Fighting Statistical Re-Identification in Human Trajectory Publication

Jiaxin Ding
Stony Brook University
jiading@cs.stonybrook.edu

Chien-Chun Ni
Stony Brook University
chni@cs.stonybrook.edu

Jie Gao
Stony Brook University
jgao@cs.stonybrook.edu

ABSTRACT

The maturing of mobile devices and systems provides an unprecedented opportunity to collect a large amount of real world human motion data at all scales. While the rich knowledge contained in these data sets is valuable in many fields, various types of personally sensitive information can be easily learned from such trajectory data. The ones that are of most concerns are frequent locations, frequent co-locations and trajectory re-identification through spatio-temporal data points. In this work we analyze privacy protection and data utility when trajectory IDs are randomly mixed during co-location events for data collection or publication. We demonstrate through both analyses and simulations that the global geometric shape of each individual trajectory is sufficiently altered such that re-identification via frequent locations, co-location pairs or spatial temporal data points is not possible with high probability. Meanwhile, a decent number of local geometric features of the trajectory data set are still preserved, including the density distribution and local traffic flow.

CCS CONCEPTS

• **Security and privacy → Privacy protections**;

KEYWORDS

Human Mobility, Trajectory, Re-identification, Privacy Protection

ACM Reference format:
Jiaxin Ding, Chien-Chun Ni, and Jie Gao. 2017. Fighting Statistical Re-Identification in Human Trajectory Publication. In *Proceedings of SIGSPATIAL'17, Los Angeles Area, CA, USA, November 7–10, 2017,* 4 pages.
https://doi.org/10.1145/3139958.3140045

1 INTRODUCTION

The past decade has witnessed the rapid development of techniques for tracking human mobility. The rich knowledge contained in human mobility data can have a huge impact in many fields ranging from transportation to health care, from civil engineering to energy management, from e-commerce to social networking, etc.

A crucial problem that remains to be addressed is the protection of personally sensitive information. Human mobility trajectories are surprisingly unique with strong personal traits. Releasing the trajectory data to the public or third party, even after names or other identifiers are removed, can raise serious safety concerns. In the following we discuss three aspects of sensitive data that can be learned from motion trajectories.

Frequent locations. It has been discovered that human trajectories show a high degree of temporal and spatial regularity. Each individual can be characterized by a time independent characteristic travel distance and probabilities to return to a few highly frequent locations [4], which results in high predictability of individual motion [7]. Locations that are frequently visited can be related to personal identifiable information. With the knowledge of home location and work location, it is easy to identify the user with the help of an external database.

Social ties and frequent co-location patterns. A few studies show that mobility patterns shape and impact social connections. It is possible to infer 95% of social ties from the motion trajectory data alone [3], since friends show distinctive temporal and spatial features in their moving patterns. Social ties learned from mobility data can increase the impact of the revelation of sensitive data – once a single individual is identified, it is easy to identify his/her friends by examining the frequent co-location events. Thus such co-location patterns must be removed or confused sufficiently.

Unique signatures. Motion trajectories are fairly unique. In a study [2] on a dataset of fifteen months of human mobility data for one and a half million individuals, with only four spatio-temporal points one can uniquely identify 95% of the individuals. As shown in [5], an adversary, when equipped with a small amount of the snapshot information, can infer an extended view of the whereabouts of a victim in a mobility trace. Protecting the privacy of individuals should also consider hiding or confusing these unique signatures.

Our Approach

In this paper we focus on analyzing solutions that can be implemented in a distributed manner on mobile devices such that for trajectories collected from these devices, frequent locations, frequent co-locations and unique signatures are removed almost surely. Meanwhile the modified traces are still useful for applications that require faithful local geometric features such as local traffic flow. To achieve that, we build upon a simple idea in the literature named "mix-and-match", similar to the idea of the "mix-zones"[1], but without allocating static geographical regions. We evaluate the algorithm in protecting user identity and utility both analytically and with real data.

2 MODELS AND ALGORITHMS

2.1 Model

We assume that users participating in data collection would not mind sharing their trajectories if the personally identifying information can be removed. For that users follow a protocol that allows

SIGSPATIAL'17, November 7–10, 2017, Los Angeles Area, CA, USA

ACM ISBN 978-1-4503-5490-5/17/11.
https://doi.org/10.1145/3139958.3140045

trajectories to be gathered while modifications are done. We assume that users have short range communication and two users that are co-located (at the same location at the same time) can directly communicate with each other. The collected data will be trajectory traces labeled by IDs.

