No More than What I Post: Preventing Linkage Attacks on Check-in Services

Fengli Xu, Yong Li, *Senior Member, IEEE*, Zhen Tu, Shuhao Chang, and Hongjia Huang

Abstract—With the flourishing of location based social networks, posting check-ins has become a common practice to document one's daily life. Users usually do not consider check-in records as violations of their privacy. However, through analyzing two real-world check-in datasets, our study shows that check-in records are vulnerable to linkage attacks. Specifically, adversary is able to uniquely re-identify over 52∼66 percent users in other anonymous mobility datasets and 60∼80 percent users have more than 60 percent probability leaking unreported mobility records. In addition, we further demonstrate that the privacy sensitivity of check-in records can be more accurately measured by including the information of additional mobility data compared with only looking at check-ins. Based on this observation, we design a *partition-and-group* framework to integrate the information of check-ins and additional mobility data to attain a novel privacy criterion—$k^{\tau,l}$*-anonymity*. It ensures adversaries with arbitrary background knowledge cannot use check-ins to re-identify users in other anonymous datasets or learning unreported mobility records. The proposed framework achieves favorable performance against state-of-art baseline in terms of improving check-in utility by 24∼57 percent while providing stronger privacy guarantee at the same time. We believe this study will open a new angle in attaining both privacy-preserving and useful check-in services.

Index Terms—Check-ins, privacy-preserving data publishing, linkage attacks, mobility data privacy

1 Introduction

Check-in service has now become a popular feature that is widely adopted by the mainstream social media platforms, such as Facebook, Twitter and Wechat. It facilitates users to document their daily activities with mobility trace and share them with public audience. Users usually do not associate the self-reported check-ins with privacy risks, since they only check-in to places they feel comfortable [1]. However, the uniqueness of human mobility often exposes their check-in records to linkage attacks, i.e., revealing their identities and unreported mobility records in other anonymous mobility datasets, such as call detail records [2], transportation records [3], and credit card records [4]. Moreover, recent researches showed that most users are unaware or not able to fully anticipate the privacy risks embedded in posting check-ins [5]. Therefore, it is a paramount task for the check-in service providers to quantify the potential privacy exposures and put forward feasible solutions.

Previous efforts attempt to address the problem of linkage attacks on mobility data by ensuring user's anonymity in anonymous mobility datasets [4], [6]. That is, making sure adversary cannot achieve unique linkages based on user's check-ins through generalizing the records in anonymous mobility datasets. However, such approach often requires unacceptable data utility degeneration [7], and cannot prevent adversary from learning additional unreported mobility records [8]. It is also unrealistic for users and check-in service providers to assume all anonymous mobility datasets have been properly sanitized, since studies repeatedly demonstrated that insecure datasets had been irreversibly spread across the Internet [9], [10]. Therefore, these findings suggest it is impractical to prevent linkage attacks by sanitizing anonymous mobility datasets. In this paper, we investigate and address this problem through a novel angel—looking at the public mobility records, i.e., check-ins.

To better understand the underlying mechanism of linkage attack, we conduct extensive experiments on two large scale real-world check-in traces in parallel with user's additional mobility data from two mainstream social media platforms—WeChat and Weibo. It allows us to make the following important observations. First, check-in records have severe privacy exposures to linkage attack. Specifically, 52 percent users in WeChat and 66 percent users in Weibo can be uniquely re-identified with their check-ins, while 60 percent users in WeChat and 80 percent users in Weibo suffer from leaking unreported mobility records with over 60 percent probability. Second, such grave privacy risks result from the high uniqueness in both check-ins and mobility data. Interestingly, we find out that the root causes for the highly unique check-ins and mobility data are exactly the opposite—too little check-ins in total and too many mobility records per users. Third, we find out that the privacy sensitivity of check-ins measured by only looking at check-in uniqueness is an upper bound of actual privacy risks. One can significantly improve the estimation

• *The authors are with Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: {xfl15, tuz16, chang-sh15, hhj15}@mails.tsinghua.edu.cn, liyong07@tsinghua.edu.cn.*

Manuscript received 26 Feb. 2019; revised 8 Aug. 2019; accepted 7 Oct. 2019. Date of publication 15 Oct. 2019; date of current version 7 Jan. 2021. (Corresponding author: Yong Li.)
Digital Object Identifier no. 10.1109/TMC.2019.2947416

of privacy sensitivity by introducing moderate amount of additional mobility data. These findings inspired us to design a better-informed privacy mechanism for check-in services by integrating the information of check-ins and additional mobility data.

In this paper, we put forward several contributions to attain both privacy-preserving and useful check-in services. First, we extend the frameworks of *k-anonymity* [11] and *l-diversity* [8] in check-in privacy preserving, and devise a novel privacy criterion $k^{\tau,l}$-anonymity. It ensures the posted check-ins cannot be exploited to distinguish user from at least other $k-1$ users in any anonymous mobility datasets, and for any time window of duration τ user's actual locations are indistinguishable from at least other $l-1$ locations. Second, we further propose a *partition-and-group* framework to optimize the check-in utility under $k^{\tau,l}$-anonymity privacy guarantee by carefully partitioning user population into small anonymity groups. Third, we conduct a thorough trace-driven evaluation on the proposed framework based on the real-world datasets. The evaluation results demonstrate that our framework significantly outperforms state-of-art baseline methods in terms of achieving 24$\sim$57 percent check-in utility improvement while providing stronger privacy guarantee in the same time. In addition, 32$\sim$62 percent check-in utility boost of our framework is achieved by introducing additional mobility data, which showcases the benefits of integrating additional mobility data in privacy-preserving check-in service. Finally, our study reveals two intriguing trade-offs between the utility and privacy in check-in services: (i) in order to achieve modest privacy gains, users need to sacrifice significant check-in utility, i.e., reducing spatio-temporal resolution of check-ins. (ii) users may increase the utility of their check-ins with same privacy level by letting check-in service providers to collect moderate amount of additional mobility data. Such findings may have direct implications on how to defend linkage attacks with the joint effort of check-in service providers and individual users.

2 Related Works

We summarize and discuss the most relevant literature from the following three aspects.

Linkage Attack. The linkage attacks were widely studied in multiple scenarios and had received increasing attention in recent years [8], [11], [12], [13]. The most prominent two branches are *re-identification attack* and *probabilistic attack* [14]. Specifically, the *re-identification attack* aims at recovering individuals' identities in anonymous datasets by achieving unique linkages with public datasets. For example, 87 percent of American population can be uniquely re-identified with the public accessible information of ZIP code, gender and date of birth [11]. Similar findings have been established in wide range of scenarios, including web browsing records [15], call detail records [7], app usage records [13], [16], social media profile [17] and so on. One popular privacy model against such attack is k-anonymity, which requires to make the records of each individual indistinguishable from at least $k-1$ others [11]. On the other hand, *probabilistic attack* is a more general form linkage attacks, which aims at improving some belief on individuals through correlating the datasets. Researchers demonstrated that by combining online social network data and sparse offline location data individual's locations can be predicted with high precision [18]. In addition, user's identities across different online social media sites can be associated with high accuracy based on the user generate content [17], [19], e.g., public profile and images. Moreover, the salary class of individuals can be accurately inferred by correlating census data with public available information [8]. To defend such attacks, l-diversity and t-closeness have been proposed to ensure the diversity on the sensitive information within each anonymity group [8], [12], [20], [21], [22].

We position our study in a novel scenario of defending against the linkage attacks on social media check-ins. We aim to provide strong privacy guarantee for check-in service against both *re-identification attack* and *probabilistic attack*, and design privacy solution compatible with unstructured spatio-temporal data.

Mobility Data Privacy. The literature in this area can be further broke down into two categories: aggregate mobility data privacy and individual mobility data privacy. As for the former, recent study found evidence that aggregate mobility data suffers from the risks of leaking individual trajectories [13], [23]. In addition, *differential privacy* has been applied on aggregate mobility data to provide provable privacy guarantees for individuals [24], [25]. As for individual mobility privacy, *geo-indistinguishability* model is devised to achieve practical privacy guarantee in individual mobility collecting [26], [27]. In addition, vast amount of literature were dedicated to ensure location anonymity in the context of geo-referenced queries in location based service [28], [29], [30]. However, these privacy models aim to prevent attackers to infer certain mobility records of individuals, but provide no privacy guarantee against the linkage attack on trajectories [31]. On the contrary, *cloaking*, *generalization* and *suppression* techniques are leveraged to achieve k-anonymity in releasing anonymous individual trajectories [6], [32], [33], [34], [35]. However, recent studies showed that such approaches will likely to result in significant data utility degeneration [7], [36].

We tackle the specific problem of designing privacy-preserving check-in service, which is closely related to prior effort in individual trajectories releasing. However, it differs from previous works that users have strong requirement for check-in utility and social desirability, which makes the differential privacy framework not applicable and poses a pressing need for novel privacy criterion.

Privacy in Check-in Services. The privacy risk in posting check-ins is more subtle than most data publishing scenarios, since users often cannot fully anticipate the privacy threat and only realize the exposure in regret [1], [5]. Researchers have shown that anonymization does not help in check-in scenario since both social relationship information and location information can be leveraged to reveal user's identity [37], [38]. Moreover, previous researches have demonstrated that linkage attack is empirically feasible on check-in records [39], [40]. Previous solutions mainly focus on application specific privacy preserving mechanisms, such as privacy-preserving personalized location based services [41], preventing social-based location inference attacks [42], and so on. One closely related work

demonstrated that check-ins can be leveraged to re-identify anonymous call detail records through linkage attack [2]. The authors suggested this problem could be addressed by generalizing the anonymous call detail records, which was in line with rich literature dedicated to achieve k-anonymity on anonymous mobility data [6], [33], [34].

