Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data

Yi Song
SAP Research & Innovation*
National University of Singapore
yi.song01@sap.com

Daniel Dahlmeier
SAP Research & Innovation
d.dahlmeier@sap.com

Stephane Bressan
National University of Singapore
steph@nus.edu.sg

ABSTRACT

We study the problem of privacy in human mobility data, i.e., the re-identification risk of individuals in a trajectory dataset. We quantify the risk of being re-identified by the metric of uniqueness, the fraction of individuals in the dataset which are uniquely identifiable by a set of spatio-temporal points. We explore a human mobility dataset for more than half a million individuals over a period of one week. The location of an individual is specified every fifteen minutes. The results show that human mobility traces are highly identifiable with only a few spatio-temporal points. We propose a modification-based anonymization approach that is based on shorting the trajectories to reduce the risk of re-identification and information disclosure. Empirical, experimental results on the anonymized dataset show the decrease of uniqueness and suggest that anonymization techniques can help to improve the privacy protection and reduce privacy risks, although the anonymized data cannot provide full anonymity so far.

1. INTRODUCTION

The availability of mobility and location data around us is exploding due to the prevalence of mobile devices such as cell phones and tablets. Mobility traces of people are now routinely collected at a large scale, for example, by cellular network operators, location-based services, and location-enabled social network platforms. The study of human mobility can potentially unlock great value for both commercial players, as well as the public sector. Location data can, for example, assist city traffic planning, and intelligent transportation [9], as human movement patterns are not likely to significantly change over time [3, 22, 20, 18]. Individuals can also directly benefit from location-based services which provide personalized services to smartphone and tablet users, such as navigation, tracking, and recommendations for entertainment or new friendships. These location-based services heavily rely on the availability of location data, for example through location information sharing and location-aware information retrieval [17].

However, serious threats are posed to users' privacy when they share their location data with location-based service providers via queries for location-based information. Moreover, with the increasing need and desire to share or publish location information, the privacy concerns are significant. Various potentially sensitive details about the users personal information can be inferred with mobility traces [1]. Removing personal identifiers, e.g., name or social security number, is not enough for privacy protection [19]. Although the published trajectories are often made anonymous in that the true identities of individuals are replaced by random identifiers, the individuals are highly identifiable when partial knowledge of their whereabouts are publicly observable or disclosed by themselves voluntarily. Interested third parties can learn such information directly or indirectly, and the privacy concern remains. Ma et al. [14] proposed several privacy attacks in which adversaries are equipped with different amounts of information about the target. Their investigation shows that a relatively small amount of snapshot information is sufficient for the adversary to re-identify a target in a set of anonymous traces or infer the whereabouts of a target either uniquely or with high probability.

Anonymization approaches have been proposed to help improve privacy protection, e.g., by reducing the granularity of location information. However, the ability of privacy-preserving mechanisms to protect privacy is in question. DeMulder et al. [6] demonstrate that even though cell locations blur the exact locations of users, a sequence of cells allows an adversary to identify individuals with a very high probability. Using real-world location traces of mobile users and measuring the rates of correct identification of anonymized traces, they assess the extent to which anonymized location records from cell-based mobile phone networks can be linked back to previously extracted user profiles. Their work concludes that removing identifiers from location information, or reducing the granularity of the location or time, does not prevent disclosure of personally identifiable information.

In this paper, we study trajectory privacy based on a large-scale mobility dataset, The dataset contains spatio-temporal points with true identifiers replaced by synthetic identifiers. All points with the same identifier form a trajectory for the corresponding user. We quantify the privacy risk by examining the uniqueness of the trajectories when the adversary has different amounts of partial knowledge. Specifically, we assume that the adversary may know a certain number of spatio-temporal points among the trace of a target user. We measure the number of trajectories that the adversary can find based on the existing knowledge. The trajectory is unique and re-identification is successful if only one trajectory is found. Our results show that human mobility traces are highly identifiable, even with only a few spatio-temporal points.

