Your Trajectory Privacy Can Be Breached Even If You Walk in Groups

Kaixin Sui, Youjian Zhao, Dapeng Liu, Minghua Ma, Lei Xu, Li Zimu, Dan Pei*
Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University

***Abstract*—The enterprise Wi-Fi networks enable the collection of large-scale users' mobility information at an indoor level. The collected trajectory data is very valuable for both research and commercial purposes, but the use of the trajectory data also raises serious privacy concerns. A large body of work tries to achieve k-anonymity (hiding each user in an anonymity set no smaller than k) as the first step to solve the privacy problem. Yet it has been qualitatively recognized that k-anonymity is still risky when the diversity of the sensitive information in the k-anonymity set is low. There, however, still lacks a study that provides a quantitative understanding of that risk in the trajectory dataset.**

In this work, we present a large-scale measurement based analysis of the low-diversity risk over four weeks of trajectory data collected from Tsinghua, a campus that covers an area of 4 km^2, on which 2,670 access points are deployed in 111 buildings. Using this dataset, we highlight the high risk of the low diversity. For example, we find that even when 5-anonymity is satisfied, the sensitive attributes of 25% of individuals can be easily guessed. We also find that although a larger k increases the size of anonymity sets, the corresponding improvement on the diversity of anonymity sets is very limited (decayed exponentially). These results suggest that diversity-oriented solutions are necessary.

I. Introduction

Nowadays, mobile Internet plays an important role in people's daily life. People are used to carrying their mobile devices everyday, and heavily depend on them for work, entertainment, shopping, social events, *etc.* Among different wireless communication techniques, Wi-Fi has become a very popular one and is widely supported by mobile devices, such as mobile phones, tablets, and portable game consoles. In recent years, universities, companies, cities, and commercial providers often deploy enterprise Wi-Fi networks for large local areas (*e.g.*, campus, buildings, airports, hotels, public transportations, and shopping centers) to offer the mobile Internet access [2, 3]. With ubiquitous carried-on mobile devices of people, the enterprise Wi-Fi networks have a good opportunity to track people at an *indoor level*, which is more fine-grained than GPS based [24, 26] or cellular basestation based [12, 14, 25] measurements. Such trajectory datasets are full of value and can be used in many fields, *e.g.*, location based social networking [10, 26], proximity marketing [4], mobility modeling [12, 19], and intelligent transportation [9, 24].

Meanwhile, however, the trajectory related information is very sensitive. For example, the top two most visited locations of a person are likely to correspond to home and work locations [25]. Users' privacy could be seriously breached if the trajectory data is not properly sanitized before being published to the public or used internally. For example, a prior study [12] shows that because of the high uniqueness of users' mobility patterns, although users' identifiers are anonymized, one can still pinpoint 95% of users in a trajectory dataset (called the *re-identification attack*) using five random spatiotemporal points.

Motivated by the above privacy risk, there is a large body of work [5–8, 14, 25] that aims to preserve privacy when publishing trajectory datasets. Among those methods, k-anonymity [21] is widely used to sanitize the trajectory dataset in order to prevent the re-identification attack. In particular, it guarantees that each individual is indistinguishable from at least $k-1$ others (hidden in an *anonymity set* no smaller than k). However, it has been **qualitatively** recognized that k-anonymity is not enough for preventing sensitive attribute disclosure [18], especially in front of the *probabilistic inference attack*. Consider the example that 10 users including the target are in the same anonymity set, but they have the same or very similar trajectory, then it is easy for the adversary to know the target's sensitive attribute (*e.g.*, top two locations) without the need of distinguishing him or her from others in the anonymity set.

The above potential risk calls for a high diversity of sensitive attributes in anonymity sets, in addition to k-anonymity, to prevent sensitive attribute disclosure. Yet, this direction has not been explored very much for the *trajectory* dataset, although there are some solutions designed for the traditional relational data, *e.g.*, l-diversity [18]. Given the fact that many existing sanitization solutions for the trajectory dataset are still k-anonymity based, it is valuable to know whether and how k-anonymity helps improve the diversity, or how much risk still remains when k-anonymity is already achieved. The answer to the above question has important implications regarding the potential privacy risk of the trajectory dataset, and also provides the motivation to design more effective diversity-oriented sanitization solutions for the trajectory dataset. To the best of our knowledge, however, there still lacks a study to give a quantitative answer to the above question.