Different from the prior works that targeted at application specific privacy mechanisms, we investigate a more general form privacy problem in check-in services, i.e., the linkage attack with arbitrary anonymous mobility data. Moreover, instead of analyzing the feasibility of linkage attack models, we aim to propose an effective privacy model. In addition, our research differs from the second branch of related works by addressing the privacy problem through a new angle—looking at the check-in data.

3 Problem Formulation

3.1 Objectives

The ultimate objective of our study is to prevent the linkage attacks on social media check-ins, and design privacy-preserving and useful check-in services. We further break it down into three parts:

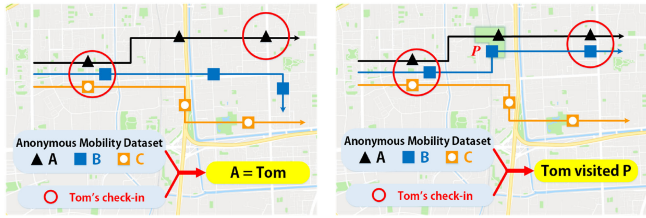
- *Preventing linkage attacks*: The posted check-ins should be privacy-preserving against linkage attacks. That is, adversary with arbitrary background knowledge is limited by user specific bound to enlarge their knowledge about individuals through accessing their check-ins.
- *Truthfulness at record level*: In the context of check-in services, social desirability plays an important role. Thus, we aim to realize *truthfulness at record level* [14], which indicates we cannot distort check-in records into mobility records that never happen or inject fake check-ins to protect user privacy.
- *Reasonable trade-off between privacy and utility*: In order to achieve practical solutions, we require the privacy mechanism to strike a reasonable trade-off between privacy and utility. It implies that the utility of check-in records should remain at desirable level when strong privacy guarantee is provided.

3.2 Attacker Model

To achieve robust privacy mechanism, we first assume that the adversary may have arbitrary background knowledge on each individual user's mobility data, including those the check-in service providers are not aware of. Then, the linkage attacks on check-in services are further classified into two categories:

Re-Identification Attack. The adversary attempts to recover user's identity in other anonymous mobility datasets by achieving a unique linkage between user's anonymous mobility data and public check-ins. The attack procedure is illustrated in Fig. 1a. It is worth pointing out that recent research successfully showcased such attack in real-world scenario, where hundreds of individuals in an anonymous call detail records are uniquely re-identified with 90 percent confidence by leveraging public accessible check-ins [2].

Probabilistic Attack. This attack aims to enlarge adversary's knowledge on individuals, i.e., learning additional mobility



(a) *Re-identification attack* (b) *Probabilistic attack*

Fig. 1. Illustration of linkage attacks on check-ins.

records they do not report. The adversary may successfully perform probabilistic attacks even if he does not achieve unique linkages, i.e., re-identifying users. For example, if two users have exactly the same trajectories in an anonymous mobility dataset, then it is impossible to distinguish them from each other based on their check-ins. However, adversary can still know for sure that the users have visited the locations shared by their trajectories. The attack procedure is illustrated in Fig. 1b. Such attack is more subtle, but no less dangerous.

3.3 Privacy Model

We first introduce the privacy framework of k-anonymity and l-diversity, which is the basis of our model. Then, we elaborate on our novel privacy criterion—$k^{\tau,l}$-anonymity.

k-anonymity. The k-anonymity framework is originally devised to defend the re-identification attacks in relational database [11]. It requires data sanitizing techniques that render each individual's attributes indistinguishable from at least other $k-1$ individuals', which forms an anonymity group that prevents any individuals within to be uniquely re-identified. However, k-anonymity is known to be powerless against *probabilistic attacks*, since each individual is hidden in a crowd that lacks of diversity [8].

l-diversity. To make up for the short-coming of k-anonymity, l-diversity is put forward to ensure users' diversity on sensitive attributes within each anonymity group [8]. Specifically, it first classifies the attributes into sensitive and non-sensitive types. Then, it requires each individual cannot be uniquely re-identified with the non-sensitive attributes, while the sensitive attributes should be of at least l different categories within each anonymity group. The idea of discriminating between sensitive information and non-sensitive information is a natural fit to our application, since the self-report check-ins are usually considered non-sensitive and unreported locations otherwise. However, l-diversity is also meant for relational database and not able to be applied on unstructured and continuous spatio-temporal data.

$k^{\tau,l}$-anonymity. Inspired by the insights and limitations of previous models, we design a novel privacy criterion *$k^{\tau,l}$-anonymity* to address the privacy issues in check-in services. Specifically, *$k^{\tau,l}$-anonymity* requires:(i) any users on social media cannot be distinguished from at least other $k-1$ users in any other anonoymous mobility datasets based on their public check-ins; (ii) for any time window of duration τ users' unreported locations cannot be discriminated from at least $l-1$ other potential locations. Therefore, the knowledge adversary can acquire through linkage attacks, i.e., users identity in other anonymous mobility datasets and unreported mobility records, is effectively

bounded by user specific parameters k, τ, l. In other words, $k^{\tau,l}$-*anonymity* is able to provide strong privacy guarantee against both *re-identification attack* and *probabilistic attack*.

Note that there is another popular privacy framework for mobility data protection, i.e., differential privacy [10]. The main reasons we do not follow this framework are two folds. First, differential privacy is designed to obfuscate certain mobility records instead of preventing trajectory linkage attack [26], [31]. That is it prevents the attackers to infer the actual location of users but does not provide rigours privacy bound for trajectory linkage, which is not applicable in our scenario. Second, current differential privacy methods mainly rely on injecting noises to mobility records [43]. Thus, it might generate false check-in records, which is against the *Truthfulness at record level* objective. On the other hand, the $k^{\tau,l}$-anonymity privacy criterion is tailored to prevent *re-identification attack* and *probabilistic attack*, and it can meet all the objectives theoretically.

4 Datasets

4.1 Data Collection

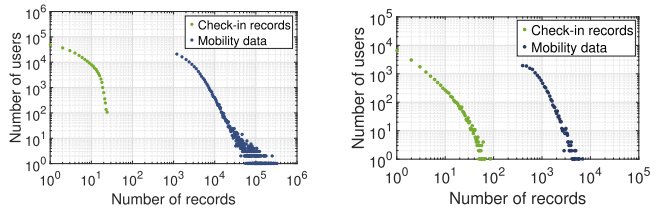
We utilize two real-world datasets collected from large scale user population in two mainstream social media platforms: WeChat and Weibo. The detailed information is discussed as follows.

WeChat Dataset.[1] WeChat platform is currently the most popular social media platform in China. This dataset consists of 530,050 check-ins collected from 100,000 users, which are randomly selected from the general user population spread across Beijing city. It covers two and a half month of usage, i.e., from Jan. 1 to Mar. 15, 2018. We also collect an additional mobility dataset including over 193 millions mobility records from same user population during same time period.

Weibo Dataset.[2] This dataset is collected by a previous research [40]. It contains 11,866,425 mobility records and 78,412 check-ins on Weibo platform from 17,425 users located in Shanghai during one week, i.e., from Apr.19 to Apr.26, 2016. Different from WeChat dataset, the Weibo dataset is collected by internet service provider by performing deep packet inspection on cellular traffic.

It is worth pointing out that the additional mobility records are collected through all the integrated location based services in these two platforms. Take the all-in-one WeChat platform as an example, whenever users invoke the embedded location based services like online map service, car hailing and takeout ordering they generate a mobility records in dataset. Therefore, the additional mobility data is a superset of check-in records, which constitutes a more fine-grained mobility trace and makes it suitable to model the arbitrary background knowledge of the attackers.

Note that the collected mobility data is a sample of user's complete mobility behavior. Therefore, the privacy exposure measured on these datasets captures a lower bound of potential exposure. However, our privacy mechanism is still set to provides privacy guarantee when arbitrary anonymous mobility data is presented. On the other hand, we perform standard prepossessing to format the datasets to accommodate the different spatio-temporal resolution results from different data sources. Specifically, we transform the GPS coordinates into grid IDs by dividing the city into grids of 1 km^2, and replace the timestamps with time slot IDs by dividing the total duration into 1 hour time slots. To demonstrate the basic statistics of datasets, we show the probability distribution function (PDF) of number of mobility records and check-in records of each user in Fig. 2. From the result, we can observe that the additional mobility records are denser then the check-in records, since number of records is 2 3 magnitude higher. In addition, both the check-in records and mobility records follow a well-defined power distribution. These observations are consistent with the previous findings on check-ins from Foursquare and twitter [44] as well as the mobility patterns extracted from call detail records [45] and credit card records [4]. It indicates our leveraged datasets are representative of typical check-in and mobility behavior, and our findings can be generalized across different platforms.



(a) WeChat dataset (b) Weibo dataset

Fig. 2. The distribution of the number of user's check-ins and mobility records.