Threat Model. We assume that the adversary knows the entire algorithm but not the random bits. The adversary may also have background knowledge on certain participants and wish to re-identify them in the collected/published trajectories. In particular, we consider the following attacks similar to those introduced in [6]:

- *Frequent Location Attack.* The adversary knows the top h popular locations H_i of each trajectory i in the input set. After the sanitization we wish that the top h popular locations of any sanitized trajectory does not match any H_i for all i. Otherwise, we say trajectory i is *revealed*.
- *Co-location Attack.* The adversary analyses the co-location events. If user i co-locates with j, it indicates a social tie between i and j. Thereafter, the adversary can identify j as long as he identifies i.
- *Unique Signature Attack.* The adversary can possibly have ℓ spatio-temporal points P_i of each trajectory i in the input data set. If there does not exist a unique trajectory that goes through all points of P_i, we consider the trajectory i to be un-identified, otherwise i is revealed.

2.2 Algorithm

Now we present the process of trajectory collection by "mix-and-match" on a user set of n individuals, $\{u_1, u_2, \cdots, u_n\}$. Initially, each user u_i carries ID i. For any u_i, u_j, if u_i co-locates with u_j, they swap the current IDs carried by them with probability p. The meeting points separate their trajectories into incoming and outgoing segments. After the swap, the outgoing segment of u_i carries the ID j while the outgoing segment of u_j carries the ID i. See Figure 1 for an example.

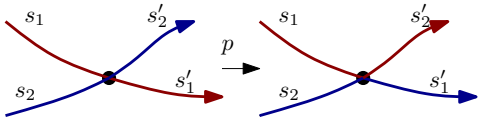


Figure 1: The original trajectories of two users are colored red and blue on the left. At the co-location point, they swap their IDs. The trajectories collected by mix-and-match is on the right.

The random mixing neither changes any location data nor does it introduce any dummy points. What is changed is how the segments are glued to one another. To evaluate the effectiveness, we introduce the *combinatorial diversity* to denote the complexity of the mixing patterns of the original trajectories, and *geometric diversity* to denote the amount of geometric changes of the generated traces from the original ones.

2.3 Combinatorial Diversity

We denote an $n \times n$ matrix $X(t)$ as the probability distribution of the IDs that each user is carrying at time t. $X_{ij}(t)$ is the probability that u_i is carrying ID j at time t. Initially, $X(0)$ is a matrix with all 1s on the diagonal and 0 otherwise. Each co-location event of two users changes the distribution matrix.

Combinatorial Diversity. Let $X_{*i}(t)$ denote the ith column vector of $X(t)$. We define the combinatorial diversity of ID i as

$$\frac{\|X_{*i}(t) - X_{*i}(0)\|}{\|X_{*i}(0)\|}, \tag{1}$$

where $\|\cdot\|$ is the ℓ_2 norm. It measures the difference between the current and the initial probability distribution of users carrying ID i. The more uniform the distribution is, the closer to $\sqrt{1-\frac{1}{n}}$ the combinatorial diversity converges to.

Theoretical Analysis. To analyze the combinatorial diversity, we introduce a simplified mobility model below. Let M denotes the co-location probability matrix derived from the mobility patterns, where M_{ij} is the probability that u_i co-locates with u_j. We assume that at the beginning of each discretized time slot t, there is a co-location event; for each co-location event, we uniformly randomly select a participant u_i, and the second participant u_j is selected with probability M_{ij}. Based on this model, we have the following theorem.

THEOREM 2.1. *For any i, $\frac{\|X_{*i}(t)-X_{*i}(0)\|}{\|X_{*i}(0)\|} \geq \sqrt{1-\frac{1}{n}} - \delta$, for $t \geq \frac{\log(\varepsilon^{-1})+2\log(\delta^{-1})}{\log \lambda_2(\overline{W^2})^{-1}}$, with probability $1-\varepsilon$, where $\overline{W^2} = I - \frac{2p(1-p)}{n}D + \frac{2p(1-p)}{n}(M+M^T)$, M is the co-location pattern described above, D is a diagonal matrix $D_{ii} = \sum_{j=1}^{n}[M_{ij}+M_{ji}]$, n is the number of users.*

The theorem shows that based on the mobility model, the probability distribution of IDs can be arbitrarily uniform with high probability given sufficient time. This guarantees the performance of the "mix-and-match" procedure.

