# You Can Hide, But Your Periodic Schedule Can't

Minghua Ma†, Kai Zhao†, Kaixin Sui‡, Lei Xu†, Yong Li†, Dan Pei† *
†Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University ‡Microsoft
{mmh16, k-zhao16}@mails.tsinghua.edu.cn, kasui@microsoft.com,
xul0815@mail.tsinghua.edu.cn, {liyong07, peidan}@tsinghua.edu.cn

***Abstract*—The enterprise Wi-Fi networks enable the collection of large-scale users' trajectory datasets, which are highly desired for both research and commercial purposes. Meanwhile, releasing these mobility data also raises serious privacy concerns. A large body of work tries to achieve $k$-anonymity as the first step to solve the privacy problem and it has been qualitatively recognized that $k$-anonymity is still risky when the diversity of sensitive information in the $k$-anonymity set is low. However, there lacks a study that provides a quantitative understanding for trajectory data. In this work, we investigate the schedule-leakage risk for the first time, by presenting a large-scale measurement based analysis of the high schedule-leakage risk over sixteen weeks of trajectory data collected from Tsinghua University, a campus with 2,670 access points deployed in 111 buildings. Using this dataset, we recognize the high risk of the schedule-leakage, *i.e.*, even when 4-anonymity is satisfied, 28% of individuals' schedules are totally disclosed, and 56% are partly disclosed.**

## I. Introduction

With the booming development of mobile technologies, Wi-Fi has become a very popular one and is widely supported by mobile devices such as mobile phones, tablets, and portable game consoles. With ubiquitous carried-on mobile devices of people, the enterprise Wi-Fi networks are able to track people at an indoor level, which is more fine-grained than GPS [1, 2] or cellular base station [3, 4, 5] localization. Such mobility datasets are full of value and can be used in many fields such as location based social networking [1, 6], proximity marketing [7], mobility modeling [4, 8], and intelligent transportation [2, 9].

However, the trajectory of human mobility is very sensitive. For example, [3] studied the top two most visited locations of a person which are likely to correspond to home and work locations. Users' privacy could be seriously breached if the trajectory data is not properly sanitized before being published. A prior study [4] shows that even if users' identifiers are anonymized, we can still re-identify 95% of users in a trajectory dataset using four random spatiotemporal points.

Motivated by the above privacy risk, there is a large body of work [3, 5, 10, 11, 12, 13, 14] that aims to publish privacy-preserving trajectory datasets. Among them, $k$-anonymity [15] is widely used to sanitize the trajectory dataset in order to prevent the aforementioned re-identification attack. In particular, it guarantees that in any anonymous set, each individual is indistinguishable from at least $k-1$ other individuals. However, it has been **qualitatively** recognized that $k$-anonymity is not enough for preventing sensitive attribute disclosure, when faced with the *attribute linkage attack* [16]. Consider the case that $k$ users including the target individual are in the same anonymous set. If they have the same or very similar trajectories, it is easy for the adversary to determine the target individual's sensitive attribute (*e.g.*, the top two location [17]) without the need of uniquely identifying the target from the anonymous set. Besides, $l$-diversity [16] is another concept on demand that in a $k$-anonymous set, the diversity of sensitive attribute should be large enough so that it is hard for the adversary to guess the sensitive attribute.

Achieving $l$-diversity needs to define sensitive attribute of trajectories, but it is not easy as there is no uniform definition. Since the periodic schedule of a person corresponds to his working or entertaining pattens that is likely to attach to his habits, interests, activities and social relationships, in this paper we choose periodic spatiotemporal schedule ($periodic\ schedule$ for short) as the sensitive attribute, which has not been studied in the context of privacy protection previously to the best of our knowledge. We design a novel attack model with regard to periodic schedule and provide a large-scale measurement based analysis concerning that, even if the trajectory dataset satisfies $k$-anonymity, the probability that the adversary successfully infers the target's schedule may still be high. We take a first step by studying campus periodic schedule, collecting sixteen weeks (one semester) of Wi-Fi trajectory data which contains 10,126 distinct students' trajectories from Tsinghua University. In campus, periodic schedule is a specific type and relatively easy to study and there are some ground truths of 4,412 course timetables of 721 volunteers from the TUNow app [18, 19] to validate the periodic schedule extracted from the trajectory data. This dataset (§II-A) offers us a valuable opportunity to analyze the risk of leaking periodic schedule information ($schedule\text{-}leakage\ risk$ for short) of large-scale users' trajectory. Our contributions are summarized as follows:

- We present the first quantitative study of the schedule-leakage risk of the trajectory dataset, to the best of our knowledge.
- Our analysis highlights a very high schedule-leakage risk in the trajectory dataset (§IV). We find that even

* Dan Pei is the corresponding author.

for the data satisfying 4-anonymity, 28% of individuals' schedules are totally disclosed, and 56% individuals' schedules are partly disclosed.

The remainder of the paper is organized as follows. §II describes the dataset we collected and provides the background of trajectory privacy we would like to study. §III introduce a novel attack model related to periodic schedule. §IV analyzes schedule-leakage risk. §V discuss the trade off between data utilities and privacy and the possible solutions. §VI reviews the related work, and §VII concludes the paper.

## II. Background

In this section, we first introduce the Wi-Fi dataset we collected from a large campus network. Then, we introduce some background about the privacy issue regarding the trajectory.

### A. Wi-Fi Trajectory Dataset

From February 22, 2016 to June 12, 2016, 16 weeks of Wi-Fi trajectory dataset are collected from Tsinghua University. The time span of this dataset is a full semester excluding the final examination weeks. The dataset keeps 10,126 distinct trajectories of undergraduate and graduate students' devices. The campus covers an area of 4km$^2$ with about 42,000 students and 11,000 faculty and staff members. There are 2,670 Cisco enterprise APs being deployed in 111 buildings, including classrooms, departments, administrative buildings, libraries, dormitories, and others.

We use the method in [19], using those association and probing logs, to obtain the trajectory information from the Tsinghua campus Wi-Fi network. In particular, the fields of the dataset contains Device ID, Time of record and Location (Associated/probed AP). We distinguish Device ID using the device MAC addresses (anonymized by hashing). As for the device location, because a device, at a time, could be seen by more than one AP either through association or probing, we need a way to represent the device location. We represent the device location using the name of the AP which receives the strongest signal from the device and only use the granularity of **building level**, because (1) building level is a type of location generalization which makes the trajectory data less sensitive than the fine grained granularity (smaller than floor). (2) the building level is precise enough to find out the students' periodic schedule which we mentioned in §III-A. Therefore, rather than *geographic locations* (*e.g.*, GPS), our trajectory data is based on *semantic locations* [20]. The time and location of a record form a spatiotemporal point, and all the spatiotemporal points belonging to a device form the trajectory of the device.

### B. Ethical Considerations

Up to now, Tsinghua University does not have IRB (Institutional Review Board), thus we discuss the ethical considerations following the best common practices. First, we get the approval to collect data from Tsinghua Network Center and the volunteers of TUNET App. Second, both the student ID and MAC address in our dataset are anonymized by randomized mapping before any analysis. Third, our conclusions are all based on statistical results rather than specific individual. In summary, we conclude that during our research we strictly protect the privacy of students and would not cause any ethical problem.

### C. Privacy Attack on Trajectory Data

To preserve the privacy of $data\ owners$ in the data publishing, the $data\ publisher$ will remove explicit identifiers, such as name and telephone numbers to achieve $Privacy\ Preserving\ Data\ Publishing$ (PPDP) [21]. However, simple anonymization cannot guarantee PPDP. There are two types of privacy attacks related to our study [21]: (1) $Record\ linkage$: the adversary aims to uniquely identify the victim in dataset with additional outside information. (2) $Attribute\ linkage$: the adversary gets the victim's sensitive attribute instead of identifying him.