4.2 Ethic Consideration

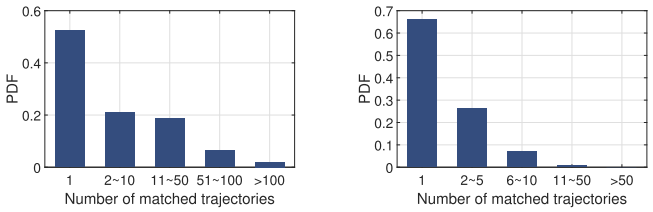
We take careful steps to address privacy issues regarding the sharing and mining of user data. First, the terms of service for WeChat and internet service provider include consent for research studies. Tencent, the parent company of WeChat, and authors of [40] shared user data after preprocessing the data to protect user privacy. All user identifiers have been replaced with the secure hashcodes to improve anonymity. Second, our research protocol has been reviewed and approved by our local university institutional board. Third, all data is stored in a secure off-line server. Only authorized members in the research team can access the datasets, and all research procedures are bound by strict non disclosure agreements.

5 Understanding the Privacy Exposure in Check-ins

5.1 Measuring Privacy Exposure

We first measure the privacy exposures to re-identification attack. Since adversary attempts to uniquely re-identify users in anonymous mobility dataset through such attack, we measure the exposures as the number of trajectories each user's check-in records match to in the anonymous mobility datasets. For example, if there is only one matched trajectory, it indicates a successful re-identification on revealing the identities of targeted users. We plot the PDF of the number of matched trajectories on both WeChat and Weibo

1. [Online]. Available: https://weixin.qq.com/
2. [Online]. Available: https://weibo.com/



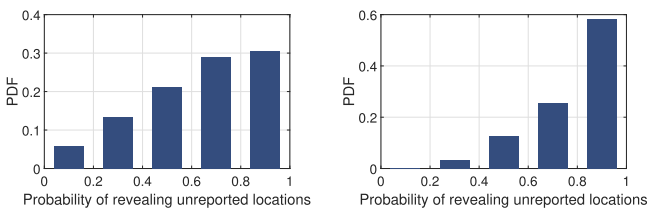
(a) WeChat dataset (b) Weibo dataset

Fig. 3. Privacy exposure to re-identification attack.

datasets in Fig. 3. For the Wechat datasets, we can observe that 52 percent users' anonymous trajectories can be uniquely re-identified with their check-ins. In addition, Fig. 3b shows 66 percent of users in Weibo can be uniquely re-identified. These results indicate that public check-ins have severe privacy exposure in leak users' identity in other anonymous mobility datasets.

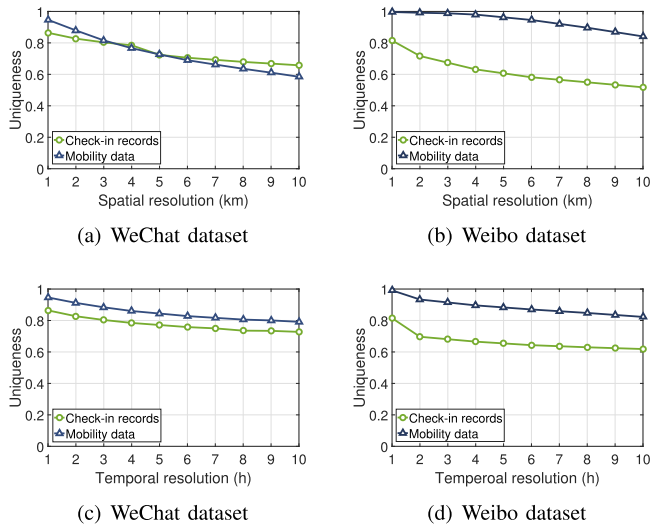
Then, we further quantify the privacy exposure to probabilistic attack in users' check-in records, i.e., the likelihood of revealing unreported mobility records. Previous analysis shows more than 30 percent users' check-ins will match to more than one anonymous trajectories. However, in a given time slot, if all the matched trajectories visit a same location, then the adversary can make a safe bet that the targeted user has been there. In other words, the more diverse visited locations of the matched trajectories in a given time slot, the lower risk of revealing unreported mobility records of the targeted users. Assume there are l different locations visited by matched trajectories, we calculate the probability of revealing unreported mobility records as $1/l$. For each user, we compute the probability of revealing unreported mobility records in each time slot and average them across the time duration to measure the exposure to probabilistic attack. Fig. 4 shows the PDF of each user's probability in revealing unreported mobility records. Take Weibo data for example, 60 percent users have a 80∼100 percent probability of revealing unreported mobility records, which means most of their records in anonymous mobility datasets can be accurately recovered through probabilistic attacks. In addition, 60 and 80 percent users have a more than 60 percent probability in revealing unreported mobility records for WeChat and Weibo data, respectively. All these results demonstrate that public check-in records also face serious privacy exposure to probabilistic attack.

The underlying reason of such grave privacy risk is user's highly unique check-in behavior and mobility data. It allows adversary to achieve high quality linkage between check-in records and anonymous mobility datasets. A common strategy to address this problem is to decrease the uniqueness by generalization [7], [36], i.e., lowering the



(a) WeChat dataset (b) Weibo dataset

Fig. 4. Privacy exposure to probabilistic attack.



(a) WeChat dataset (b) Weibo dataset

(c) WeChat dataset (d) Weibo dataset

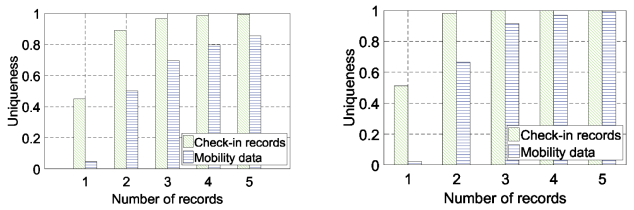
Fig. 5. Naive spatio-temporal generalization does not work in preventing linkage attacks on check-ins.

spatial or temporal resolution. However, Fig. 5 shows simple generalization does not provide effective privacy enhancement. Here, the uniqueness is measured by the percentage of users can be uniquely re-identified with their check-in records or random ten mobility records. For WeChat check-in records, we can observe that a 9 km decrease in spatial resolution only reduces the uniqueness by 23 percent, which is still over 65 percent. In addition, a 9h decrease of temporal resolution can only bring a 15 percent decrease of the uniqueness. For Weibo dataset, the decrease of spatial and temporal resolution also provides similar small decrease in uniqueness. These results indicate simple generalization does not work in preserving check-in privacy, since it requires significant utility loss to achieve modest privacy improvement.

5.2 Finding a Novel Angle

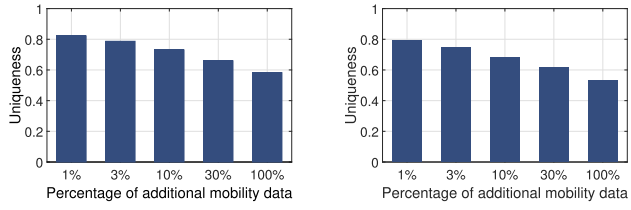
Previous observations motivate us to explore a novel angle to attain both privacy-preserving and useful check-in services. We start by investigating the root causes of such high uniqueness in both check-ins and mobility data.

Although check-ins and mobility data have similar uniqueness level, they are significantly different in the number of records per user. Therefore, we first look at the impact of record number on data uniqueness. Specifically, we randomly sample 1∼5 records from each user's check-ins and mobility data respectively, and present the uniqueness among the sampled data in Fig. 6. From the results, we can observe that the uniqueness of check-ins remains relatively high when the number of records per user decreases from 5 to 1 in both WeChat and Weibo datasets. On the contrary, the uniqueness of mobility data reduces significantly as the number of records decreases. As the number of records decreases from 5 to 1, the level of uniqueness decreases from 86 to 5 percent in WeChat mobility dataset and from 99 to 2 percent in Weibo mobility dataset, respectively. It indicates that the number of records per user is the deciding factor in the uniqueness of mobility data. In another word, the root cause of highly unique mobility data is each individual has too many mobility records.



(a) WeChat dataset (b) Weibo dataset

Fig. 6. Uniqueness of check-in and mobility data on different number of records.



(a) WeChat dataset (b) Weibo dataset

Fig. 7. The impact of additional mobility data on check-in uniqueness.

On the other hand, check-in records are inherently different from mobility data in terms of number of record, where check-ins are two magnitude fewer than mobility records in both datasets. It inspires us that the root cause for highly unique check-ins could be too few "crowd" for users to hide into, i.e., the number of check-ins in total is too small. To validate this effect, we add in additional mobility records randomly sampled from mobility datasets into check-in dataset, and present the uniqueness of check-ins with different amount of additional mobility records in Fig. 7. From the results, we find out that the uniqueness of check-ins decreases rapidly when moderate amount of additional mobility records are added in. Specifically, when $1 \sim 100$ percent total mobility records are added, the uniqueness of check-ins decreases from 82 to 58 percent in WeChat dataset and from 79 to 52 percent in Weibo dataset, respectively. These results indicate the root cause of highly unique check-ins is the sparsity of check-ins, which is in consistent with our assumptions. These analyses reveal an insightful finding that the root causes for high uniqueness in mobility data and check-ins are the exact opposite—too many mobility data and too few check-in records.

Inspired by these observations, we propose to measure the privacy sensitivity of check-ins as the uniqueness of check-ins in the context of additional mobility data. Specifically, a user's check-ins are unique if they do not co-locate with any other users' check-ins or unreported mobility records. We remark that the check-in uniqueness in the context of additional mobility data is a more fine-grained measurement on the privacy sensitivity compared with only looking at check-ins. The reason is that adversary cannot achieve unique linkages on users when their check-ins co-locate with other users' unreported mobility records, even if they are unique by only looking at check-ins. The uniqueness measured solely on check-ins is actually an upper bound of privacy sensitivity estimated with incomplete information.

