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

Fengli Xu
Beijing National Research Center for Information Science and Technology Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
liyong07@tsinghua.edu.cn

Zhen Tu
Beijing National Research Center for Information Science and Technology Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
liyong07@tsinghua.edu.cn

Hongjia Huang
Beijing National Research Center for Information Science and Technology Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
liyong07@tsinghua.edu.cn

Shuhao Chang
Beijing National Research Center for Information Science and Technology Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
liyong07@tsinghua.edu.cn

Funing Sun
Tecent Inc., Beijing 100084, China
liyong07@tsinghua.edu.cn

Diansheng Guo
Tecent Inc., Beijing 100084, China
liyong07@tsinghua.edu.cn

Yong Li
Beijing National Research Center for Information Science and Technology Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
liyong07@tsinghua.edu.cn

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. To address this problem, 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% 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.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313506

CCS CONCEPTS

• **Security and privacy** → **Privacy protections**; *Social network security and privacy*; • **Networks** → *Social media networks.*

KEYWORDS

Check-ins, privacy-preserving data publishing, linkage attacks, mobility data privacy

ACM Reference Format:
Fengli Xu, Zhen Tu, Hongjia Huang, Shuhao Chang, Funing Sun, Diansheng Guo, and Yong Li. 2019. No More than What I Post: Preventing Linkage Attacks on Check-in Services. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308558.3313506

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 [4]. 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 [5], transportation records [8], and credit card records [12]. Moreover, recent researches show

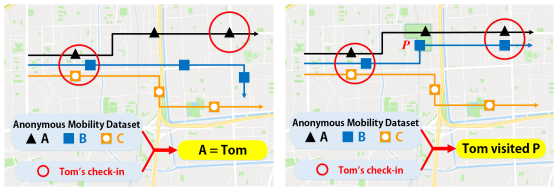
that most users are unaware or not able to fully anticipate the privacy risks embedded in posting check-ins [23]. Therefore, it is a paramount task for the check-in service providers to quantify the potential privacy exposures and put forward feasible solutions.

Previous efforts attempted to address the problem of linkage attacks on mobility data by ensuring user's anonymity in anonymous mobility datasets [12, 15]. 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 [35], and cannot prevent adversary from learning additional unreported mobility records [18]. 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 [13, 22]. 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.

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* [26] and *l-diversity* [18] into 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 two real-world datasets. The evaluation results demonstrate that our framework significantly outperforms state-of-art baseline method in terms of achieving 24%∼57% check-in utility improvement while providing stronger privacy guarantee in the same time. In addition, 32%∼62% 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

Linkage Attack: The linkage attacks were widely studied in multiple scenarios and had received increasing attention in recent years [17, 18, 26, 27]. The most prominent two branches are *re-identification attack* and *probabilistic attack* [7]. Specifically, the *re-identification attack* aims at recovering individuals' identities



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

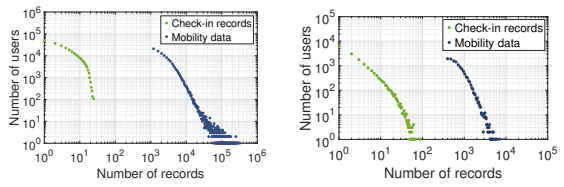
Figure 1: Illustration of linkage attacks on check-ins.

in anonymous datasets by achieving unique linkages with public datasets. For example, 87% of American population can be uniquely re-identified with the public accessible information of ZIP code, gender and date of birth [26]. Similar findings have been established in wide range of scenarios, including web browsing records [25], call detail records [35], app usage records [27, 32] 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 [26]. On the other hand, *probabilistic attack* is a more generic linkage attack, 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 [19]. In addition, the salary class of individuals can be accurately inferred by correlating census data with public available information [18]. To defend such attacks, *l*-diversity and *t*-closeness have been proposed to ensure the diversity on the sensitive information within each anonymity group [17, 18, 29, 30].

