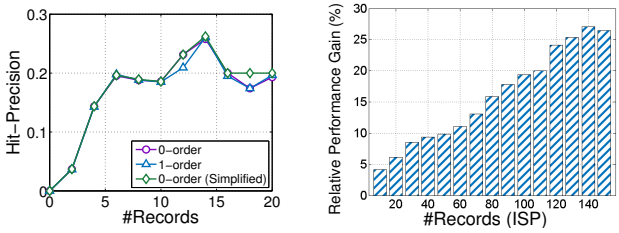
Impact of the Parameters in GMM. Fig. 10 shows the sensitivity of GMM's performance against different parameter settings. Fig. 10(a) shows the average hit-precision of the GM



(a) Different maximum tolerant delay H_u

(b) Different temporal mismatching distribution π

Fig. 10. Performance vs. different parameters in GMM.



(a) Different Order

(b) Location Context

Fig. 11. Performance vs. different components in the Markov model.

algorithm with different maximum tolerate delay H_u. We use Weibo's app-level trajectories with 2+ distinct locations. As we can observe, the hit-precision is improved slowly with a very slow rate with H_u. For a better performance, H_u should be set to a large value. However, as mentioned in Section VII-B, the computational complexity of GM algorithm increases with H_u. Thus, we should compromise between accuracy and computational complexity in real de-anonymization attack.

Next, we examine the impact of parameter $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ in the GMM. For an adversary without a detailed ground-truth dataset, these parameters cannot be estimated by the EM algorithm. To this end, instead of using parameters produced by the EM algorithm, we apply different parameters from the empirical distribution fitting: $\sigma(p)$ is set to be 0.5km for all p, and $\pi(p)$ is set to be the power-law or exponential distribution. Then, we compare their performance in Fig. 10(b).

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 performance 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.

Impact of the Parameters in Markov Mobility Model. The key parameter of the Markov mobility model is the component. Below, we evaluate the impact of the order of Markov and location context.

In Fig. 11(a), we show the impact of using 0-order Markov or 1-order Markov, as well as ignoring the dependency between external records. Specifically, we use 0-order (simplified) to represent the GM algorithm with 0-order Markov mobility model ignoring dependency between external records. In addition, maximum tolerate delay H_u is set to be 1 hours, and $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ use the value estimated by EM algorithm. As shown in Fig. 11(a), 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 performance. In addition, Fig. 11(b) shows the relative performance gain for GM algorithm with location context compared with it without location context. As we can observe, by utilizing the location context, over 25% relative performance gain is achieved, demonstrating its effectivity.

Experiment Limitation. As mentioned in Section V-B, for each trajectory in the external datasets, there must exist a matched trajectory in the ISP dataset. In practice, however, the external dataset may contain users that are not in the ISP dataset. To this end, the performance of all de-anonymization algorithms (including ours) is an upper bound. The above experiments have demonstrated the advantage of our proposed algorithms based on the relative comparisons with existing algorithms.

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 of the given service (GM-B algorithm). Specifically, the total performance gain in terms of hit-precision is more than 17% compared with the existing algorithms. Further, by adding historical check-ins and location context, another 30% to 150% relative gain can be achieved. Finally, we show that the proposed algorithms are robust against the parameter settings of the models. The result suggests that even without ground-truth data to estimate parameters, our proposed algorithms will stay robust using empirical parameters.

IX. Discussion & Conclusions

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 underestimate 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 have shown promising performance, which confirms our insights.

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. On the other hand, in order to provide better location privacy protections, practical factors should also be considered. Our result shows that both user mobility patterns and location context 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 and location

context to provide stronger privacy guarantees (*e.g.*, using differential privacy [12]). As for the further work, we plan to investigate de-anonymization attacks by considering other types of external information, $e.g.$, social graphs [19], [20], [29], [35] or user's home and work addresses and designing better privacy protection mechanisms.

ACKNOWLEDGMENT

The authors want to thank the anonymous reviewers for their helpful comments. This work was in part supported by the NSF grant CNS-1717028.