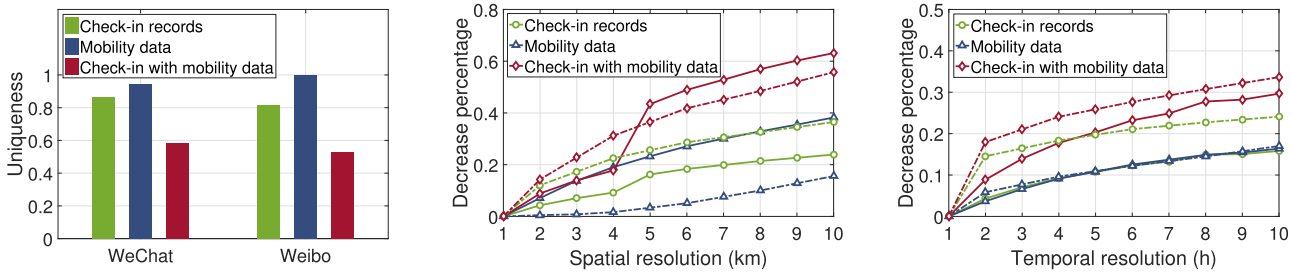
To showcase the potential performance gain by including additional mobility data, we present the comparison of privacy sensitivity estimation in Fig. 8. From the result in Fig. 8a, we find out that in WeChat dataset the check-in uniqueness measured with additional mobility data is 58 percent, which is significantly lower than the 86 percent measured among solely check-in records and 95 percent measured among solely mobility data. Similar results are also observed in Weibo dataset. In addition, Fig. 8b and 8c demonstrate that in both WeChat and Weibo dataset the uniqueness of check-ins measured with mobility data decreases significantly faster when spatial and temporal generalization is applied. These fine-grained privacy sensitivity measurements not only help us to better quantify the privacy exposure in posting check-ins, but also has the potential to facilitate better-informed privacy preserving mechanisms.

6 Solution

We first formally denote the variables we use throughout the paper. Then, we describe three basic sanitizing operations and check-in utility cost function that our system is built upon. Finally, we design an algorithm, denoted by $k^{\tau,l}$-*merge*, to efficiently implement $k^{\tau,l}$-anonymity on check-in data, and further propose a *partition-and-group* algorithm to optimize check-in utility under privacy guarantee.

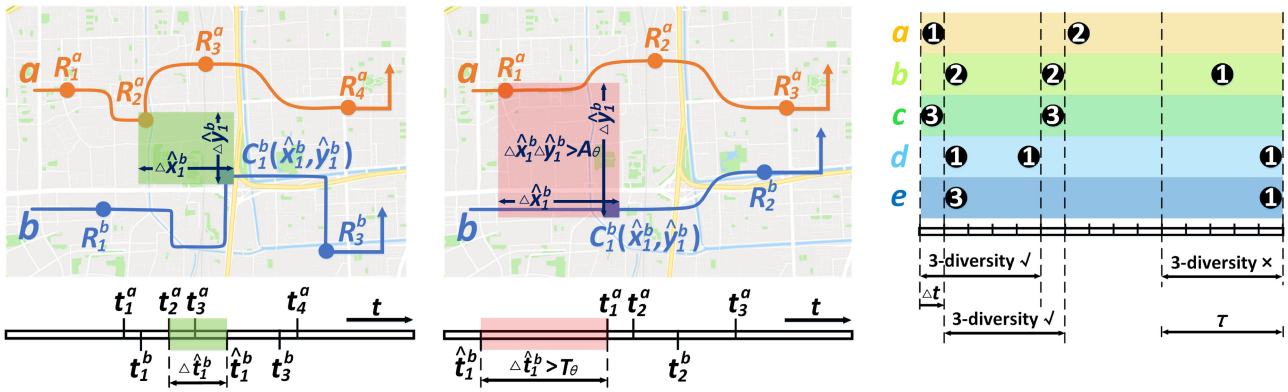
6.1 Definitions

Formally, we define the additional mobility data of user i as $R^i = \{r_m^i\}$, where r_m^i is the mth record of user i. It can be expressed as a tuple $r_m^i = (x_m^i,\ y_m^i,\ t_m^i)$, with x_m^i, $y_m^i$and t_m^i denoting the longitude, latitude and time stamp, respectively. On the other hand, we denote the *check-in records* as $C^i = \{c_m^i\}$, where c_m^i is the mth check-ins of user i. Since the check-in records after sanitization may have various spatial and temporal resolution, c_m^i is defined as $(\hat{x}_m^i,\ \Delta\hat{x}_m^i,\ \hat{y}_m^i,\ \Delta\hat{y}_m^i,\ \hat{t}_m^i,\ \Delta\hat{t}_m^i)$, with $[\hat{x}_m^i,\ \hat{x}_m^i + \Delta\hat{x}_m^i] \times [\hat{y}_m^i,\ \hat{y}_m^i + \Delta\hat{y}_m^i]$



(a) Original uniqueness (b) Spatial generalization (c) Temporal generalization

Fig. 8. Check-in uniqueness measured with additional mobility data. Solid line denoting WeChat dataset and dash line denoting Weibo dataset.



(a) *Generalization* (b) *Suppression* (c) *Diversity check*

Fig. 9. The illustrations of three basic sanitizing operations.

and $[\hat{t}_m^i,\ \hat{t}_m^i + \Delta\hat{t}_m^i]$ denoting the coverage in spatial and temporal dimensions, respectively.

6.2 Basic Operations

To accommodate the requirement of *truthfulness at record level*, we limit our data sanitizing techniques to *generalization* and *suppression*, i.e., addressing the privacy problem by reducing check-in's spatiotemporal resolution or leaving out check-ins. Such operations avoid adding noises to check-in records that may displace users to places they never been to or injecting fabricated check-ins, which maintains the integrity of check-ins and avoid compromising their social figures. On the other hand, to effectively defend against *probabilistic attack*, we also define a *diversity check* operation to ensure the diversity on sensitive information within anonymity groups. Specifically, the basic operations are described as follows:

Generalization. *Generalization* is to reduce spatial and temporal resolution of check-ins so that they overlap with other user's check-ins or unreported mobility records. In this way, the adversary can no longer uniquely link the check-ins with anonymous mobility data, which effectively prevents the *re-identification attacks*. We define the *generalization* operation as $G(c^\star, r^\star)$, where $c^\star$ and $r^\star$ are check-in and other user's mobility record, respectively. This operation will output *generalized* check-ins, which is demonstrated in Fig. 9a.

Suppression. When the spatial and temporal resolution of check-in records are too low, their utility is diminished. In real-world scenario, some "outlier" check-ins may require significant *generalization* to prevent *re-identification attacks*, which renders the check-ins useless. Specifically, *suppression* operation $S(c^\star)$ will return *true* for leaving out the check-ins $c^\star$ when spatial coverage exceed A_θ or temporal coverage exceed T_θ. That is, the system will recommend users not to post such check-ins. The *suppression* operation is demonstrated in Fig. 9b. Without loss of generality, A_θ and T_θ is set to 1000 km^2 and 120 hours, respectively.

Diversity Check. We define *diversity check* operation as $D(\{R^\star\}, \tau, l)$, with $\{R^\star\}$ denoting the unreported mobility records of the inspected anonymity group. The illustration of *diversity check* is presented in Fig. 9c. Specifically, the operation search the total time duration with a sliding time window of duration τ and step length of minimal time resolution Δt. Then, it computes the number of distinct locations in each time window with each individual contribute at most one distinct location. If there is a time window with less than l distinct locations than the operation returns *false* for failing the diversity check, otherwise it returns *true* for passing.

6.3 Cost Function

Specifically, the cost function is defined as a linear combination of the spatial and temporal coverage of the investigated check-in, which can be computed as follows,

$$U(c_m^i) = \begin{cases} \lambda_a \cdot \sqrt{A} + \lambda_t \cdot T, & if\ A < A_\theta\ and\ T < T_\theta, \\ \lambda_a \cdot \sqrt{A_\theta} + \lambda_t \cdot T_\theta, & otherwise, \end{cases}$$

where $A = |\Delta\hat{x}_m^i| \times |\Delta\hat{y}_m^i|$ and $T = |\Delta\hat{t}_m^i|$ denote the spatial and temporal coverage of generalized check-in. λ_a and λ_t are designed to weigh the spatial and temporal factors. In this study, we set both λ_a and λ_t to 0.5, which indicates 1 km diameter of spatial coverage and 1 hour temporal coverage map to similar cost. In addition, since the check-ins are *suppressed* if their spatial coverage exceed A_θ or temporal coverage exceed T_θ, we set cost function at maximum value to represent complete lost in utility.

6.4 Achieving $k^{\tau,l}$-anonymity

Now, we describe how to attain $k^{\tau,l}$-*anonymity* within a given user group. The remaining problem is finding the optimal set of mobility records each check-in should *generalized* upon, in order to minimize the overall cost on check-in utility. To that end, we propose a greedy algorithm $k^{\tau,l}$-*merge* to achieve $k^{\tau,l}$-*anonymity* while minimizing utility cost, which is presented in Algorithm 1.

The algorithm first performs *diversity check* on the given user group, and only proceed to merge if they pass the check. The diversity on unreported mobility records is ensured by searching for legitimate user groups on whole user population, which is beyond the scope of $k^{\tau,l}$-*merge* algorithm and will be addressed by following components. After that, the algorithm enumerates through all the users within the group, and find an optimal mobility record from each user for the check-in to *generalize* with. This process effectively ensures the check-ins of each individual cannot be exploited to distinguish them from the rest of the group, while keeping the diversity on unreported mobility records within the group, i.e., achieving $k^{\tau,l}$-*anonymity*.