In order to bridge this gap in **quantitatively** understanding the low-diversity risk of the trajectory data in the wild, we provide a large-scale measurement based analysis. To this end, we collect four weeks of Wi-Fi trajectory data from Tsinghua University, a campus that covers an area of 4 km^2, on which about 42,000 students and 11,000 faculty and staff members are living. There are 2,670 Cisco enterprise APs (Access Points) being deployed in 111 buildings of the campus. This dataset (§II-A) offers us a valuable opportunity to analyze the diversity of large-scale users' trajectory at an *indoor level*.

*Dan Pei is the corresponding author.

Our contributions are summarized as follows:

- In this paper, we present what we believe to be the first quantitative study of the low-diversity risk of the trajectory dataset at an indoor level.
- Our analysis highlights a very high risk of the low diversity in the trajectory dataset (§III). For example, we find that even for the data satisfying 5-anonymity (*i.e.*, all individuals are hidden in a group of no less than 5), the sensitive attributes of 25% of individuals can be easily guessed. Worse still, the problem keeps getting more serious for smaller k.
- We also characterize the relationship between the diversity and k-anonymity, and find that increasing k helps improve the diversity, but its effectiveness decays exponentially as k goes up. In addition, achieving larger k ($\geq$ 5) can significantly destroy the utility of the data. Thus, we cannot expect to use k-anonymity with a large k to solve the problem, but need to design diversity-oriented solutions.

The rest 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 analyzes the diversity and its relationship with k-anonymity. §IV reviews the related work, and §V concludes the paper.

II. Background

In this section, we first describe the Wi-Fi trajectory dataset we collected from a large enterprise wireless local area network — Tsinghua campus Wi-Fi network. Then we introduce the background regarding the trajectory privacy issue we study in this paper.

A. Wi-Fi Trajectory Dataset

We collect a four-week Wi-Fi trajectory dataset from Tsinghua University. This dataset contains 154,354 distinct devices' trajectories. The Tsinghua campus covers an area of 4km^2, on which about 42,000 students and 11,000 faculty and staff members are living. There are 2,670 Cisco enterprise APs being deployed in 111 buildings, including classrooms, departments, administrative buildings, libraries, dormitories, and others. Those APs are controlled by 14 Cisco wireless controllers.

In order to obtain the trajectory information from the Tsinghua campus Wi-Fi network, we locate devices by polling the device *association* and *probing* logs from the APs via SNMP every five minutes[1]. Therefore, our dataset contains the position snapshot of both the devices that are associated to the APs, and the devices, not associated, but whose 802.11 probing can be sensed by the APs. This measurement enables us to discover more devices in the campus, regardless of the fact whether the devices are connecting to the APs or not. Moreover, in this way, we are able to measure a more complete trajectory of devices. For example, when a device occasionally uses rogue APs [20] such as personal hotspots, its broadcasted probing can still be tracked by the APs.

Using those association and probing logs, we generate the trajectory dataset of devices. In particular, Table I lists the fields of the dataset. We distinguish devices using the device mac addresses (anonymized by hashing). As for the device location, because one device could be seen by several APs either through association or probing at a time, we need a way to represent the device location. One possible solution is to use indoor localization technologies, such as triangulation [17]. However, obtaining the absolute position of those 2,670 APs would take a huge effort. Instead, we take advantage of the fact that each AP in Tsinghua is carefully named by operators using its semantic location "*building-floor-room-AP*". We represent the device location using the name of the AP which receives the strongest signal from the device. This method gives a reasonable estimation of the nearest AP to the device [23]. Therefore, rather than *geographic locations* (*e.g.*, GPS), our trajectory data is based on *semantic locations*, which is similar to [22]. 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.

TABLE I
SELECTED FIELDS OF THE DATASET.