2.4 Geometric Diversity

When the combinatorial diversity is high enough to ensure that the IDs are mixed well, we examine the geometric diversity measured by the expectation of distance between the original and the generated trajectories after ID swaps.

Denote by r the time parameter for the trajectories and denote by $d_{ij}(r)$ the distance between the original u_i's trajectory and u_j's trajectory at time r. The geometric diversity of u_i's trajectory at time r is defined as the expected distance between the generated trajectory with ID i and the original u_i's trajectory at time r:

$$\sum_{j=1}^{n} d_{ij}(r) X_{ij}(r) \tag{2}$$

2.5 Protection from Re-identification

Now we design experiments to show how the "mix-and-match" procedure protects user privacy on sensitive information attacks.

Frequent locations. We first find out the number of frequent locations h needed to identify an individual in the original trajectories. We estimate the chance of defeating the attack, i.e., the adversary cannot find a trajectory after "mix-and-match" with the h frequent locations matching that of any user's original trajectory.

Frequent co-location patterns. We first analyse the distribution of cumulative time all pairs of individuals co-locate in the original trajectories and choose the threshold above which the co-location time can indicate a social tie. A graph can be constructed from the trajectories where there is an edge between two individuals if

they co-locate longer than the threshold. We compare the degree distribution and the max clique size between the graph of the original trajectories and the one of the generated trajectories after the "mix-and-match" procedure.

Unique signatures. On the last measure, we first examine the number of random spatio-temporal data points ℓ needed to identify an individual in the original trajectory set. Then, using ℓ random spatio-temporal locations of a user, we check the chance that the adversary can find any matching trajectory in the generated trajectories.

2.6 Utility

The "mix-and-match" process preserves all the location information and most localized linkages between the adjacent data points of the same trajectory. Therefore any data mining algorithm that focus on the local geometric patterns will be able to retrieve high quality data from the anonymized trajectories. We look at two utility measures evaluated in the experiment section.

Segment Time Durations. In the "mix-and-match" procedure, the trajectories are divided into segments at the co-location events. We analyze the time durations of these segments. A longer time duration can keep more information of the original trajectory.

Traffic Flow Estimation. Using the generated trajectories, we can estimate the traffic flow, as well as mining sequential patterns. To be specific, we can estimate the amount of traffic visiting a sequence of k locations consecutively.

3 EXPERIMENTS

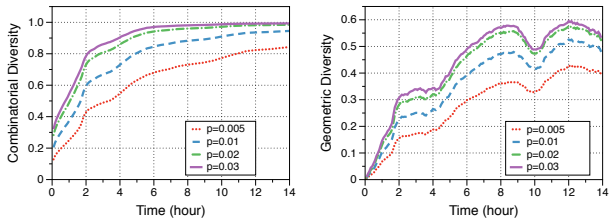
We analyze the "mix-and-match" procedure with real human mobility data in one week. In the data set, we collect trajectories of 3, 629 students with the WiFi Access Points on campus for one week. The radius of the campus is 1 km, and WiFi Access Points are deployed in 106 buildings of the campus. The connections between mobile devices and WiFi Access Points are recorded to form the trajectories. The traces are in the resolution of buildings and sampled every 5 minutes. The trajectory of each individual is in the form of $[(t_1, Loc_1), (t_2, Loc_2), \cdots, (t_i, Loc_i), \cdots]$, where t_i is time and Loc_i is the building visited at time t_i. For each individual, the number of spatio-temporal sample points ranges from 605 to 2, 028 for the week. The maximum number of buildings one individual visits during the week is 29. At most 192 people appear simultaneously in the same building and on average 26 individuals co-locate at each time slot.