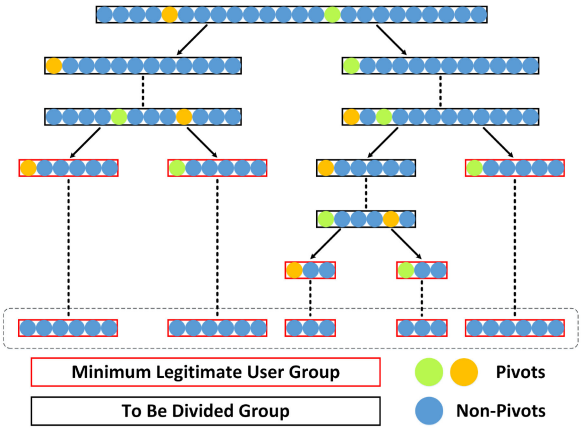


Fig. 10. Illustration of *partition-and-group* framework.

Algorithm 1. $k^{\tau,\,l}$-merge Algorithm

Input: Check-in data $\mathbb{C}$, mobility data $\mathbb{R}$
Input: Anonymity k, diversity l, time window τ
Output: Sanitized check-in data $\mathbb{C}$
if $D(\mathbb{R},\ l,\ \tau) == false$ or $|\mathbb{C}| < k$ **then**
  Return $false$;
else
  $\mathbb{C} \leftarrow \mathbb{C}$;
  foreach $i,\ j \in \mathbb{C},\ j \neq i$ **do**
    foreach $c^* \in \mathbb{C}[i]$ **do**
      $r^* \leftarrow \text{argmin}_{\forall r \in \mathbb{R}[j]} U(G(c^*, r))$;
      $c^* \leftarrow G(c^*, r^*)$;
      $\mathbb{C}[i].update(c^*)$;
    foreach $c^* \in \mathbb{C}[j]$ **do**
      $r^* \leftarrow \text{argmin}_{\forall r \in \mathbb{R}[i]} U(G(c^*, r))$;
      $c^* \leftarrow G(c^*, r^*)$;
      $\mathbb{C}[j].update(c^*)$;
  Return $\mathbb{C}$;

6.5 Partition-and-Group Framework

One key problem in optimizing the privacy mechanism on large scale check-ins is how to partition the user population into optimal anonymity groups. The check-in utility will be significantly improved by carefully classifying the users into numerous small anonymity group that passes *diversity check* compared with putting all of them in one group. We use the word "legitimate" to refer to the anonymity groups that pass the *diversity check*. Achieving the optimal partition of user population requires enumerate all the legitimate anonymity group, which is a NP-hard problem and cannot be readily solved in real-world scenario. We design a novel *partition-and-group* framework to efficiently optimize the check-in utility through a "divide-and-conquer" manner. The idea is to iteratively break down the user population into two small subsets until the *minimum legitimate anonymity groups* are met, which is illustrated in Fig. 10. An important problem is determining whether a anonymity group is *minimum legitimate anonymity group*, i.e., the anonymity group cannot be divided into smaller subsets that all pass *diversity check*. We exploit a convenient property of *diversity check* to address this problem, which is formally described in the following proposition.

Proposition 1. *If an anonymity group does not pass the* diversity check, *then any subsets of this anonymity group will not pass* diversity check.

Proof. Suppose the unreported mobility records of an anonymity group do not pass the *diversity check* of parameters (τ, l). Based on the definition of *diversity check*, there exist at least one time window $[t, t+\tau]$ that the number of distinct locations is less than l. Since the number of distinct locations increases with number of users monotonically, any subsets of inspected anonymity group will have less than l distinct locations in $[t, t+\tau]$. Therefore, any subsets of inspected anonymity group will not pass the *diversity check*. □

Algorithm 2. Partition-and-Group Algorithm

Input: Check-in data $\mathbb{C}$, mobility data $\mathbb{R}$
Input: Anonymity k, diversity l, time window τ
Output: Sanitized check-in data $\mathbb{C}$
foreach $i,\ j \in \mathbb{C},\ j \neq i$ **do**
  $\mathbb{C}^* \leftarrow k^{\tau,\,l}$-merge$(\mathbb{C}[\{i,j\}], \mathbb{R}[\{i,j\}], 2, 0, 0)$;
  $W[a,b] \leftarrow sum([U(c^*) | \ \forall\ c^* \in \mathbb{C}^*])$;
$checkin\_stack.insert(\mathbb{C})$; $mobility\_stack.insert(\mathbb{R})$;
$stop \leftarrow false$;
while $stop \neq false$ **do**
  $\mathbb{C}^* \leftarrow checkin\_stack.pop()$;
  $\mathbb{R}^* \leftarrow mobility\_stack.pop()$;
  if ! divide-2-group$(\mathbb{C}^*, \mathbb{R}^*, W, k, l, \tau)$ **then**
    $checkin\_group.insert(\mathbb{C}^*)$;
    $mobility\_group.insert(\mathbb{R}^*)$;
  else
    $\mathbb{C}_1, \mathbb{C}_2, \mathbb{R}_1, \mathbb{R}_2 \leftarrow$ divide-2-group$(\mathbb{C}^*, \mathbb{R}^*, W, k, l, \tau)$;
      $checkin\_stack.insert(\{\mathbb{C}_1, \mathbb{C}_2\})$;
    $mobility\_stack.insert(\{\mathbb{R}_1, \mathbb{R}_2\})$;
  if $checkin\_stack == \emptyset$ **then**
    $stop \leftarrow true$;
while $checkin\_group \neq \emptyset$ **do**
  $\mathbb{C}^* \leftarrow checkin\_stack.pop()$;
  $\mathbb{R}^* \leftarrow mobility\_stack.pop()$;
  $\mathbb{C}^* \leftarrow k^{\tau,\,l}$-merge$(\mathbb{C}^*, \mathbb{R}^*, k, l, \tau)$;
  $\mathbb{C}.insert(\mathbb{C}^*)$;
Return $\mathbb{C}$;

The above proposition guarantees that an anonymity group is *minimum legitimate anonymity group* if it cannot be further divided into two legitimate subsets, since any subsets of anonymity groups that cannot pass *diversity check* will not pass the *diversity check*. Build upon this proposition, we design the *partition-and-group* framework with the pseudocode presented in Algorithm 2. Specifically, it first computes the cost matrix W, with $W[i,j]$ filled with the cost of achieving 2-anonymity on the check-ins of user i and j with $k^{\tau,\,l}$-*merge* algorithm. Then, it iteratively partition each anonymity group into two subsets with *divide-2-group* algorithm, and when an anonymity group cannot be divided further it is considered as a final anonymity group. Finally, we apply $k^{\tau,\,l}$-*merge* algorithm on each final anonymity group to ensure all users are protected by $k^{\tau,\,l}$-*anonymity*.

In addition, the pseudocode of *divide-2-group* algorithm is presented in Algorithm 3. It first selects out a user i with maximum total cost to other users within the group, and it further selects out a user j with maximum cost to user i. The

selected out users are considered as the pivots of two subsets. Then, the algorithm iteratively picks one remaining user with minimum cost to these two pivots to join their groups until the input anonymity group is equally divided. After that, the *diversity check* is performed on both generated anonymity groups. If both anonymity groups are legitimate, then return them as partition result. Otherwise, if both groups fail the *diversity check*, the input anonymity group is deemed unable to be further divided according to Proposition 1. On the other hand, if only one anonymity group passes the *diversity check*, the failed group keeps borrowing one most distant user from the succeed group, until them both pass or fail the *diversity check*.

6.6 Complexity Analysis

Partition-and-group framework consists of three stages: (i) computing the cost matrix W. (ii) deriving final anonymity groups through iterative divisions. (iii) attaining $k^{\tau,l}$*-anonymity* on each anonymity group with $k^{\tau,l}$*-merge* algorithm. We denote the number of user as N, the average number of check-ins and mobility records of each user as $\bar{q}$ and $\bar{p}$, respectively. Now, we analyze the computation complexity of each stage.

Algorithm 3. Divide-2-Group Algorithm

Input: Check-in data $\mathbb{C}$, mobility data $\mathbb{R}$
Input: Cost matrix W
Input: Anonymity k, diversity l, time window τ
Output: Groups of records $\mathbb{C}_1, \mathbb{C}_2, \mathbb{R}_1, \mathbb{R}_2$
$i \leftarrow \text{argmax}_m\ sum(W[m,:])$;
$j \leftarrow \text{argmax}_m\ W[i,m]$;
$\mathbb{C}_1, \mathbb{C}_2, \mathbb{R}_1, \mathbb{R}_2 \leftarrow partition\_equally(\mathbb{C}, \mathbb{R}, W, i, j)$;
$stop \leftarrow false$;
while $stop \neq false$ **do**
  if $|\mathbb{R}_1| < k$ *or* $|\mathbb{R}_2| < k$ **then**
    Return $false$;
  else if ! $(D(\mathbb{R}_1, \tau, l)$ *or* $D(\mathbb{R}_2, \tau, l))$ **then**
    Return $false$;
  else if ! $D(\mathbb{R}_1, \tau, l)$ **then**
    $\mathbb{C}_1, \mathbb{R}_1 \leftarrow borrow\_one\_user(\mathbb{C}_2, \mathbb{R}_2)$;
  else if ! $D(\mathbb{R}_2, \tau, l)$ **then**
    $\mathbb{C}_2, \mathbb{R}_2 \leftarrow borrow\_one\_user(\mathbb{C}_1, \mathbb{R}_1)$;
  else
    $stop \leftarrow true$;
Return $\mathbb{C}_1, \mathbb{C}_2, \mathbb{R}_1, \mathbb{R}_2$;