Field	Value Example	
Device ID (anonymized MAC address)	118974	
Time of record	2015-11-20 18:05	
Location (Associated/probed AP receiving the strongest signal from the device)	Building Floor Room AP	Main building 5^{th} #2 #1
Building Type	Administrative building	

In summary, our dataset differs from others in several aspects. (1) Temporal sampling rate: our trajectory is derived from Wi-Fi data, which enables us to sample the trajectory at a consistent and high frequency (*i.e.*, every five minutes). In contrast, the sampling interval of another commonly-used data source for the trajectory, namely Call Data Records (CDR) [12, 14, 25], is varying and long (depending on whether users call or not). (2) Location type and granularity: our dataset can describe the semantic location at indoor level. For example, a device can be located at "Main building, 5th floor, room #2, close to AP #1". Neither GPS based [24, 26] or cellular basestation based [12, 14, 25] trajectory data can provide such accurate and meaningful indoor information. (3) Data scale of Wi-Fi based trajectory datasets: we collect data from over 150,000 devices in a campus where 2,670 APs are deployed; a past dataset [22] contains only 6,000 devices and 623 APs. (4) Data format: our data is the typical "trajectory-only" data, that is, the data only contains the trajectory of each individual. This kind of data is relatively easy to collect, and is also suitable for studying the privacy issue of the trajectory. Some datasets consist of the trajectory and other sensitive attributes, such as diseases [11], but they are outside the scope of this paper.

[1]The interval of 5 minutes is much more fine-grained than prior trajectory datasets [12, 25], whose interval could be hours. We do not sample the data at a higher frequency in order to avoid overloading the wireless controllers.

B. Privacy Attack on Trajectory Data

When publishing trajectory datasets, preserving user privacy is a critical task. Among several privacy risks [13], we study a typical one — *sensitive attribute disclosure*. In particular, an adversary knows that the target individual i is in the dataset. For example, a student of Tsinghua (who uses Wi-Fi of mobile devices) should be in our dataset. Then, the adversary wants to find out the *sensitive attributes* of i from its trajectory, such as home or work locations. Since our focus is on trajectory-only data, we only consider sensitive attributes regarding the trajectory.

We formally describe the problem as follows (Fig. 1). The adversary's knowledge about i forms the *quasi-identifiers* of i, denoted by QID_i. Similar to sensitive attributes, for the trajectory-only data, QID_i should be only related to i's trajectory, such as random three spatiotemporal points of i. Then, adversary can determine the individuals whose quasi-identifiers are the same with QID_i. These individuals form the *anonymity set* of i, denoted by S_i. The goal of the adversary is to infer the sensitive attributes of i, for example the top two locations of i, based on S_i.

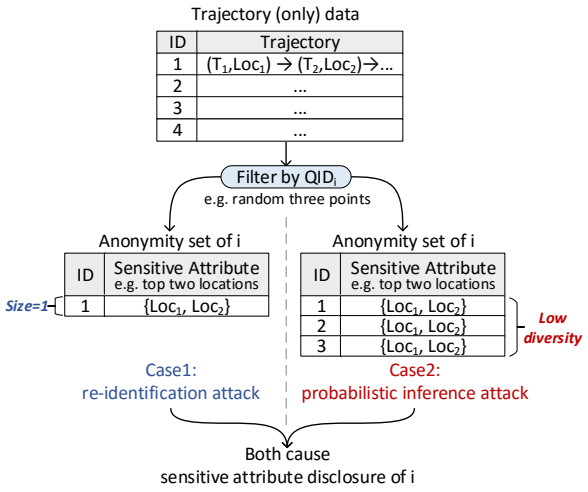


Fig. 1. Sensitive attribute disclosure of i on the trajectory-only data.

To succeed in finding the sensitive attributes, one special case is that $|S_i| = 1$, which is called the *re-identification attack* (case 1 in Fig. 1). It indicates that the adversary can identify i in the dataset. Prior studies show that this case could be very common because of users' highly unique mobility patterns. For example, [12] shows that using five random spatiotemporal points can pinpoint 95% of users uniquely. Because the re-identification attack definitely leads to sensitive attribute disclosure, many studies [5, 6, 14] try to prevent the re-identification attack as an initial step towards privacy-preserving trajectory data publishing. The key idea of those methods is to sanitize the dataset to realize k-anonymity, *i.e.*, $|S_i| \geq k, \forall i$ in the data. The larger k is, the more robust the data is against the re-identification attack.