In the type of "trajectory-only" datasets, previous works concentrate on $record\ linkage$ [5, 10, 12]. However, solving $record\ linkage$ does not ensure complete security of privacy, for example, the adversary can implement an $attribute\ linkage$ attack, which we will have a specific example to explain in §III. The attribute linkage attack assumes that, an adversary knows that the victim individual $i$ is in the dataset. For example, a student of Tsinghua University (who always switches on Wi-Fi of his phone) should be in our dataset. Then, the adversary wants to find out *sensitive attributes*, such as his curriculum schedule, of $i$ from the trajectories without identifying him.

## III. Problem and Positioning

People always move by a periodic schedule so that the periodic schedule is very sensitive. However, attack model regarding to the periodic schedule disclosure in trajectory dataset still remain unknown to the best of our knowledge. In this section, we want to answer the following questions: (a) What is periodic schedule disclosure? This will be described with the ground truth compare in §III-A. (b) How is the periodic schedule disclosed when data publishing, even the dataset is satisfied $k$-anonymity? We propose an attack model (§III-B) regarding to this sensitive attribute.

### A. Periodic Schedule

Periodic schedule is a common phenomenon that is sensitive to the individual. The periodic schedule of an individual is formally defined as the individual's pauses at a certain location with the regular time interval. In campus, periodic schedule is relatively easy to define. For each student, they have curriculum schedules with the regular time interval of a week during a semester. We treat the periodic schedule as the studied sensitive attribute, considering for the following reasons. First, the periodic schedule contains information about spatial and temporal dimension, while the top two locations [17], *i.e.* home and working place just in a spatial dimension. The adversary can use the periodic schedule of an individual to gain more knowledge about when and where to attack. Let alone

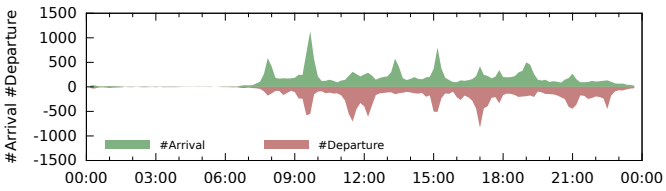



Fig. 1. Flows in teaching buildings.

that periodic schedule probably includes top locations. Second, it is easy for the adversary to extract the schedule from the trajectory data. There are many existing researches about how to extract periodic schedule from trajectory datasets. [22] proposed an algorithm $Periodica$ to discover periodic behavior in trajectories. Third, the trajectory dataset we employed in this paper is collected from campus, where most individuals are undergraduate students and graduate students. They all have relatively specific periodic schedules. Fortunately, we crowdsource some ground truths of course timetables of volunteers from the TUNow app [18, 19]. The 721 volunteers of TUNow registered to 4,412 courses in total. These ground truths can be used to validate the schedule we inferred. In the following analysis, we are based on the spring semester in 2016.

We discuss how to generate the periodic schedule from the published trajectory dataset. The attackers are assumed to just gain the knowledge from the published trajectory dataset. To start with aggregated temporal mobility metrics of a building: the arrival/departure numbers of devices, these metrics are summarized based on the association and probing logs. Students attend classes according to the curriculum schedule of a day in classroom buildings. The curriculum schedule, which is shown in Table I, is easy to find out in Fig. 1. People's movements always follow certain regulations which can be extracted from the trajectories. Thus, our method can be applied on other datasets in which individuals have periodic schedules. The individual's trajectory duration is divided into $T = 30$ time slots on a weekday schedule. The records of every individual can be mapped to a time slot using the time of the records. After mapping all 16 weeks' data, for each time slot, we count the number of each individual $(i)$'s appearances at each location. The location may not always a distinct one, *e.g.*, $i$ attends class at $location\ a$ for 15 weeks, while there is a week which is holiday. So he is in $location\ b$ for one week. Then we select the most frequent location, *i.e.* $location\ a$ in the above example, to form the periodic schedule of $i$.