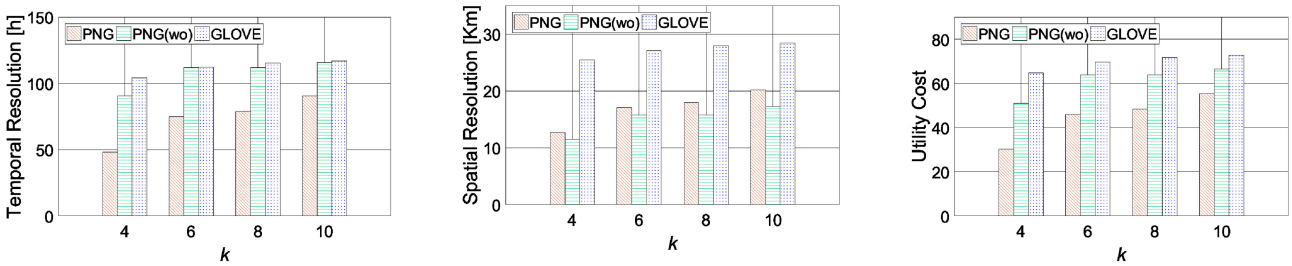
In the first stage, the cost matrix is computed by invoking $k^{\tau,l}$*-merge* for N^2 times. Each time the $k^{\tau,l}$*-merge* need to go through the mobility records and check-ins of two users. Therefore, the overall computation complexity for the first stage is $\mathcal{O}(N^2\bar{q}\bar{p})$. Instead of going through all the mobility records, the $k^{\tau,l}$*-merge* algorithm can only considered the feasible choices, i.e., the mobility records within spatial coverage A_θ and temporal coverage T_θ from target check-ins. Such queries can be achieved in constant time with a hash map function. In addition, the number of check-ins per users is often several magnitude smaller than user number N in practice. Therefore, the actual complexity of the first stage can be approximated as $\mathcal{O}(N^2)$. In the second stage, the number of layers in the division tree is less than $log_2(N)$ because of the binary division, and each layer can be computed in linear time of user number N. Therefore, the overall complexity of the second stage is $\mathcal{O}(Nlog_2(N))$. Finally, we assume the users population are partitioned into m groups, and the average number of users in each anonymity group is N/m. With previous mentioned hash map functions, the average computation complexity of $k^{\tau,l}$*-merge* on each anonymity group is $\mathcal{O}(N^2/m^2)$, and hence the overall complexity is $\mathcal{O}(N^2/m)$. Therefore, the overall computation complexity of *partition-and-group* framework is $\mathcal{O}(N^2)$. The proposed solution can be readily deployed in quadratic time of user population. More importantly, all three stages of the proposed framework are highly parallelizable, which ensures they are scalable to large user population.

7 Evaluation

7.1 Performance Comparison

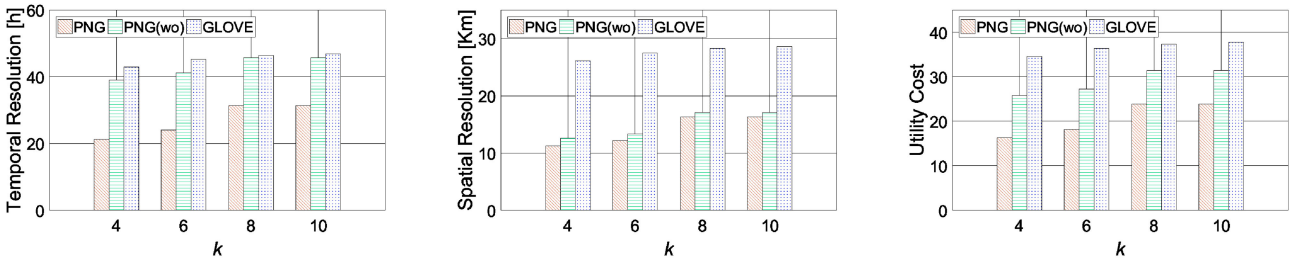
Our solution, denoted by *PNG*, aims to achieve $k^{\tau,l}$*-anonymity* to prevent both *re-identification attack* and *probabilistic attack*. In order to show its superiority, we consider two baselines, i.e., *PNG(wo)* and *GLOVE*. *PNG(wo)* is a degraded version of *PNG*, in the condition that only k*-anonymity* is guaranteed to defend *re-identification attack*. On the other hand, *GLOVE* [6] is a state-of-art solution to achieve same privacy guarantee as *PNG(wo)*. Note that there are several differential privacy based privacy techniques in mobility data protection [26], [31]. We do not compare these models because they are dedicated to protecting certain mobility records, and provide no guarantee against linkage attack, e.g., achieving k*-anonymity*. To compare the performance of these three solutions, we utilize three metrics of average temporal resolution, average spatial resolution and average utility cost of the sanitized check-ins. Note that *GLOVE* cannot be scalable to large populations due to the high computation complexity. In order to ensure fair comparison, we measure the performance of these three solutions based on two subsets with 5,000 users that are randomly sampled from the two investigated datasets, respectively. Note that the Weibo dataset covers a period of one week, while WeChat dataset covers a period of one month. It allows us to evaluate the model's performance on datasets with different characteristics.

We show the performance comparison of these three solutions with different values of k and $l(=k/2)$ in Figs. 11 and 12. We can observe that our *PNG* solution outperforms the other two baselines in all privacy settings. With 4-anonymity and 2-diversity on Weibo and WeChat datasets, the average temporal resolution of sanitized check-ins is 23 h and 48 h, while the spatial resolution is 11 km and 12 km, respectively. Such spatial and temporal resolution is sufficient to meet user's need in documenting their daily life. However, the average spatial and temporal resolution for *PNG(wo)* is much higher, and most of sanitized check-ins from *GLOVE* are too coarse-grained to use with the average temporal resolution reaches as much as 104h. Similar results are observed in average spatial resolution. Furthermore, when it comes to the average utility loss, *PNG* has 24 and 53 percent improvements in the comparison with *PNG(wo)* and *GLOVE* on WeChat dataset. In addition, *PNG* has 27 and 57 percent improvements in the comparison with *PNG (wo)* and *GLOVE* on Weibo dataset. In summary, all these



(a) Temporal resolution (b) Spatial resolution (c) Utility cost

Fig. 11. The performance comparison between our solution and baseline on WeChat data.



(a) Temporal resolution (b) Spatial resolution (c) Utility cost

Fig. 12. The performance comparison between our solution and baseline on Weibo data.

results have demonstrated that our proposed *PNG* solution can significantly reduce check-in utility loss even with a stricter privacy criterion $k^{\tau,\, l}$-*anonymity* is met.

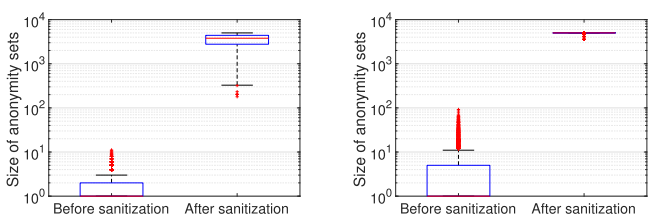
To showcase the privacy gain of the proposed model, we compare the privacy exposure to the re-identification attack and probabilistic attack in Figs. 13 and 14, respectively. From Fig. 13, we observe that the medium value of anonymity set size increases from 1 to 3795 in WeChat dataset and from 1 to 5000 in Weibo dataset. It indicates the privacy exposure to re-identification attack is dramatically reduced, since an average user is protected by an anonymity set of more than 3795 users after the sanitization which makes it improbable for adversaries to uniquely re-identify them. As for the probabilistic attack, Fig. 14a shows 87.9 percent of unreported locations can be inferred with 60 percent accuracy in original WeChat datasets which is reduced to 42.2 percent after the sanitization. Similar observation is made in Fig. 14b that the percentage of locations can be inferred with 60 percent accuracy reduced from 92.1 to 0.6 percent after the sanitization. It showcases the proposed model is effective to prevent probabilistic attack.

To shed light on the empirical runtime, we conduct experiments on the proposed *PNG* model and *GLOVE* on datasets with different number of users and parameter k, and display the empirical runtime in Fig. 15. Note that the experiments are performed with a Intel Xeon E5-2650 CPU. We can observe that the *PNG* model is consistently more efficient than the *GLOVE* model in terms of consuming less CPU runtimes across all datasets and parameter settings. In addition, Fig. 15a shows the runtimes of both *PNG* and *GLOVE* model increase with the number of users, and more importantly the ratio between them increases from 2.8 to 34.74 as the number of users increases from 2500 to 10000. It indicates our proposed model can further save computation time as the dataset scales up. On the other hand, Fig. 15b shows the parameter k does not significantly impact on the overall computation runtimes, which is consistent with our previous complexity analysis.

7.2 Impact of System Parameters

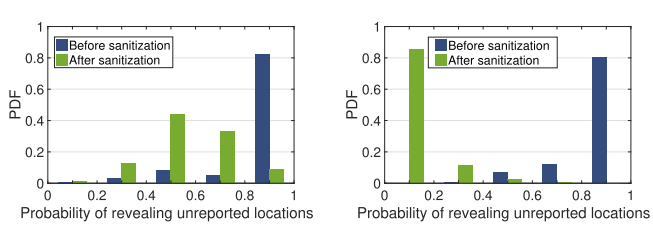
Now we analyze the impact of three key system parameters, i.e., k, l and the hyper-parameters of the cost function, on the performance of our *PNG* solution. To avoid redundancy, we only demonstrate results on WeChat dataset, similar observations are made in Weibo dataset.