However, it has been recognized that k-anonymity is not enough for preventing sensitive attribute disclosure [18], especially in front of the *probabilistic inference attack* (case 2 in Fig. 1). Consider an extreme example that $|S_i| = 3$, but all the three individuals in S_i have the same sensitive attributes, so it is easy to infer the sensitive attribute of i without the need of distinguishing i from others in S_i. This potential risk calls for a high *diversity of sensitive attributes* (diversity for short) of the anonymity sets to prevent sensitive attribute disclosure.

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

C. Motivation and Problem

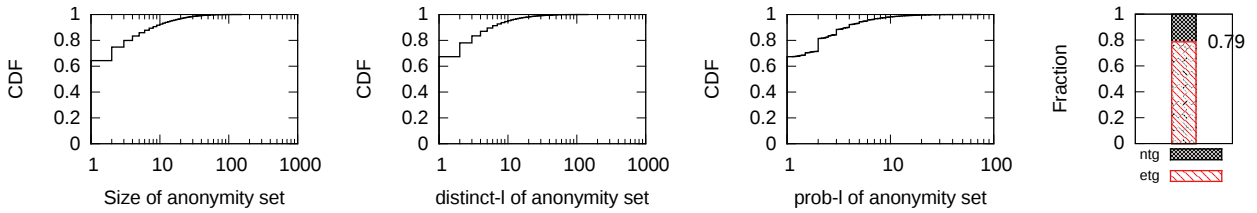
In the domain of traditional relational data, where the quasi-identifiers and the sensitive attributes are independent, there already exists data sanitization methods considering the diversity, such as l-diversity [18]. Realizing the goal is, however, more challenging on the trajectory-only data, because the quasi-identifiers and the sensitive attributes are both derived from the trajectory information, and can thus be inter-dependent. For example, the quasi-identifier is the random three spatiotemporal points, and the sensitive attribute is the top two locations. In this context, realizing the similarity on the quasi-identifiers, and the diversity on the sensitive attributes are conflicting, and thus more challenging to satisfy (as briefly discussed in [5]). Therefore, many existing data sanitization methods for trajectory datasets [5, 6, 14] only focus on k-anonymity without considering the diversity.

The good news is that, intuitively, k-anonymity might help improve the diversity as it limits the minimal size of anonymity sets, thus the anonymity sets could potentially be more diverse. But what is the case in the wild? Given the fact that many solutions are k-anonymity based, it is valuable to know whether and how the size of anonymity set helps improve the diversity, or how serious problems can still be caused by the low diversity when k-anonymity is already satisfied. For example, if the problem is trivial, then trajectory datasets are relatively safe after using existing k-anonymity based data sanitization solutions; otherwise, we should be aware of the risk of the low diversity, and such a result can also motivate the design of more effective diversity-oriented sanitization solutions for trajectory datasets.

To the best of our knowledge, however, there still lacks a measurement study to provide a **quantitative** understanding on the above questions. Our work aims at providing such knowledge by analyzing our large-scale Wi-Fi trajectory dataset. In particular, we would like to answer the following questions:

- What is the diversity of the trajectory dataset?
- What is the relationship between k and the diversity? Does a larger k help improve the diversity? If so, will the diversity problem be solved by k-anonymity?

Note that, we do know that the data utility is another major concern of the data. It is often a trade-off between the data utility and the privacy [14, 25]. However, in this work we focus on studying the diversity and its relationship with k-anonymity. We consider the data utility as our future work.



(a) CDF of size of anonymity set (b) CDF of $distinct\text{-}l$ (c) CDF of $prob\text{-}l$ (d) Fraction of etg/ntg

Fig. 2. Size and diversity of anonymity set (quasi-identifier: random three spatiotemporal points, sensitive attribute: top two locations)

III. Diversity Analysis of the Data

In this section, we analyze the dataset described in §II-A to answer the first two questions. We use different metrics to evaluate the diversity of the dataset, and analyze the relationship between the diversity and k.