REFERENCES

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

[2] ——, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *Proc. IEEE ICDE*, 2008.

[3] G. Acs and C. Castelluccia, "A case study: Privacy preserving release of spatio-temporal density in paris," in *Proc. ACM KDD*, 2014.

[4] E. Akim and D. Tuchin, "Gps errors statistical analysis for ground receiver measurements," in *Proc. ISSFD*, 2003.

[5] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proc. ACM CCS*, 2013.

[6] N. Banerjee, A. Rahmati, M. Corner, S. Rollins, and L. Zhong, "Users and batteries: interactions and adaptive energy management in mobile systems," *Proc. ACM UbiComp*, 2007.

[7] C. M. Bishop, "Pattern recognition," *Machine Learning*, vol. 128, pp. 1–58, 2006.

[8] A. Cecaj, M. Mamei, and F. Zambonelli, "Re-identification and information fusion between anonymized cdr and social network data," *Journal of Ambient Intelligence and Humanized Computing*, vol. 7, no. 1, pp. 83–96, 2016.

[9] 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, p. 1376, 2013.

[10] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel, "Identification via location-profiling in gsm networks," in *Proc. ACM WPES*, 2008.

[11] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem, "Dynamic population mapping using mobile phone data," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 111, no. 45, pp. 15 888–15 893, 2014.

[12] C. Dwork, "Differential privacy: A survey of results," in *Proc. TAMC*, 2008.

[13] L. Edmond, "Traité de criminalistique," *Lyon.: Joannes DESVIGNE et ses FILS*, 1931.

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

[15] 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. WWW*, 2013.

[16] 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. ACM KDD*, 2015.

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

[18] M. Gramaglia and M. Fiore, "Hiding mobile traffic fingerprints with glove," in *Proc. ACM CoNext*, 2015.

[19] 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. NDSS*, 2015.

[20] S. Ji, W. Li, M. Srivatsa, and R. Beyah, "Structural data de-anonymization: Quantification, practice, and implications," in *Proc. ACM CCS*, 2014.

[21] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE ICDE*, 2007.

[22] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 6, pp. 1067–1080, 2007.

[23] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao, "Privacy vulnerability of published anonymous mobility traces," *IEEE/ACM Transactions on Networking (TON)*, vol. 21, no. 3, pp. 720–733, 2013.

[24] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 3, 2007.

[25] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press, 2008, vol. 1, no. 1.

[26] 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. ACM KDD*, 2016.

[27] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Where you are is who you are: User identification by matching statistics," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 11, no. 2, pp. 358–372, 2016.

[28] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE SP*, 2008.

[29] ——, "De-anonymizing social networks," in *Proc. IEEE SP*, 2009.

[30] 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 CCS*, 2017.

[31] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *Proc. WWW*, 2016.

[32] L. Rossi and M. Musolesi, "It's the way you check-in: identifying users in location-based social networks," in *Proc. ACM WOSN*, 2014.

[33] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *Proc. IEEE SP*, 2011.

[34] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *SIGKDD Explor. Newsl.*, vol. 18, no. 2, 2017.

[35] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: Using social network as a side-channel," in *Proc. ACM CCS*, 2012.

[36] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[37] J. A. Thomas and T. M. Cover, *Elements of information theory*. John Wiley & Sons, 2006.

[38] D. Valcarce, J. Parapar, and Á. Barreiro, "Additive smoothing for relevance-based language modelling of recommender systems," in *Proc. CERI*, 2016.

[39] 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. ICWSM*, 2016.

[40] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proc. ACM MobiCom*, 2011.

[41] Z. Zhang, L. Zhou, X. Zhao, G. Wang, Y. Su, M. Metzger, H. Zheng, and B. Y. Zhao, "On the validity of geosocial mobility traces," in *Proc. ACM HotNets*, 2013.

[42] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5g," *IEEE Network*, vol. 30, no. 1, pp. 44–51, 2016.