To reduce the privacy risks, we propose a simple and effective anonymization method. The main idea of the method is to "cut" long trajectories into several short trajectories

*This work was done during an internship at SAP.

according to different time windows. These shorter trajectories are then assigned different user identifiers for each time window. We show that the uniqueness measured on the anonymized trajectories is reduced, and thus the privacy risk is being decreased.

Anonymization always comes at the cost of data utility [2, 13]. Thus, the success of the anonymization method heavily depends on the success of preserving the data utility. Trajectory anonymization techniques are expected to preserve privacy while retaining data utility to support useful queries, e.g., aggregated analysis and temporal queries [4]. Our anonymization method maintains the a high level of data utility by retaining all the original information in terms of the spatio-temporal points within each time window and adding no dummy points or false information.

The rest of the paper is organized as follows. Section 2 presents the related work on privacy and anonymization of location and trajectory data. Section 3 describes our anonymization approach. Section 4 describes the dataset and pre-processsing. Section 5 presents experiments and results. Section 6 discusses the findings. We conclude the paper in Section 7.

2. RELATED WORK

There has been a significant amount of work that has studied privacy and anonymity in location-based data, including methods to quantify privacy risks and anonymization techniques to counter such risks. In this section, we provide an overview of prior work that is most relevant to our work.

Zang et al. [21] examine the human mobility by the *top N* locations for each individual in a large-scale dataset of call data records. They consider anonymization techniques based on generalization of the granularity level in the time domain or in the spacial domain. They compare the number of users that can be uniquely identified by a given set of such locations at different granularity levels both before and after anonymizing the data. They find that releasing location data anonymized with their method still carries a high risk of privacy breach or the data needs to be very coarse in the time domain or space domain, in which case the data utility decreases significantly. They measure the utility of the anonymized traces by the cumulative density function and the entropy of the locations visited by each user at different granularity levels.

Golle and Partriage [10] examine the re-identification risk based on home and work locations based on a very large dataset representative of the whole U.S. working population. The concept of anonymity set is defined for measuring privacy. They point out that the location traces are at great risk when both the home and work locations can be deduced. To prevent the traces from re-identification, a considerable amount of location obfuscation, coarsening the data spatially, is needed before release.

Neigiz et al. [15] adopt the notion of k-anonymity [19] to trajectories and propose a novel generalization-based approach for trajectory anonymization as well as a randomization-based reconstruction algorithm for releasing anonymized trajectory data. The effectiveness of the proposed techniques is tested on both real and synthetic data and measured by a log distance. The method is effective in that every trajectory is indistinguishable from k-1 other trajectories.

Gao et al. [8] propose an anonymization model based on the same notion, k-anonymity. They consider trajectory similarity and direction for finding optimal anonymity sets and trajectory distance for data utility. Their experiments on synthetic data shows the effectiveness of their model regarding both privacy protection and data utility.

Freudiger et al. [7] focus on quantifying the privacy risks induced by using location-based services. They evaluate the success of location-based services in predicting the true identities of pseudonymous users and their points of interest on real mobility traces. They confirm the ability of location-based service providers to uniquely identify users based on a small number of location samples observed from the users. The effects of data type and quantity to the identification ability are explored. Our work is different from theirs as we further investigate the effects of the anonymization method that we propose.

Shin et al. [17] give an overview of the existing privacy protection schemes. Privacy is classified into two groups: query privacy e.g., whether a user can be identified, and location query e.g., whether a user can be accurately located. The privacy protection schemes are reviewed from three categories: policy, location perturbation and obfuscation, and private information retrieval based approaches.

A good survey of state-of-the-art privacy-preserving techniques can be found in [4]. The authors give an overview of location privacy, trajectory privacy, and the anonymization techniques respectively, e.g., false locations and space transformation for location anonymization. Spatial cloaking, mix-zones for trajectory anonymization. They summarize and categorize several anonymization techniques, e.g., [12, 11, 16].