We compare the schedules extracted by our method to the ground truth, *i.e.*, the 721 volunteers' curriculum schedules, to illustrate that the trajectory dataset can be used to find the periodic schedule. Specifically, split by comma, the first number in a cell of Table I represents the total number of schedules extracted by our method, while the second number in a cell is the total number of schedules of the ground truth. We can tell that Tsinghua University does not arrange any course in Thursday afternoon and the total number of courses $3,548$ extracted by us is close to the ground truth $4,412$ (takes up more than 80%). The periodic schedule extracted from

TABLE I
CURRICULUM SCHEDULE IN TSINGHUA

| | Mon. | Tue. | Wed. | Thu. | Fri. |
|---|---|---|---|---|---|
| 08:00 ∼ 09:35 | 132, 147 | 148, 161 | 232, 247 | 124, 136 | 48, 54 |
| 09:50 ∼ 12:15 | 163, 183 | 269, 306 | 340, 384 | 218, 261 | 201, 243 |
| 13:30 ∼ 15:05 | 46, 136 | 106, 201 | 95, 194 | 0, 0 | 38, 133 |
| 15:20 ∼ 16:55 | 190, 209 | 252, 278 | 165, 183 | 0, 1 | 96, 126 |
| 17:05 ∼ 18:40 | 5, 13 | 24, 46 | 55, 89 | 0, 2 | 3, 6 |
| 19:20 ∼ 21:45 | 81, 96 | 223, 243 | 111, 123 | 155, 176 | 28, 35 |

trajectory dataset is not 100% the same as the curriculum for these reasons: on the one hand, some classroom buildings are without AP so the student's cell phone may associate to AP in the building nearby; on the other hand, students may not take cell phones to classes or may turn off Wi-Fi they are having classes.

### B. Attack Model

We formally describe the problem as illustrated in Fig. 2. In order to attack a target individual $i$, the adversary must have some information which is helpful to identify the individual $i$ in the dataset, and we call it *quasi-identifier*, denoted by $QID$. The adversary's knowledge about the individual $i$ is represented as $QID_i$. Indeed, it is typically easy for adversary to know some locations that $i$ has been to. Here we assume $QID_i$ is formed by random $M$ spatiotemporal points of $i$. For example, the adversary can get the spatiotemporal information from $i$'s posts on social networks, or that the adversary met $i$ several times. The adversary searches the dataset with random $M$ spatiotemporal points and get a set of trajectories, denoted as $S_i$, in which the trajectories have the same $QID$ with $i$. Our attack model is based on $k$-anonymity. Trajectories which have several same spatiotemporal points are extracted as $k$-anonymous set. This is a part-length trajectory anonymization method [17]. As the dataset already achieves $k$-anonymity, the size of $S_i$ is larger than one, which indicates that he cannot reidentify the individual. However, the adversary is still able to infer the sensitive attribute of $i$ from these trajectories.

As for the sensitive attribute, we introduce the notion of *periodic schedule* of an individual. If the periodic schedule of an individual is breached, the adversary may gain more chance to know exactly the individual's habit, interests, activities and social relationships.

We achieve $k$-anonymity by just select individuals with same $QID$ for the following reason. There are several existing algorithms [5, 10, 12], but we do not utilize these algorithms. Because these algorithms achieve full-length trajectory anonymization, which means trajectories in the same $k$-anonymous set are exactly the same. Thus, their periodic schedule are also the same, which leads to serious schedule-leakage. As shown in Fig. 3, 56% of individuals have very different trajectories from others, and the corresponding anonymous sets are of size 1. As a result, these individuals can be uniquely identified. In this paper, we mainly focus on anonymous sets which contain at least two individuals' trajectories.

In an anonymous dataset, when $k$-anonymity is well satisfied, the adversary cannot uniquely identify the individual

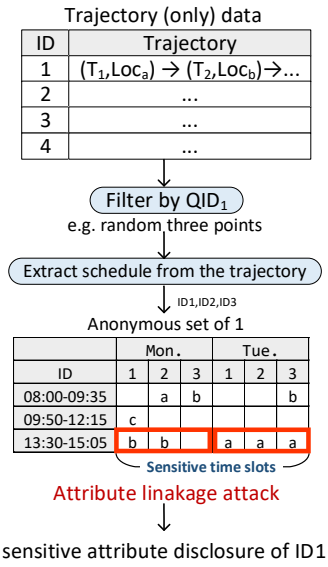



Fig. 2. Sensitive attribute disclosure of ID1 on the trajectory-only data.