Our major results of the "mix-and-match" process with the swapping probability $p = 0.02$ are summarized as follows:

- **Defeat Adversary Attacks.** With our mixing ID approach, the chance that the adversary failed the frequent location attack, co-location attack, and unique signature attack is higher than 99%.
- **High Utility.** In the generated mixing ID approach data set, the estimation of the number of people in the building and the traffic flow between buildings are 100% accurate.

3.1 Combinatorial Diversity

We choose the exchange probabilities $p = 0.005, 0.01, 0.02, 0.03$. In Figure 2(a), the higher the swapping probability p is, the faster the combinatorial diversity converges. During the first 14 hours, all the average combinatorial diversities for the different swapping probability p converge to the theoretical lower bound, approximate 1, meaning that the IDs are mixed up uniformly.



(a) Average combinatorial diversity (b) Average geometric diversity

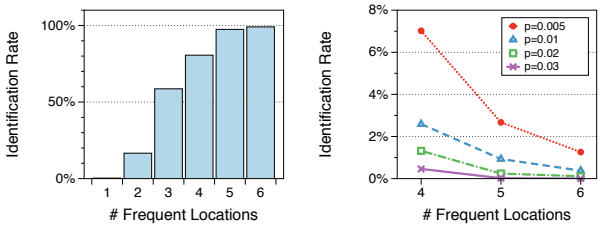
Figure 2: The combinatorial diversity, and the geometric diversity distribution among all trajectories during 14 hours.

3.2 Geometric Diversity

In Figure 2(b), the value of geometric diversity goes up with the increase of p; the average geometric diversities increase sharply in the first 2 hours, then continue to increase slightly; after the 6th hour, the geometry diversities concentrate around 0.6km, which is more than half of the radius of the campus, indicating an expected large shift between the generated and the original trajectories.

3.3 Frequent Location Attack

We first analyze the number of frequent locations h we need to uniquely identify one trajectory in the original data. In Figure 3(a), we show that if fewer than 3 most frequent locations are revealed to the attacker, the identification rate is lower than 60%, while for $h \geq 4$, the attacker's identification rate is higher than 80%. As shown in Figure 3(b), even the adversary knows the top 4 frequent locations, he/she only has 7% chance to recover the correct identity, and the chance goes even lower with larger number of frequent locations revealed. The identities are well protected from frequent location attack.



(a) The identification rate with top h frequent locations provided from the original data set.

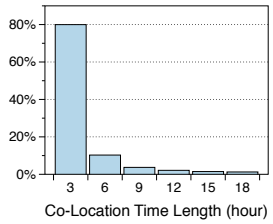
(b) The identification rate in the newly generated trajectories with top h frequent locations in the original trajectories.

Figure 3: Frequent location attack.

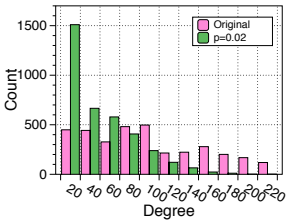
3.4 Co-location Attack

We study the distribution of the total co-location time lengths between two individuals in the original trajectories during the week, shown in Figure 4(a). Here, over 80% of co-locations are within 3 hours and only 3.89% are above 12 hours. We assume that the co-location length of 12 hours is enough to indicate a relationship.

When constructing the co-location graph, every identity is denoted as a node, and there is an edge between node i and j if the total co-location time length between user i and j is above the threshold of 12 hours. We compare the co-location graph with the original trajectories and the generated. Notice that the results for different p are similar in the experiment, therefore, we only show the statistics for $p = 0.02$. We denote the original graph as G_1, the graph with generated data as G_2. The number of edges in G_1 is 177, 195, while the number of G_2 is only 67, 407, 63% less. The degree distributions of nodes in G_1 and G_2 are shown in Figure 4(b). The degrees of 90% nodes in G_2 are below 80, while that in G_1 have a longer "tail": over 27% nodes in G_1 have degrees over 160. We further examine the size of the maximum clique for the two graphs. The value for G_1 is 146, while the one for G_2 is much smaller, 35. G_2 is dissimilar with G_1. In summary, the co-location attack can be well defended.



(a) The histogram of co-location time lengths.

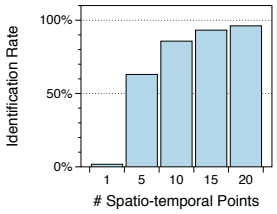


(b) The histogram of degrees of the original relationship graph and the graph with generated trajectories.