A. Methodology

As aforementioned in §II-B, to measure the diversity of a dataset, first we need to determine the quasi-identifiers and the sensitive attributes. In this paper, we choose *random three spatiotemporal points* of an individual as the quasi-identifier, since such information is relatively easy for the adversary to acquire [8]. *E.g.*, the adversary can get the spatiotemporal information from i's posts on social networks, or that the adversary met i several times. As for the sensitive attribute, we choose *top two locations* of an individual. According to [25], top locations have significant implications of people's mobility patterns, thus causing serious concerns of privacy. *E.g.*, the top two locations are likely to correspond to home and work locations [25]. That is, the adversary knows i's random three spatiotemporal points, and wants to find out i's top two locations from the dataset.

Given the above quasi-identifier and sensitive attribute, we measure the diversity of each individual's anonymity set as follows. For each individual i, first we get its quasi-identifier, namely the random three spatiotemporal points. Then, we search the dataset to find the individual(s) that have the same value of the quasi-identifier with i. These individuals form i's anonymity set S_i. To quantify the diversity of the sensitive attribute in S_i, we adopt three metrics. The first two are commonly-used *numerical* metrics, and the third one is a *categorical* metric defined by us.

- $\boldsymbol{distinct\text{-}l}$: Distinct l-diversity [15]. There are $distinct\text{-}l$ distinct values for the top two locations in S_i.
- $\boldsymbol{prob\text{-}l}$: Probabilistic l-diversity [16]. The frequency of the most frequent value of the top two locations is $\frac{1}{prob\text{-}l}$ in S_i. $prob\text{-}l$ reaches the maximum, *i.e.*, $|S_i|$, when the top two locations in S_i are all different from each other.
- $\boldsymbol{etg}$ or $\boldsymbol{ntg}$: Easy to guess or not. We categorize S_i into two classes. If the top two location of i has the highest frequency in S_i, we say that S_i is etg (not diverse), otherwise ntg (diverse). The intuition is that with no assumption of the background knowledge of i, it is natural to guess the most frequent top two locations in S_i as i's top two locations. This simple guess will be correct if S_i is etg.

B. Diversity of the Dataset

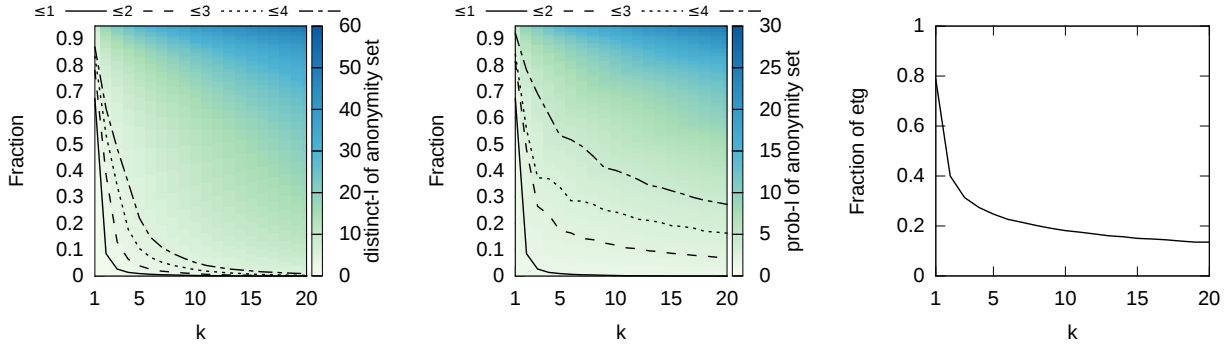
Since the diversity of an anonymity set is the lowest (regardless of metrics) when the size of the anonymity set is one, we first would like to know how many anonymity sets belong to this extreme case. In Fig. 2(a), we see that 65% of anonymity sets have size of one, which means that those individuals are highly unique in their mobile patterns and can be re-identified easily with only three of their random spatiotemporal points. This observation agrees with previous studies carried out on CDR datasets [12, 14, 25]. The nature of the uniqueness of the trajectory will no doubt significantly impact the diversity as well. In particular, those 65% of anonymity sets have the lowest diversity: their $distinct\text{-}l$ and $prob\text{-}l$ both equal one, and they belong to etg. The distribution of the three diversity metrics in Fig. 2(b), Fig. 2(c), and Fig. 2(d) confirms this: the fractions of $distinct\text{-}l = 1$, $prob\text{-}l = 1$, and etg are larger than 65%.