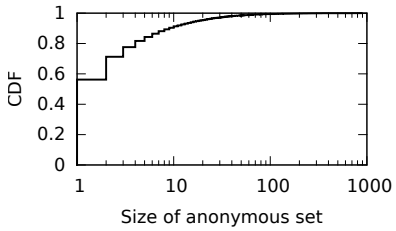



Fig. 3. CDF of size of anonymous set

TABLE II
EXAMPLE TO CALCULATE PROB(ABILITY)

| Time slot \ ID | 1 | 2 | 3 | Prob |
|---|---|---|---|---|
| Mon. 08:00 - 09:35 | $a:8$ | $a:14$ | $b:12$ | 0.46 |
| Mon. 09:50 - 12:15 | $c:14$ | $c:5$ | $a:8$ | 0.40 |
| Mon. 13:30 - 15:05 | $b:14$ | $b:12$ | - : 0 | 0.54 |
| Tue. 08:00 - 09:35 | $c:5$ | $c:2$ | $b:16$ | 0.33 |
| Tue. 09:50 - 12:15 | $c:8$ | - : 0 | $b:6$ | 0.17 |
| Tue. 13:30 - 15:05 | $a:14$ | $a:12$ | $a:15$ | 0.85 |

that he want to find out. He may find multiple individuals due to the $k$-anonymity. As the example given in Fig. 2, the adversary knows individual ID1's several spatiotemporal points and ID1 is in the dataset. With this information, he finds three individuals ID1, ID2, ID3 which means they all have these spatiotemporal points in their trajectories. Apparently, their periodic schedules are not identical but similar. For example, during 13:30 to 15:05 on Tuesday, they all appear in $Location\ a$ , hence the adversary can make a conclusion that location $a$ must be a periodic location in ID1's periodic schedule. This privacy leakage is caused by the low diversity of their periodic schedule.

In summary, in order to preserve more privacy, trajectories in a dataset should be similar in terms of quasi-identifiers (realizing $k$-anonymity) and diverse in terms of sensitive attributes (against the attribute linkage attack).

## IV. MEASUREMENT STUDY OF SCHEDULE-LEAKAGE RISK

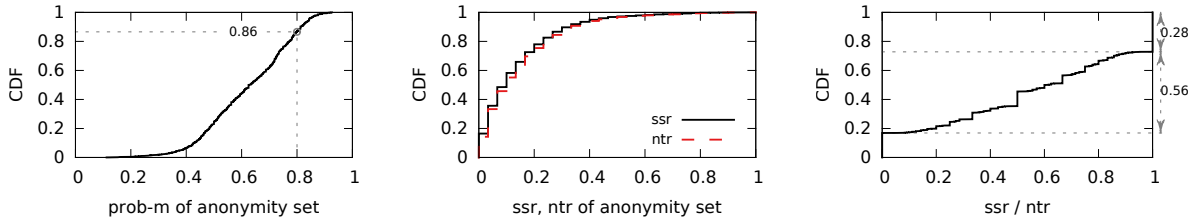
In this section, we measure the schedule-leakage risk of trajectory dataset. We first illustrate that how $k$-anonymity is achieved and propose two metrics to measure the schedule-leakage risk. Then we analyze the risk of the dataset.

### A. Methodology

As aforementioned in §III-B, in this paper, we choose *random $M$ spatiotemporal points* of an individual as the quasi-identifier. Then, we want to explore how serious the privacy problem is. The first step is to obtain the anonymous set of each individual. For each individual $i$, we search the dataset with his quasi-identifier, namely the random $M$ spatiotemporal points to find the individual(s) that have the same quasi-identifier with $i$. By this way, we get $i$'s anonymous set $S_i$. We set $k = 4$ for the following experiments. Thus the size of $S_i$ is $k$ ($k > 3$). 4-anonymity is a relatively strong protection for trajectory privacy. In previous studies, full-length trajectory anonymization [5] only applicable to $k$ no more than 5, otherwise the anonymized dataset becomes hardly exploitable.