First, we measure the performance of *PNG* with different settings of k and l, and show the results in Fig. 16. With a fixed 2-*diversity*, the average temporal resolution, spatial



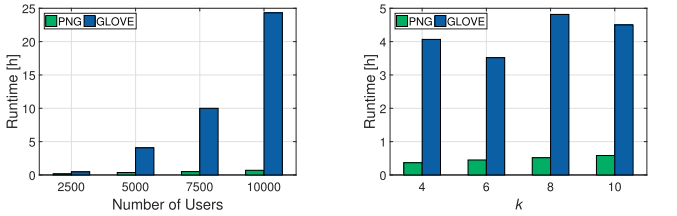
(a) WeChat dataset (b) Weibo dataset

Fig. 13. The comparison on privacy exposure to re-identification attack before and after sanitization.



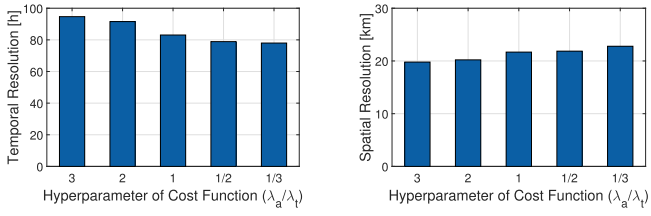
(a) WeChat dataset (b) Weibo dataset

Fig. 14. The comparison on privacy exposure to probabilistic attack before and after sanitization.



(a) Impact of dataset size (b) Impact of parameter k

Fig. 15. Empirical runtime on datasets with different number of users and parameter k.



(a) Temporal Resolution (b) Spatial Resolution

Fig. 17. Impact of hyper-parameters of the cost function.

resolution and utility cost increase monotonously as k grows from 2 to 14. However, further increase of k does not result in a significant check-in utility degeneration, suggesting that achieving a stricter privacy guarantee will only cause limited margin check-in utility loss. In other words, it indicates our solution can achieve favorable check-in utility when strong privacy protection is needed. As for *probabilistic attack*, a lager l indicates stronger privacy protection. For both WeChat and Weibo datasets, a larger l will also cause additional check-in utility loss. However, the additional utility cost for preventing *probabilistic attack* is much smaller when k is of higher value. It indicates the *PNG* framework provides efficient solution to defend both *re-identification attack* and *probabilistic attack*.

Second, we also evaluate the impact of the cost function hyper-parameter on the model's performance, i.e., the impact of λ_a and λ_t. Since only the ratio between them matters in the optimization process, we measure the performance of *PNG* with the ratio ranging from 3 to 1/3, which is demonstrated in Fig. 17. As the ratio between λ_a and λ_t decreases from 3 to 1/3, we can observe the temporal resolution decreases from 94.69 to 77.99 and the spatial resolution increases from 19.79 to 22.79. It indicates the hyper-parameter λ_a and λ_t can indeed tune the *PNG* model to optimize the spatial resolution and temporal resolution, respectively.

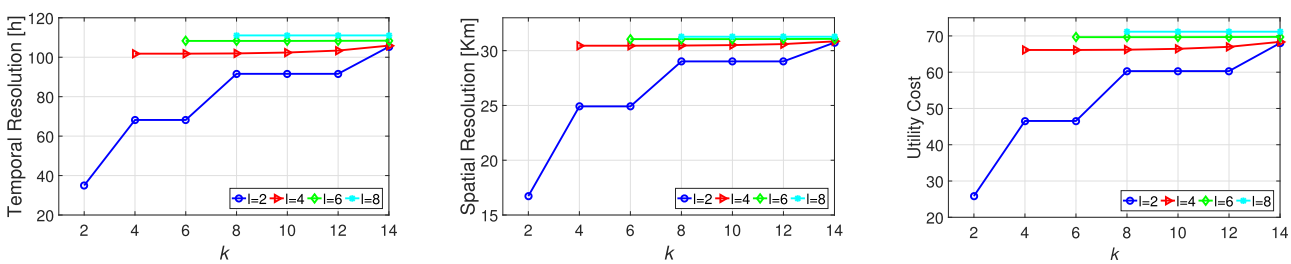
7.3 Performance on Datasets with Different Characteristics

First, we measure the proposed *PNG* model's performance on the datasets with different characteristics to examine its robustness in different scenarios. Fig. 18a shows the utility cost consistently decreases from 56.57 to 42.28 as the number of users increase. It indicates users can enjoy higher check-in utility with same privacy criterion when more users join the service, which is probably because each user can find a more suitable anonymity group as the overall user population increases. On the other hand, we measure user's mobility characteristics as the number of check-in records and radius of gyration, which is each user's activity area measured by their mobility records' standard deviation from the central locations of them. We show the performance variation with these two characteristics in Fig. 18b and 18c, respectively. We can observe that users with only one check-in and 0 radius of gyration have higher utility cost, while the utility cost on other user groups does not vary significantly. It implies it is more difficult to protect users with only one check-in, but performance generally does not vary with the mobility characteristics.

Second, we also evaluate the impact of the amount of additional mobility data. Generally speaking, with more complete knowledge about user's mobility behavior, the system is able to better measure the privacy sensitive of each check-in record and derive better privacy solutions. The results of different percentages of additional mobility data are shown in Fig. 19. In Fig. 19a, we can observe that only 20 percent additional mobility data in WeChat dataset grants the system a 28.6 percent performance boost in check-in utility. In addition, when more than 60 percent additional mobility data is provided the performance of system gradually reaches a relative high point, with 30.9 percent utility improvement compares with no additional mobility data. Similar results are observed on Weibo dataset, which is shown in Fig. 19b. To conclude, the above evaluation verifies our intuition that moderate amount of additional mobility data can lead to significant check-in utility improvement, which showcases the feasibility of our system in real-world scenario.

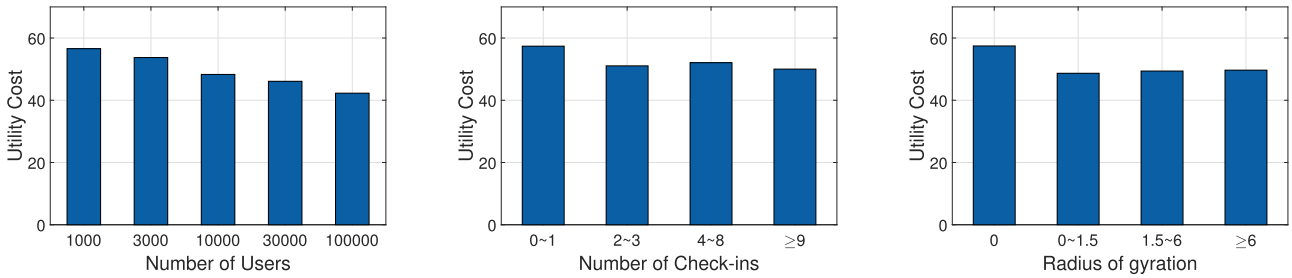
7.4 Discussion

Implications for Check-In Service Providers. The immediate implication of our study is that publilc check-ins suffer severe privacy exposure to linkage attacks. Adversary is able to learn additional knowledge about users beyond the check-ins they post by correlating the check-ins with other anonymous



(a) Temporal resolution (b) Spatial resolution (c) Utility cost

Fig. 16. The performance of our algorithm under different k and l on WeChat dataset.



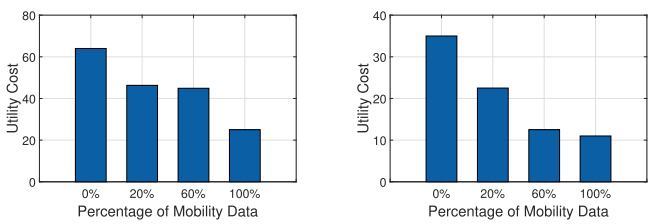
(a) Number of Users (b) Number of Check-ins (c) Radius of Gyration

Fig. 18. The model performance on datasets with different characteristics.

mobility datasets. Naive generalization techniques cannot effectively prevent such attacks while maintaining reasonable check-in utility. However, we also demonstrated that check-in service providers can more accurately measure the privacy sensitivity of check-ins by leveraging additional mobility records of users. Furthermore, we showcase the information of additional mobility data can be exploited to significantly improve check-in utility.

Implications for Individual Users. The most insightful implication for individual users is that there exists two trade-offs in privacy-preserving check-in service. First, individual may choose to improve the privacy level of their check-ins by posing check-ins with coarse-grained spatio-temporal resolutions. However, moderate privacy gain often requires significant check-in utility sacrifice. Second, users may effectively increase the utility of check-ins while enjoying the same privacy level by allowing check-in service providers to collect some additional mobility data, which can be limited on insensitive areas. Such findings may transform users' attitude toward check-in service providers' data collection, and form new kind of partnership between users and check-in service providers to defend linkage attacks.

Limitations. First, our analysis on check-in utility is still coarse-grained. For instance, check-in records in different regions, e.g., rural area and urban area, could have different contribution to the overall utility. Such fine-grained modelling requires in-depth user study, which we leave as our future work. Second, the currently proposed privacy framework is not specialized for longitude attacks. For example, a current secure check-in may be vulnerable in the future due to unpredictable human mobility. An intuitive method to address this problem is to design a sliding window algorithm to bound the privacy leakage in each window by a given budget.