In addition, we find that the uniqueness does not explain all the low diversity. For example, Fig. 2(d) shows that 79% of anonymity sets belong to etg. That means, excluding those 65% caused by the high uniqueness, there are still 40% ($\frac{79\%-65\%}{100\%-65\%}$) of anonymity sets belonging to etg even though their sizes are no less than two! This observation shows that k-anonymity could have a high risk of the low diversity, and also motivates the following analysis of the relationship between the diversity and k.

C. Relationship Between the Diversity and k

After seeing the evidence of low-diversity risk in the k-anonymity dataset, it is natural to ask how serious the problem is? Does this risk disappear if k increases? We answer these questions by analyzing the relationship between the diversity and k. To this end, we focus on the anonymity sets whose sizes are no less than k ($k = 1, 2, ..., 20$, respectively). Then, we characterize how the diversity distribution changes over these anonymity sets. The results are shown in Fig. 3, and three main takeaways are as follows:

The diversity of anonymity sets are very low for a small k. We observe that when k is relatively small (*e.g.*, $k \leq 5$), the diversity of many anonymity sets is very low, thus posing a high risk for k-anonymity. In particular, Table II shows how many anonymity sets have the low diversity when $k = 5$ (corresponding to $k = 5$ in Fig. 3). We observe that, although the anonymity sets contain no less than five individuals, at most four distinct sensitive attributes can be found in 21% of



(a) Fraction of *distinct-l* (b) Fraction of *prob-l* (c) Fraction of *etg* (*ntg* is the rest)

Fig. 3. Distribution of the three diversity metrics as k increases. $k = 1$ also represents the original dataset, which itself is already 1-anonymity. In (a) and (b), the fractions of *distinct-l* and *prob-l* no more than 1, 2, 3, 4 (relative low diversity) are shown with lines.

anonymity sets (*distinct-l* ≤ 4), and at most three distinct sensitive attributes can be found in 10% of anonymity sets (*distinct-l* ≤ 3). For *prob-l*, the most frequent top two locations account for more than $1/4$ (*prob-l* ≤ 4) in 53% anonymity sets, and even more than $1/2$ (*prob-l* ≤ 2) in 17% of anonymity sets. Worse still, 25% anonymity sets belong to *etg*, meaning that the top two locations of 25% of individuals can be easily guessed from the most frequent top two locations in their anonymity sets. These results **quantitatively** demonstrate the high risk of the low diversity for k-anonymity.

TABLE II
HOW MANY ANONYMITY SETS HAVE THE LOW DIVERSITY WHEN $k = 5$.

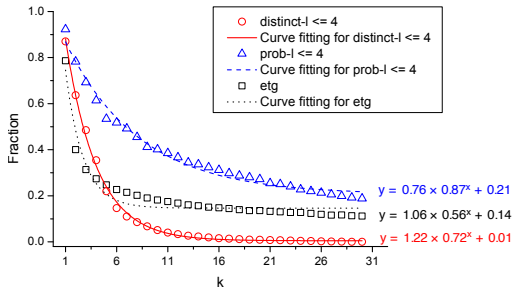
x	Fraction of		
	distinct-l $\leq x$	*prob-l* $\leq x$	*etg*
1	1%	1%	25%
2	4%	17%	
3	10%	33%	
4	21%	53%	

***k*-anonymity helps improve the diversity of anonymity sets, but less effective as *k* increases.** Overall, we see that in Fig. 3, as k increases, the diversity of anonymity sets becomes better. For example, the fractions of anonymity sets with *distinct-l* ≤ 4, *prob-l* ≤ 4, and belonging to *etg* are getting smaller. However, the curves visually present a exponential decay rather than an ideal linear relationship. As a result, the improvement on diversity brought by k-anonymity is very limited. Add to that the fact that achieving large k (≥ 5) greatly destroys the utility of the data [14], one cannot expect to use k-anonymity with large k to improve the diversity. Instead, some more effective diversity-oriented solutions should be specifically designed.