Based on the novel sensitive attribute, *i.e.*, *periodic schedule*, the records of every individual are mapped to the $T = 30$ time slots. The data length is denoted as $C$, here $C$ is 16 (weeks). To quantify schedule-leakage risk in $S_i$, we propose two metrics, since this problem has not been studied before. One is *Max Attack Probability* ($prob$-$m$), represents the successfully attack probability of the most vulnerable time slot in a $k$-anonymous set. The other one is *Sensitive Schedule Rate* ($ssr$), which is the proportion of time slots of $i$'s schedule that can be easily guessed.

**Max Attack Probability** ($prob$-$m$): For individuals in $S_i$, their periodic schedules can be merged as a $T \cdot k$ matrix $L$. Each row of $L$ is a time slot of the schedule, and each column of $L$ is an individual in $S_i$. We present an example in Table II. According to §III-A, the elements of $L$ are $location$ and $appear\ times$, *e.g.*, in the first cell of the above schedule, $L_{n,j} = (a:8)$ means that at time slot $n$, individual $j$ appeared 8 times at location $a$ during the studied 16 weeks. Then for each row of $L$, we classify the elements according to the location and sum up the $appear\ times$ in the same category. The location with largest $appear\ times\ A_n$ is assumed to be the attack location. For example, in the first row of Table II, we can get two categories, *i.e.*, location $a$ and location $b$, which contribute to the $appear\ times$ of 22 $(8+14)$ and 12, separately, so $A_n = 22$. Since there are $k$ individuals in the $k$-anonymous set and each individual has $C$ weeks, $A_n/(k{\cdot}C)$ denotes that the adversary has a probability of $A_n/(k \cdot C)$ to successfully acquire the victim's periodic location in this time slot. Then $prob\text{-}m = \max_{n=1}^{T}\{A_n/(k \cdot C)\}$ represents the attack probability of the most vulnerable time slot in the schedule. In the above example of schedule table, it can be found that the $prob$-$m$ is 0.85. This means the adversary has the largest attack probability of 0.85 to say that he can find the victim in location $a$ when Tuesday from 13:30 to 15:05.




(a) CDF of $prob\text{-}m$ (b) CDF of $ssr$ (c) CDF of $ssr$ / $ntr$

Fig. 4. Attack metrics of anonymous set (quasi-identifier: random three spatiotemporal points, sensitive attribute: periodic schedule)

**Sensitive Schedule Rate** ($ssr$) is the proportion of time slots in an individual $i$'s periodic schedule that is sensitive and easy to guess. Suppose that individual $i$ is the target that the adversary wants to attack. Before calculating $ssr$, we preprocess $L$ to a matrix by retaining the periodic locations at which $appear\ times/C$ is greater than $60\%$. Since there are several Chinese holidays and Tsinghua Anniversary on which there is no lecture during a semester and students may miss lectures sometimes, we believe $60\%$ is a reasonable threshold. As $C = 16$, we keep the location in the time slot if its appear times is more than 9 (weeks). For those locations with appear times less than 9 (weeks) we assume that they are not periodic locations throughout the semester. For each time slot within the $k$-anonymous set, if the location of individual $i$ appears the most times, we say this time slot is sensitive for $i$, *i.e.* the sensitive time slot. The intuition is that it's natural to guess the most frequent location as $i$'s periodic location. Then we count the number of sensitive time slots which represents how many time slots of $i$ can be easily attacked. Namely, for time slots of $T$, $ssr = \#(sensitive\ time\ slots)/T$.

An example is given in Fig. 2. The original time table is Table II. Here, $ssr = 0.33$. From the processed table the adversary can infer that individual ID1 appears at location $a$ during 13:30-15:05 every Monday and at location $b$ during 13:30-15:05 every Tuesday.

### B. Schedule-Leakage Risk of the Dataset

Now, we analyze the risk of leaking periodic schedule from a $k$-anonymous dataset. Here we define $M$ = 3 which means an adversary knows exactly three spatiotemporal points of victim's trajectory. After searching the dataset, the adversary acquires a $k$-anonymous set containing several individuals. If $k = 1$, the adversary can identify the victim and obtain all the trajectory information, let alone sensitive attributes. However, if $k > 1$, he cannot know who is the victim, but he still can extract some privacy information from their similar periodic schedule.