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 broken down into two categories: aggregated mobility data privacy and individual mobility data privacy. As for the former, recent study found evidence that aggregated mobility data suffered from the risks of leaking individual trajectories [28, 33]. In addition, *differential privacy* has been applied on aggregated mobility data to provide provable privacy guarantees for individuals [1, 14]. As for individual mobility privacy, *geo-indistinguishability* model is devised to achieve practical privacy guarantee in individual mobility collecting [2, 6]. In addition, vast amount of literature were dedicated to ensure location anonymity in the context of geo-referenced queries in location based service [3, 10, 34]. On the other hand, *cloaking*, *generalization* and *suppression* techniques are leveraged to achieve *k*-anonymity in releasing anonymous individual trajectories [15, 16, 20, 21, 24]. However, recent studies showed that such approaches will be likely to result in significant data utility degeneration [11, 35].

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 poses significant challenges on privacy solutions.



(a) WeChat dataset (b) Weibo dataset

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

3 PROBLEM FORMULATION

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 [26]. In check-in privacy scenario, 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, which is illustrated in Figure 1(a). Therefore, 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 from being uniquely re-identified.

l-diversity: It is put forward to ensure users' diversity on sensitive attributes within each anonymity group [18]. Specifically, it is designed to prevent *probabilistic attack*, where the adversary aims to learn additional mobility records the users do not report, which is illustrated in Figure 1(b). It requires each individual cannot be uniquely re-identified with the non-sensitive attributes (i.e., check-ins), while the sensitive attributes (i.e., unreported mobility records) should be of at least l different categories within each anonymity group.

$k^{\tau,l}$-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*.

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. The mobility records are collected when users invoke location based services in WeChat, such as posting check-ins and using map services.

Weibo Dataset[2]: This dataset is collected by a previous research [31]. 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.

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 Figure 2. From the results, we can observe that they all follow a fat-tailed distribution, which echos with the findings in previous researches [9]. It indicates that our datasets are representative of typical check-in behaviors.

5 SOLUTION

5.1 Definitions

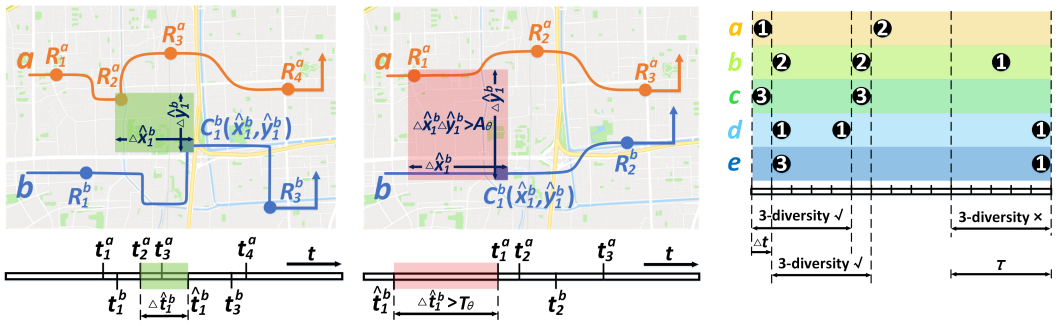
Formally, we define the additional mobility data of user i as $R^i = \left\{r_m^i\right\}$, where r_m^i is the m-th record of user i. It can be expressed as a tuple $r_m^i = (x_m^i,\ y_m^i,\ t_m^i)$, with x_m^i, $y_m^i$and t_m^i denoting the longitude, latitude and time stamp, respectively. On the other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

5.2 Basic Operations

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. In addition, we also define a cost function to measure the check-in utility loss in attaining privacy criterion. The basic operations are described as follows:

Generalization: It 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

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



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

Figure 3: The illustrations of three basic data sanitizing operations.

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 Figure 3(a).

Suppression: When the spatial and temporal resolution of check-in records is 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 Figure 3(b). Without loss of generality, A_θ and T_θ are set to $1000km^2$ 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 Figure 3(c). 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 then the operation returns $false$ for failing the diversity check. Otherwise, it returns $true$ for passing.

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

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

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. 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.