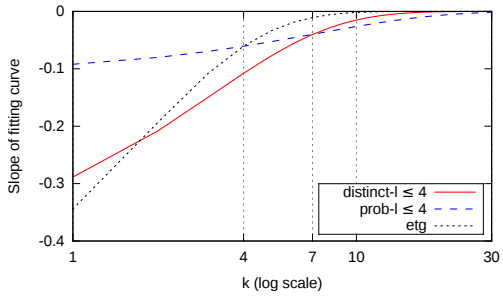
***k*-anonymity affects the three diversity metrics differently, and has the least impact on *prob-l*.** We notice that the relationships between k and different diversity metrics are different. Since the low diversity represents a high risk and thus is of more interest, without loss of generality, we focus on modeling the relationships between k and *distinct-l* ≤ 4, *prob-l* ≤ 4, and *etg* (curves in Fig. 3). Given the visually exponential trends, we use the asymptotic regression model [1] in the form $y = a \times b^x + c$ to characterize the relationships. We fit the range of k from 1 to 30. The fitting curves and corresponding functions are shown in Fig. 4(a). The residual sum of squares (RSS) is 0.003, 0.013, and 0.028 for *distinct-l* ≤ 4, *prob-l* ≤ 4, and *etg*, respectively. The slopes of the fitting curves are shown in Fig. 4(b). We observe that when k is less than 4, *prob-l* ≤ 4 drops much slower than *etg*; when k is less than 7, *prob-l* ≤ 4 drops slower than *distinct-l* ≤ 4. For larger values of k, their slopes are much similar with each other. In addition, the fitting curve of *prob-l* ≤ 4 also converges to a larger c (*i.e.*, 0.21) than the other two, which means in about 20% of anonymity sets, there are over $1/4$ individuals having the same top two locations (*prob-l* ≤ 4), although these anonymity sets are already larger than 30 ($k = 30$)! The above result shows that k-anonymity is more vulnerable to the attack based on *prob-l*.

IV. RELATED WORK

k-anonymity [21] commends that each individual is indistinguishable from at least $k-1$ others, *i.e.*, walking in a group of k. It is an effective countermeasure against the re-identification attack, however, it fails to provide strong privacy protection under probabilistic inference attacks, which motivate l-diversity [18] and t-closeness [15]. Our work deals with privacy preservation of trajectory dataset. As claimed in [14], this is a very different problem from ensuring anonymity in relational micro-data. Within the domain of movement micro-data, *a.k.a.* trajectory datasets, present widely used k-anonymity based algorithms [5, 6, 14] are unsatisfactory when the diversity of the sensitive information in the k-anonymity set is low. The risk of k-anonymity when confronted to attacks that aim at revealing the sensitive attribute has been qualitatively recognized in the context of moving object data [5]. Our work provides a quantitative understanding of that risk in the trajectory dataset. Inspired by l-diversity [18], we use the *distinct-l* [15], *prob-l* [16], and *etg* or *ntg* (defined by us) rather than the size of anonymity sets [25] to evaluate the privacy vulnerability of the dataset.



(a) Curve fitting



(b) Slope of fitting curves.

Fig. 4. Curve fitting for the relationship between k and the fraction of anonymity sets with the low diversity.

V. Conclusion

As people increasingly rely on mobile Internet and applications for their daily life, the trajectory data is now much easier to collect and more important than before. Mining such data can provide a huge research and commercial value, such as helping understand the mobility pattern of users, and providing useful insights for system design. However, for all of its advantages, publishing trajectory data is a double-edged sword in the sense that it raises serious privacy concerns. Many sanitization methods have been proposed to protect the privacy contained in trajectory data, and most of them are based on k-anonymity, which is qualitatively known for being vulnerable to the risk of the low diversity. However, there still lacks a study that provides a quantitative understanding of that risk in the trajectory data.