First, Fig. 4(a) shows the attack probability of the most likely to be successfully attacked time slot in a $k$-anonymous set. There are $14\%(1 - 0.86)$ individuals whose $prob\text{-}m$ is larger than $0.8$. The larger $prob\text{-}m$ is, the more chance of his periodic schedule could be breached. And we calculate the average $prob\text{-}m$ of all anonymous sets is 0.65. This means the schedule-leakage risk is very high, because the diversity of the location in the time slots of anonymous set is low.

Second, we analyze the $ssr$ using the preprocessed $L$ (explained in the definition of $ssr$). Non-empty time slot means in that time slot the individual has a certain location where he appears more than 9 weeks. For example, in Fig. 2, 8:00-9:35 Monday is a non-empty time slot for individual 2 as he is at the location $a$. Since attacking the empty time slots is meaningless to the adversary, the number of non-empty time slots is considered. The non-empty time slots rate is defined as $ntr = \#(non\text{-}empty\ time\ slots)/T$.

In Fig. 4(b), there are two curves representing the $ssr$ and the rate of non-empty time slots in periodic schedule of the victim, respectively. We can see that the CDF of $ssr$ is very close to the CDF of $ntr$, which means most of the non-empty time slots are very sensitive. To be more precise, we calculate the rate of successfully attacked time slot numbers to non-empty time slot numbers which equals to $ssr/ntr$. The result is shown in Fig. 4(c). From this figure we can tell that more than half of time slots are easy to guess even when $4$-anonymity is satisfied. 28% of individuals' schedules are totally disclosed, and 56% individuals' schedules are partly disclosed, which raises a serious privacy concern when publishing trajectory data.

## V. Discussion

The trajectory datasets incontestable open up many issues with respect to security and privacy. In this section, we discuss the issues of trajectory dataset publishing.

### A. Trade off between data utilities and privacy

Wi-Fi provider can collect register's location data in real time or retrospectively to physically locate the phone with high degrees of accuracy. GPS enabled phones enable precise outdoor location placement. Such trajectory datasets are full of value and can be used in many fields. There are many public-available trajectory datasets [1, 2, 20].Using these data may compromise the privacy of the data owners, *i.e.* the mobile users. To protect privacy, data needs to be modified or encrypted before being published. However, this usually results in a decline of data utility. Existing solutions [5] already satisfies data utility only to meet the needs of $k$-anonymity. If the protection should be strong enough to against the above attack, data utility will be further reduced. It is often a trade-off between the data utility and the privacy [3, 5]. This would be a big challenge in the future work.

### B. Possible Solutions

As the analysis above infers, the schedule-leakage risk is mainly due to the low diversity of schedules. From this perspective, $l$-diversity may be tackled to solve the schedule-leakage problem. One possible solution is that making trajectories of individuals who have different schedules similar and making those who have similar schedules different. However, this is difficult because, on one hand, similar trajectories usually have similar schedules. On the other hand, achieving $l$-diversity may cause huge damage to data utility.

## VI. Related work

In the area of $Privacy\ Preserving\ Data\ Publishing$, a widely used privacy criterion is $k$-anonymity [15]. Within the domain of movement meta-data *a.k.a.* trajectory datasets, many algorithms [5, 10, 12] have been proposed based on $k$-anonymity. The $k$-anonymity criterion is effective in preventing individuals from being re-identified, but it fails to provide strong protection against attribute linkage attacks. This shortcoming motivates $l$-diversity [16], which demands that in a $k$-anonymous set, the diversity of sensitive attribute should be large enough so that the adversary can hardly guess the sensitive attribute. There are studies that try to protect dataset from attribute linkage such as [23], but it assumes that the trajectories are not sensitive and the sensitive attribute is something else that links to the trajectory *e.g.*, diagnosis. However, trajectory-only data is different from traditional data such as medical data and census data which has separate quasi-identifiers and sensitive attributes. Quasi-identifiers and sensitive attributes in trajectory data are both extracted from the trajectory. Inspired by this, [17] analyzes the diversity of top two locations as quasi-identifiers [3] in a $k$-anonymous set and finds that increasing $k$ cannot improve the diversity significantly. Differential privacy [24] is one of the state of the art techniques for privacy protection. It mainly solve the problems of database queries such as publishing spatiotemporal density of each cell [14] rather than publishing trajectories, which is the situation of this paper.