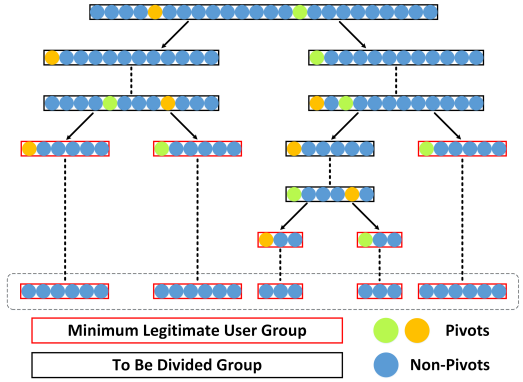


Figure 4: Illustration of *partition-and-group* framework.

5.3 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 to 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 Figure 4. 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 5.1. *If an anonymity group does not pass the* diversity check, *then any subsets of this anonymity group will not pass the* diversity check.

Proof. Suppose the unreported mobility records of an anonymity group do not pass the *diversity check* of parameters (τ, l). Based

Algorithm 1: 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 $\hat{\mathbb{C}}$
foreach $i,\ j\ \in\ \mathbb{C},\ j \neq i$ **do**
  $\mathbb{C}^\star \leftarrow k^{\tau,\ l}$-merge$(\mathbb{C}[\{i,j\}], \mathbb{R}[\{i,j\}], 2, 0, 0)$;
  $W[a,b] \leftarrow sum([U(c^\star)|\ \forall\ c^\star \in \mathbb{C}^\star])$;
checkin_stack.insert$(\mathbb{C})$; *mobility_stack.insert*$(\mathbb{R})$;
stop $\leftarrow$ *false*;
while *stop* $\neq$ *false* **do**
  $\mathbb{C}^\star \leftarrow$ *checkin_stack.pop*();
  $\mathbb{R}^\star \leftarrow$ *mobility_stack.pop*();
  if *! divide-2-group*$(\mathbb{C}^\star, \mathbb{R}^\star, W, k, l, \tau)$ **then**
    checkin_group.insert$(\mathbb{C}^\star)$;
    mobility_group.insert$(\mathbb{R}^\star)$;
  else
    $\mathbb{C}_1, \mathbb{C}_2, \mathbb{R}_1, \mathbb{R}_2 \leftarrow$ divide-2-group$(\mathbb{C}^\star, \mathbb{R}^\star, W, k, l, \tau)$;
    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**
  $\mathbb{C}^\star \leftarrow$ *checkin_stack.pop*();
  $\mathbb{R}^\star \leftarrow$ *mobility_stack.pop*();
  $\mathbb{C}^\star \leftarrow k^{\tau,\ l}$-merge$(\mathbb{C}^\star, \mathbb{R}^\star, k, l, \tau)$;
  $\hat{\mathbb{C}}$*.insert*$(\mathbb{C}^\star)$;
Return $\hat{\mathbb{C}}$;

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*. □

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* algorithm with the pseudocode presented in Algorithm 1. The elementary building block is $k^{\tau,\ l}$*-merge* function, which first performs *diversity check* on the given user group and then enumerates through all the users within the group to find optimal mobility records from each user for the check-in to *generalize* with. The *partition-and-group* algorithm 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. The *divide-2-group* algorithm equally divides the user group into 2 anonymity groups based on their distance to two pivot users that have maximum overall distance to other users. If both groups fail the *diversity check*, the input anonymity group is deemed unable to be further divided. On the other hand, the failed group keeps borrowing one most distant user from the succeed group, until they both pass or fail the *diversity check*. Finally, we apply $k^{\tau,\ l}$*-merge* function on each final anonymity group to ensure all users are protected by $k^{\tau,\ l}$*-anonymity*.