In this paper, we present what we believe to be the first **quantitative** study of the low-diversity risk of the trajectory dataset at the indoor level. We collect four weeks of trajectory data from the Wi-Fi network of Tsinghua University, a 4 km^2 campus with 2,670 APs deployed in over 100 buildings. This dataset gives us a valuable opportunity to analyze the diversity risk over a large-scale trajectory data at an indoor level. Our study highlights a very high risk of the low diversity. For example, we find that even for 5-anonymity, the sensitive attributes of 25% of individuals can be easily guessed. We believe that our study provides a good quantitative evidence of the low-diversity risk in the trajectory data. Given the fact that the trajectory data is important for both research and business, we argue that more sophisticated diversity-oriented sanitization solutions should be designed to further preserve the privacy of the trajectory data, and we will investigate this direction in our future work.

Acknowledgments

We thank the anonymous reviewers for their thorough comments. We strongly appreciate Jiaxin Ding, Jie Gao, Zhi Wang, and Yong Li for their valuable suggestions. We are grateful to Yousef Azzabi for his elaborative proofreading.

This work was partly supported by the Key Program of NSFC under grant 61233007, the NSFC under grant 61472214 & 61472210, the 863 Program under grant 2013AA013302, 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] Asymptotic regression model. http://www.originlab.com/doc/Origin-Help/Asymptotic1-FitFunc.
[2] Ewlan market. http://www.delloro.com/news/enterprise-wireless-lan-market-to-expand-70-percent-by-2018.
[3] Ewlan overview. http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Mobility/emob41dg/emob41dg-wrapper/ch1_Over.html.
[4] Proximity marketing. https://www.marketingtechblog.com/proximity-marketing/.
[5] O. Abul, F. Bonchi, and M. Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *ICDE*. IEEE, 2008.
[6] O. Abul, F. Bonchi, and M. Nanni. Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 2010.
[7] G. Acs and C. Castelluccia. A case study: privacy preserving release of spatio-temporal density in paris. In *SIGKDD*. ACM, 2014.
[8] F. Bonchi, L. V. Lakshmanan, and H. W. Wang. Trajectory anonymity in publishing personal mobility data. *ACM Sigkdd Explorations Newsletter*, 2011.
[9] L. Chen, A. Mislove, and C. Wilson. Peeking beneath the hood of uber. In *IMC*. ACM, 2015.
[10] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *SIGKDD*. ACM, 2011.
[11] C.-Y. Chow and M. F. Mokbel. Trajectory privacy in location-based services and data publication. *ACM Sigkdd Explorations Newsletter*, 2011.
[12] 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*, 2013.
[13] B. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)*, 2010.
[14] M. Gramaglia and M. Fiore. Hiding mobile traffic fingerprints with glove. *ACM CoNEXT*, 2015.
[15] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*. IEEE, 2007.
[16] T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In *SIGKDD*. ACM, 2009.
[17] H. Liu, H. Darabi, P. Banerjee, and J. Liu. Survey of wireless indoor positioning techniques and systems. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 2007.
[18] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *TKDD*, 2007.
[19] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 2010.
[20] K. Sui, Y. Zhao, D. Pei, and L. Zimu. How bad are the rogues' impact on enterprise 802.11 network performance? In *INFOCOM*. IEEE, 2015.
[21] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.
[22] K. Tan, G. Yan, J. Yeo, and D. Kotz. Privacy analysis of user association logs in a large-scale wireless lan. In *INFOCOM*. IEEE, 2011.
[23] J. Teng, C. Xu, W. Jia, and D. Xuan. D-scan: Enabling fast and smooth handoffs in ap-dense 802.11 wireless networks. In *INFOCOM*. IEEE, 2009.
[24] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: driving directions based on taxi trajectories. In *SIGSPATIAL*. ACM, 2010.
[25] H. Zang and J. Bolot. Anonymization of location data does not work: A large-scale measurement study. In *MobiCom*. ACM, 2011.
[26] Y. Zheng, Y. Chen, X. Xie, and W.-Y. Ma. Geolife2. 0: a location-based social networking service. In *MDM*. IEEE, 2009.