## VII. Conclusion

In this paper, we present what we believe to be the first **quantitative** study of the schedule-leakage risk of the trajectory dataset. Our study highlights a very high risk of the schedule-leakage. For example, we find that even for 4-anonymity, 28% of individuals' schedules are totally disclosed, and 56% individuals' schedules are partly disclosed. We plan to extend to periodic schedule in other environments, not just the campus, but the attack model and the metric of schedule-leakage risk is general applicable.

## VIII. Acknowledgments

Thorough comments and valuable feedbacks from the reviewers also helped us improve the work. This work was partly supported by the Key Program of NSFC under grant 61233007, the NSFC under grant 61472214 & 61472210, the 973 Program under grant 2013CB329105, the Tsinghua National Laboratory for Information Science and Technology key projects, the Global Talent Recruitment (Youth) Program. and the Cross-disciplinary Collaborative Teams Program for Science & Technology & Innovation of Chinese Academy of Sciences-Network and system technologies for security monitoring and information interaction in smart grid.

## References

[1] Y. Zheng, Y. Chen, X. Xie, and W.-Y. Ma, "Geolife2. 0: a location-based social networking service," in *Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on*. IEEE, 2009, pp. 357–358.

[2] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: driving directions based on taxi trajectories," in *Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems*. ACM, 2010, pp. 99–108.

[3] H. Zang and J. Bolot, "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, 2011, pp. 145–156.

[4] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, 2013.

[5] M. Gramaglia and M. Fiore, "Hiding mobile traffic fingerprints with glove," *ACM CoNEXT*, pp. 1–13, 2015.

[6] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 1082–1090.

[7] "Proximity marketing," https://www.marketingtechblog.com/proximity-marketing/.

[8] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[9] L. Chen, A. Mislove, and C. Wilson, "Peeking beneath the hood of uber," in *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*. ACM, 2015, pp. 495–508.

[10] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. Ieee, 2008, pp. 376–385.

[11] M. E. Nergiz, M. Atzori, and Y. Saygin, "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, 2008, pp. 52–61.

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

[13] F. Bonchi, L. V. Lakshmanan, and H. W. Wang, "Trajectory anonymity in publishing personal mobility data," *ACM Sigkdd Explorations Newsletter*, vol. 13, no. 1, pp. 30–42, 2011.

[14] G. Acs and C. Castelluccia, "A case study: privacy preserving release of spatio-temporal density in paris," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1679–1688.

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

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *TKDD*, vol. 1, no. 1, p. 3, 2007.

[17] K. Sui, Y. Zhao, D. Liu, M. Ma, L. Xu, L. Zimu, and D. Pei, "Your trajectory privacy can be breached even if you walk in groups."

[18] M. Zhou, M. Ma, Y. Zhang, K. Sui, D. Pei, and T. Moscibroda, "Edum: classroom education measurements via large-scale wifi networks," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 316–327.

[19] M. Zhou, K. Sui, M. Ma, Y. Zhao, D. Pei, and T. Moscibroda, "Mobicamp: a campus-wide testbed for studying mobile physical activities."

[20] K. Tan, G. Yan, J. Yeo, and D. Kotz, "Privacy analysis of user association logs in a large-scale wireless lan," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 31–35.

[21] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys (CSUR)*, vol. 42, no. 4, p. 14, 2010.

[22] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye, "Mining periodic behaviors for moving objects," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1099–1108.

[23] B. D. M. Mohammed, Noman Fung, "Walking in the crowd: anonymizing trajectory data for pattern analysis," in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 1441–1444.

[24] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.