6 EVALUATION

6.1 Performance Comparison

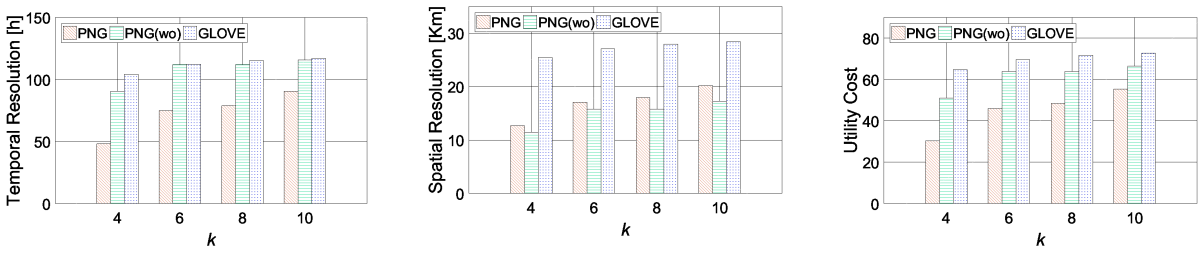
Our solution, denoted by *PNG*, aims to achieve $k^{\tau,\ l}$*-anonymity* to prevent both *re-identification attack* and *probabilistic attack*. In order to show its superiority, we consider two baselines, *i.e.*, *PNG(wo)* and *GLOVE*. *PNG(wo)* is a degraded version of *PNG*, in the condition that only *k-anonymity* is guaranteed to defend *re-identification attack*. On the other hand, *GLOVE* [15] is a state-of-art solution to achieve same privacy guarantee as *PNG(wo)*. 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 our two datasets for one-month duration.

We show the performance comparison of these three solutions with different values of k and l(=k/2) in Figure 5 and Figure 6. 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 resolutions of sanitized check-ins are 23h and 48h, while the spatial resolutions are 11km and 12km, respectively. Such spatial and temporal resolution is sufficient to accomodate user's need in documenting their daily life. However, the average spatial and temporal resolutions for *PNG(wo)* are 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% improvements in the comparison with *PNG(wo)* and *GLOVE* on WeChat dataset. In addition, *PNG* has 27% and 57% improvements in the comparison with *PNG(wo)* and *GLOVE* on Weibo dataset. In summary, all these results have demonstrated that our proposed *PNG* solution can significantly reduce check-in utility loss even when a stricter privacy criterion $k^{\tau,\ l}$*-anonymity* is met.

6.2 Impact of System Parameters

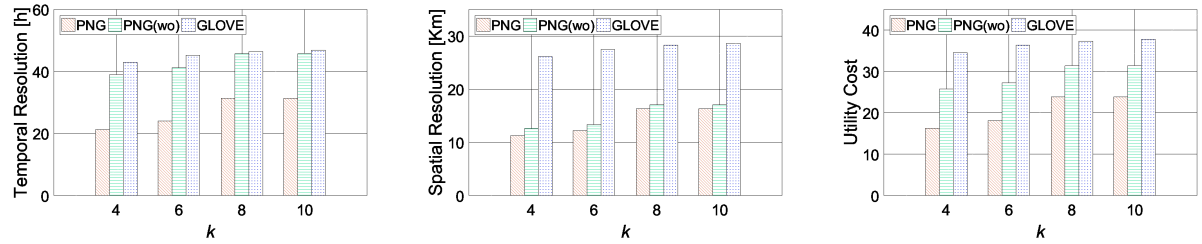
Now we analyze the impact of three key system parameters, *i.e.*, k, l and the amount of available additional mobility records, on the performance of our *PNG* solution.

First, based on both WeChat and Weibo datasets, we measure the performance of *PNG* with different settings of k and l, and show the results in Figure 7 and Figure 8. Take WeChat dataset for example, with a fixed 2-*diversity*, the average temporal resolution, spatial 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



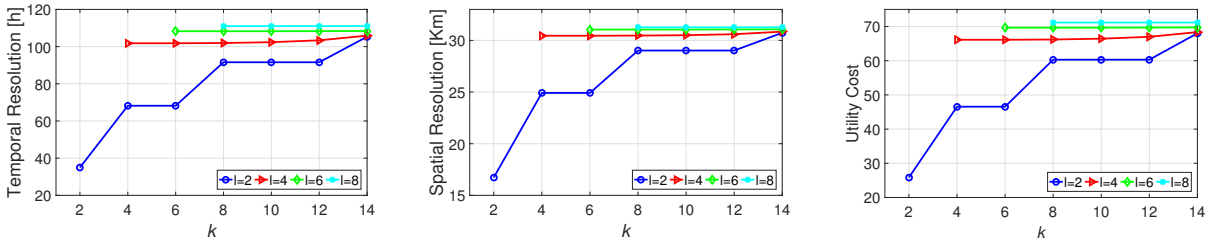
(a) Temporal resolution (b) Spatial resolution (c) Utility cost

Figure 5: The performance comparison between our solution and baseline on WeChat data.