(a) WeChat Dataset (b) Weibo Dataset

Fig. 19. The performance on dataset with different amount of additional mobility data.

8 Conclusion

In this paper, we investigate the problem of understanding and defending the linkage attacks on check-in services. Through extensive empirical analysis on two real-world datasets, we make important observations that the actual privacy sensitivity of check-ins is significantly smaller than one would expect from the uniqueness of check-ins, which can be better measured by introducing additional mobility data. Inspired by these findings, we design a novel *partition-and-group* framework that integrates the information of check-ins and additional mobility data to provide privacy-preserving and useful check-in service. Evaluation results show that the proposed framework significantly outperforms state-of-art baseline in terms of improving the check-in utility by 24∼57 percent and providing stronger privacy guarantee in the same time. We believe our study opens a new angle on measuring and preserving user privacy on check-in services.

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 U1936217, 61971267, 61972223, 61941117, 61861136003, Beijing Natural Science Foundation under L182038, Beijing National Research Center for Information Science and Technology under 20031887521, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] I. Bilogrevic, K. Huguenin, S. Mihaila, R. Shokri, and J. P. Hubaux, "Predicting users' motivations behind location check-ins and utility implications of privacy protection mechanisms," in *Proc. 22nd Netw. Distrib. Syst. Secur. Symp.*, Internet Society, 2015.

[2] A. Cecaj, M. Mamei, and N. Bicocchi, "Re-identification of anonymized cdr datasets using social network data," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, 2014, pp. 237–242.

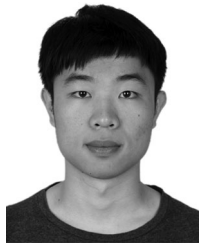
[3] R. Chen, B. Fung, B. C. Desai, and N. M. Sossou, "Differentially private transit data publication: A case study on the montreal transportation system," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 213–221.

[4] Y.-A. De Montjoye, L. Radaelli, V. K. Singh, et al., "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Sci.*, vol. 347, no. 6221, pp. 536–539, 2015.

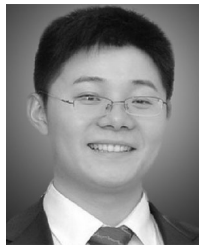
[5] S. Patil, G. Norcie, A. Kapadia, and A. J. Lee, "Reasons, rewards, regrets: Privacy considerations in location sharing as an interactive practice," in *Proc. 8th Symp. Usable Privacy Security*, 2012, pp. 1–15.

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

[7] 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.
[8] 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.
[9] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *Ucla L. Rev.*, vol. 57, 2009, Art. no. 1701.
[10] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata, Languages, Program.*, 2006, pp. 1–12.
[11] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 05, pp. 557–570, 2002.
[12] 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.
[13] Z. Tu, F. Xu, Y. Li, P. Zhang, and D. Jin, "A new privacy breach: User trajectory recovery from aggregated mobility data," *IEEE/ACM Trans. Netw.*, vol. 26, no. 3, pp. 1446–1459, Jun. 2018.
[14] B.-C. Chen, D. Kifer, K. LeFevre, A. Machanavajjhala, et al., "Privacy-preserving data publishing," *Foundations Trends® Databases*, vol. 2, no. 1–2, pp. 1–167, 2009.
[15] J. Su, A. Shukla, S. Goel, and A. Narayanan, "De-anonymizing web browsing data with social networks," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1261–1269.
[16] P. Welke, I. Andone, K. Blaszkiewicz, and A. Markowetz, "Differentiating smartphone users by app usage," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 519–523.
[17] Y. Shen, F. Wang, and H. Jin, "Defending against user identity linkage attack across multiple online social networks," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 375–376.
[18] C. Marlow, C. Marlow, and C. Marlow, "Find me if you can: Improving geographical prediction with social and spatial proximity," in *Proc. Int. Conf. World Wide Web*, 2010, pp. 61–70.
[19] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *ACM SIGKDD Explorations Newslett.*, vol. 18, no. 2, pp. 5–17, 2017.
[20] Z. Tu, K. Zhao, F. Xu, Y. Li, L. Su, and D. Jin, "Protecting trajectory from semantic attack considering k-anonymity, l-diversity and t-closeness," *IEEE Trans. Netw. Service Manage.*, vol. 16, no. 1, pp. 264–278, Mar. 2018.
[21] Z. Tu, K. Zhao, F. Xu, Y. Li, L. Su, and D. Jin, "Beyond k-anonymity: Protect your trajectory from semantic attack," in *Proc. 14th Annu. IEEE Int. Conf. Sens., Commun., Netw.*, 2017, pp. 1–9.
[22] F. Bonchi, L. V. Lakshmanan, and H. W. Wang, "Trajectory anonymity in publishing personal mobility data," *ACM SIGKDD Explorations Newslett.*, vol. 13, no. 1, pp. 30–42, 2011.
[23] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1241–1250.
[24] 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.
[25] C. Gao, C. Huang, Y. Yu, H. Wang, Y. Li, and D. Jin, "Privacy-preserving cross-domain location recommendation," in *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, 2019, Art. no. 11.
[26] M. E. Andrs, 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.
[27] K. Chatzikokolakis, C. Palamidessi, and M. Stronati, "A predictive differentially-private mechanism for mobility traces," in *Proc. Int. Symp. Privacy Enhancing Technol. Symp.*, 2013, vol. 8555, pp. 21–41.
[28] B. Bamba, L. Liu, P. Pesti, and T. Wang, "Supporting anonymous location queries in mobile environments with privacygrid," in *Proc. Int. Conf. World Wide Web*, 2008, pp. 237–246.
[29] T. Xu and Y. Cai, "Location anonymity in continuous location-based services," in *Proc. 15th Annu. ACM Int. Symp. Adv. Geographic Inf. Syst.*, 2007, Art. no. 39.
[30] C.-Y. Chow and M. F. Mokbel, "Trajectory privacy in location-based services and data publication," *ACM SIGKDD Explorations Newslett.*, vol. 13, no. 1, pp. 19–29, 2011.
[31] 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.
[32] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, "Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories," *Proc. IEEE Conf. Comput. Commun.*, 2017, pp. 1–9.
[33] A. Monreale, G. L. Andrienko, N. V. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel, "Movement data anonymity through generalization," *Trans. Data Privacy*, vol. 3, no. 2, pp. 91–121, 2010.
[34] M. E. Nergiz, M. Atzori, and Y. Saygin, "Towards trajectory anonymization: A generalization-based approach," in *Proc. SIGSPATIAL ACM GIS Int. Workshop Security Privacy GIS LBS*, 2008, pp. 52–61.
[35] V. Primault, S. B. Mokhtar, C. Lauradoux, and L. Brunie, "Time distortion anonymization for the publication of mobility data with high utility," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, 2015, vol. 1, pp. 539–546.
[36] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 3, 2013, Art. no. 1376.
[37] 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.
[38] L. Rossi, M. Musolesi, and A. Torsello, "On the k-anonymization of time-varying and multi-layer social graphs." in *Proc. 9th AAAI Int. Conf. Weblogs Social Media*, 2015, pp. 377–386.
[39] H. Wang, Y. Li, G. Wang, and D. Jin, "You are how you move: Linking multiple user identities from massive mobility traces," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 189–197.
[40] 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. 25th Annu. Netw. Distrib. Syst. Secur. Symp.*, 2018.
[41] D. Yang, D. Zhang, and B. Qu, "Privcheck: Privacy-preserving check-in data publishing for personalized location based services," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 545–556.
[42] Z. Huo, X. Meng, and R. Zhang, "Feel free to check-in: Privacy alert against hidden location inference attacks in GeoSNs," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2013, pp. 377–391.
[43] C. Dwork, "Differential privacy," *Encyclopedia Cryptography Security*, H. C. A. van Tilborg and S. Jajodia, Eds. Boston, MA: Springer US, pp. 338–340, 2011, doi: 10.1007/978-1-4419-5906-5_752.
[44] Z. Cheng, J. Caverlee, K. Lee, and D. Sui, "Exploring millions of footprints in location sharing services," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2783
[45] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, 2008, Art. no. 779.



Fengli Xu received the BS degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2015, and he is currently working toward the PhD degree from Electronic Engineering Department, Tsinghua University, Beijing, China. His research interests include human mobility, mobile big data mining and user behavior modeling.



Yong Li (M'09-SM'16) 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, Beijing, China. He has served as general chair, TPC chair, 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 6700. Among them, ten 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 ResearchersYoung Talent Program of China Association for Science and Technology and The National Youth Talent Support Program. He is a senior member of the IEEE.

Zhen Tu received the BS degree in electronics and information engineering and the second BS degree in economics both from Wuhan University, Wuhan, China, in 2016, and currently she is working towards the master degree from Electronic Engineering Department, Tsinghua University, Beijing, China. Her research interests include mobile big data mining, user behavior modeling, data privacy, and security.

Shuhao Chang received the BS degree in electronics information science and technology from Tsinghua University, Beijing, China, in 2019, and he is currently working toward the master's degree in Computer Science and Engineering Department, University of California San Diego, San Diego, California. His research interests include mobile big data mining, human mobility, and deep transfer learning.

Hongjia Huang received the BS degree from Electronic Engineering Department, Tsinghua University, Beijing, China, in 2019. Currently, he is working toward the master's degree from Electrical and Computer Engineering Department, University of California, Los Angeles, California. His research interests include mobile computing and ubiquitous computing.