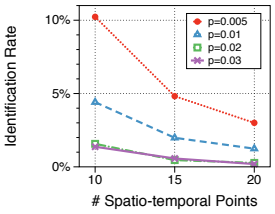
Figure 4: Co-location attack.

3.5 Unique Signature Attack

We demonstrate the fraction of individuals that can be uniquely identified with ℓ random spatio-temporal points in the original data set, shown in Figure 5(a). With $\ell \geq 10$, the identification rate is over 80%. Assume the adversary collects ℓ spatio-temporal points on the original trajectory i. If the adversary finds a match in the generated data set, we define that the trajectory i is identified. The identification rate is plotted in Figure 5(b). The highest identification rate 10% appears when $\ell = 10, p = 0.005$. Even in that case, the generated trajectory identified by the adversary as trajectory i only share 14% spatio-temporal points with the original trajectory i. If we choose $p \geq 0.02$, the identification rate is smaller than 0.5% for $\ell \geq 15$. The unique signature attack is defeated.



(a) The identification rate with ℓ random spatio-temporal points provided from the original data set.

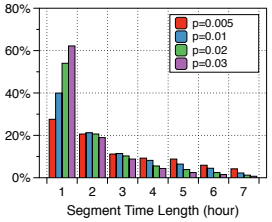


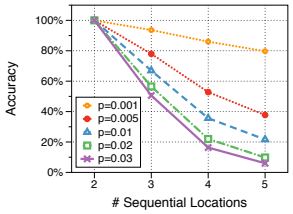
(b) The identification rate in the generated trajectories with ℓ random spatio-temporal points.

Figure 5: Unique signature attack.

3.6 Utility Analysis

Segment time durations. We analyze the time durations of the segments, in Figure 6(a). For $p = 0.03$, 62% of the segment time durations are within 1 hour, while for $p = 0.005$, only 27% are within 1 hour, and over 53% are greater than 3 hours. The smaller the swapping probability is, the longer the time durations of the segments are. We can carefully choose p to get a trade off between the privacy and utility.





(a) The histogram of time durations of segments without ID swaps for different p.

(b) The accuracy of traffic flow estimation of k sequential locations with the generated trajectories.

Figure 6: Utility.

Traffic flow estimation. With the "mix-and-match" algorithm, the queries of the number of individuals in each building and the traffic between two buildings reach 100% accuracy . The accuracy of the traffic estimation of k sequential locations is demonstrated in Figure 6(b). When $k = 3$, we can still have an accuracy over 55% for all p. As k increases, the estimation accuracy decreases sharply for $p \geq 0.005$. In the case of $p = 0.001$, even for $k = 5$, the accuracy is still above 80%. The utility for sequential pattern mining is high.

4 CONCLUSION

This paper studies the mixing ID approach for removing identifying features from a trajectory set, with theoretical analyses and performance demonstrated in real human mobility data sets. We show that the generated trajectories defeat the attacks from the adversary and still have good performances for local traffic analysis.

ACKNOWLEDGMENTS

The authors would like to acknowledge support through NSF DMS-1418255, CCF-1535900, CNS-1618391, DMS-1737812 and AFOSR FA9550-14-1-0193.

REFERENCES

[1] A.R. Beresford and F. Stajano. 2003. Location privacy in pervasive computing. *Pervasive Computing, IEEE* 2, 1 (Jan 2003), 46–55.

[2] 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 (25 March 2013).

[3] Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* 106, 36 (2009), 15274–15278.

[4] Marta C. González, César A. Hidalgo, and Albert-Laźló Barabási. 2008. Understanding Individual Human Mobility Patterns. *Nature* 453 (June 2008).

[5] Chris Y.T. Ma, David K.Y. Yau, Nung Kwan Yip, and Nageswara S.V. Rao. 2010. Privacy Vulnerability of Published Anonymous Mobility Traces. In *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking (MobiCom '10)*. ACM, New York, NY, USA, 185–196.

[6] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. 2011. Quantifying location privacy. In *Security and privacy (sp), 2011 ieee symposium on*. IEEE, 247–262.

[7] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of Predictability in Human Mobility. *Science* 327, 5968 (2010), 1018–1021.