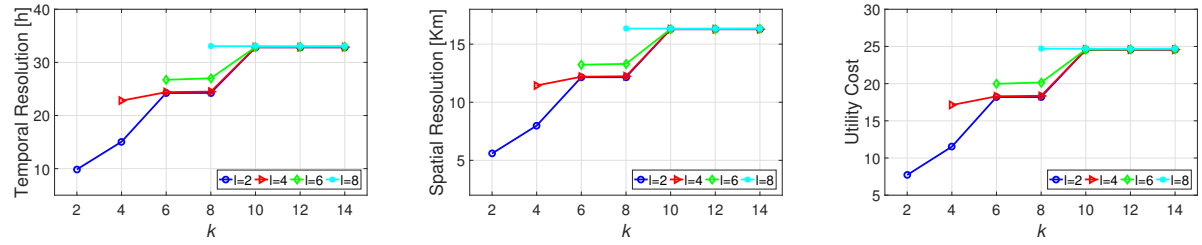
(a) Temporal resolution (b) Spatial resolution (c) Utility cost

Figure 6: The performance comparison between our solution and baseline on Weibo data.



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

Figure 7: The performance of our algorithm under different k and l on WeChat dataset.

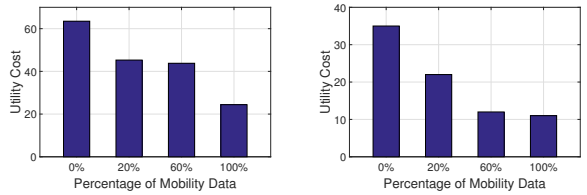


(a) Temporal resolution (b) Spatial resolution (c) Utility cost

Figure 8: The performance of our solution under different k and l on Weibo dataset.

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 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 Figure 9. In Figure 9(a), we can observe



(a) WeChat Dataset (b) Weibo Dataset

Figure 9: Impact of the amount of additional mobility data.

that only 20% additional mobility data in WeChat dataset grants the system a 28.6% performance boost in check-in utility. In addition, when more than 60% additional mobility data is provided the performance of system gradually reaches a relative high point, with 30.9% utility improvement compares with no additional mobility data. Similar results are observed on Weibo dataset, which is shown in Figure 9(b). 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 CONCLUSION

In this paper, we investigate the problem of understanding and defending the linkage attacks on check-in services. 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% 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.

ACKNOWLEDGMENT

This work was supported in part by The National Key Research and Development Program of China under grant 2017YFE0112300, the National Nature Science Foundation of China under 61861136003, 61621091 and 61673237, and Beijing National Research Center for Information Science and Technology under 20031887521.

REFERENCES

[1] Gergely Acs and Claude Castelluccia. 2014. A case study: privacy preserving release of spatio-temporal density in paris. (2014), 1679–1688.
[2] Miguel E. Andrs, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability:differential privacy for location-based systems. 901–914.
[3] Bhuvan Bamba, Ling Liu, Peter Pesti, and Ting Wang. 2008. Supporting Anonymous Location Queries in Mobile Environments with PrivacyGrid. In *International Conference on World Wide Web*. 237–246.
[4] Igor Bilogrevic, KÃĪvin Huguenin, Stefan Mihaila, Reza Shokri, and Jean Pierre Hubaux. 2015. Predicting Users' Motivations behind Location Check-Ins and Utility Implications of Privacy Protection Mechanisms. *Network and Distributed System Security Symposium* (2015).
[5] Alket Cecaj, Marco Mamei, and Nicola Bicocchi. 2014. Re-identification of anonymized CDR datasets using social network data. In *IEEE International Conference on Pervasive Computing and Communications Workshops*. 237–242.
[6] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. 2013. A Predictive Differentially-Private Mechanism for Mobility Traces. 8555 (2013), 21–41.
[7] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre, Ashwin Machanavajjhala, et al. 2009. Privacy-preserving data publishing. *Foundations and Trends® in Databases* 2, 1–2 (2009), 1–167.
[8] Rui Chen, Benjamin Fung, Bipin C Desai, and Nériah M Sossou. 2012. Differentially private transit data publication: a case study on the montreal transportation system. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 213–221.
[9] Z. Cheng. 2011. Exploring millions of footprints in location sharing services. *Icwsm* (2011).
[10] Chi-Yin Chow and Mohamed F Mokbel. 2011. Trajectory privacy in location-based services and data publication. *ACM Sigkdd Explorations Newsletter* 13, 1 (2011), 19–29.
[11] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports* 3 (2013), 1376.
[12] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, et al. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347, 6221 (2015), 536–539.
[13] Cynthia Dwork. 2006. Differential Privacy. In *International Colloquium on Automata, Languages, and Programming*. 1–12.
[14] Chen Gao, Chao Huang, Yue Yu, Huandong Wang, Yong Li, and Depeng Jin. 2019. Privacy-preserving Cross-domain Location Recommendation. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*.
[15] Marco Gramaglia and Marco Fiore. 2015. Hiding mobile traffic fingerprints with GLOVE. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*. ACM, 26.
[16] Marco Gramaglia, Marco Fiore, Alberto Tarable, and Albert Banchs. 2017. Preserving Mobile Subscriber Privacy in Open Datasets of Spatiotemporal Trajectories. *IEEE INFOCOM* (2017).
[17] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 106–115.
[18] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3.
[19] Cameron Marlow, Cameron Marlow, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *International Conference on World Wide Web*. 61–70.
[20] Anna Monreale, Gennady L Andrienko, Natalia V Andrienko, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, and Stefan Wrobel. 2010. Movement Data Anonymity through Generalization. *Trans. Data Privacy* 3, 2 (2010), 91–121.
[21] Mehmet Ercan Nergiz, Maurizio Atzori, and Yucel Saygin. 2008. Towards trajectory anonymization: a generalization-based approach. In *Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS*. ACM, 52–61.
[22] Paul Ohm. 2009. Broken promises of privacy: Responding to the surprising failure of anonymization. *Ucla L. Rev.* 57 (2009), 1701.
[23] Sameer Patil, Greg Norcie, Apu Kapadia, and Adam J. Lee. 2012. Reasons, rewards, regrets:privacy considerations in location sharing as an interactive practice. 1–15.
[24] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. 2015. Time distortion anonymization for the publication of mobility data with high utility. In *Trustcom/BigDataSE/ISPA, 2015 IEEE*, Vol. 1. IEEE, 539–546.
[25] Jessica Su, Ansh Shukla, Sharad Goel, and Arvind Narayanan. 2017. De-anonymizing Web Browsing Data with Social Networks. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1261–1269.
[26] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.
[27] Zhen Tu, Runtong Li, Yong Li, Gang Wang, Di Wu, Pan Hui, Li Su, and Depeng Jin. 2018. Your Apps Give You Away: Distinguishing Mobile Users by Their App Usage Fingerprints. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 138.
[28] Zhen Tu, Fengli Xu, Yong Li, Pengyu Zhang, and Depeng Jin. 2018. A New Privacy Breach: User Trajectory Recovery From Aggregated Mobility Data. *IEEE/ACM Transactions on Networking* 26, 3 (2018), 1446–1459.
[29] Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, and Depeng Jin. 2017. Beyond k-anonymity: protect your trajectory from semantic attack. In *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.
[30] Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, and Depeng Jin. 2018. Protecting Trajectory from Semantic Attack Considering k-Anonymity, l-diversity and t-closeness. *IEEE Transactions on Network and Service Management* (2018).
[31] Huandong Wang, Chen Gao, Yong Li, Gang Wang, Depeng Jin, and Jingbo Sun. 2018. De-anonymization of Mobility Trajectories: Dissecting the Gaps between Theory and Practice. In *Network and Distributed System Security Symposium*.
[32] Pascal Welke, Ionut Andone, Konrad Blaszkiewicz, and Alexander Markowetz. 2016. Differentiating smartphone users by app usage. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 519–523.
[33] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. 2017. Trajectory Recovery From Ash: User Privacy Is NOT Preserved in Aggregated

Mobility Data. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1241–1250.

[34] Toby Xu and Ying Cai. 2007. Location anonymity in continuous location-based services. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*. ACM, 39.

[35] Hui Zang and Jean Bolot. 2011. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 145–156.