Montjoye et al. [5] who find that human mobility traces are highly unique by studying fifteen months of human mobility data for one and a half million individuals. Quantifying the privacy risk by measuring the uniqueness of human mobility traces on both original data and spatially and temporally coarsen data, they conclude that even coarse datasets provide little anonymity. Our work is inspired by their work. Similar to their approach, we quantify privacy risk by estimating the uniqueness of trajectories. However, our work differs in that we calculate the average uniqueness based on the *whole* dataset rather than based on a random sample from the dataset. We also propose an anonymization method and further investigate the effectiveness of the method.

3. TRAJECTORY ANONYMIZATION

A trajectory tr_i consists of an ordered set of spatio-temporal points, denoted as $\{< p_1, t_1 >, < p_2, t_2 >, ..., < p_j, t_j > , ..., < p_n, t_n >\}$ where each tuple $p_j =< x_j, y_j >$ represents a point location with geographic coordinates x_j as longitude and y_j as latitude. t_j represents the corresponding timestamp. The number of spatio-temporal points n equals to the size of the set $|tr_i|$, i.e., the total number of points in trajectory tr_i. The total number of trajectories in the dataset is denoted as N.

3.1 Problem Definition

Let $\{tr_i\}_{i=1,...,N}$ be the original trajectory dataset consisting of a set of N static trajectories, i.e., the trajectories are fixed and are not changing or being extended any more. Each trajectory tr_i is associated with a unique synthetic identifier u_i. We assume an adversary whose goal is to re-identify one or more trajectories in the dataset, that is the

adversary's goal is to be able to map one or more synthetic identifiers to real user identities. In this paper, we assume that the adversary is equipped with additional knowledge of the partial trajectory of a user in the form of a limited number of points that that the user has visited at particular times. In other words, we assume that the adversary knows part of the trajectory which is denoted as a number of m spatio-temporal points $\{p_j|1 \leq j \leq n\}$.

The anonymity or privacy risk associated with the release of dataset is quantified by the *uniqueness* of the trajectories [5]. Uniqueness of trajectories is defined as follows:

$$uniqueness = \frac{\sum \delta_i}{N} \quad (1)$$

where $\delta_i = 1$ if $|\{tr_i|\{p_j|1 \leq j \leq n\} \subseteq tr_i\}| = 1$. Otherwise $\delta_i = 0$ It measures how likely a trajectory can be re-identified by the adversary. High *uniqueness* indicates a high probability of success of re-identification, and thus high privacy risk. The goal of anonymizing the trajectory dataset is making it difficult for an adversary to re-identify trajectories by decreasing the uniqueness of the trajectories. In the following section, we show a simple and efficient algorithm to decrease the uniqueness of the trajectories.

3.2 Anonymization method

To reduce the privacy risk, we propose a simple but effective method for anonymizing trajectory data. Our method is based on the insight that the uniqueness of a user's trajectory increases with the length of the trajectory. Take, for example, the trajectory of a single user over a duration of 24 hours. For the trajectory to be not unique, there has to be at least one other user who has been in the same location as the first user for every point in time during that 24 hour interval. It is obvious that the chance of such a set of other users existing is low and that the chance is diminishing the longer the trajectory is. On the other hand, for a short period of time, let's say a few hours, we can expect that there is a good chance of other users being in the same location, at least in a densely-populated urban environment. Instead of reducing the resolution of the location information, we disintegrate the trajectories into a set of shorter sub-trajectories for different time windows by "cutting" the original trajectories into shorter sub-trajectories that we expect to have lower uniqueness. Note that our method provides a simple mechanism to balance privacy and utility of the trajectories. At one extreme, we can cut all trajectories into sub-trajectories of length one, essentially reducing the trajectories to a density map which has high privacy guarantees but destroys all information about the movement patterns of the users, at the other extreme we can decide not to cut the trajectories at all, keeping the original data with all information but without adding any privacy.

Formally, a trajectory tr_i with points $\{< p_1, t_1 >, < p_2, t_2 > , ..., < p_j, t_j >, ..., < p_n, t_n >\}$ can be divided into k sub-trajectories $\{tr_{ij}\}_{j=1,...,k}$ according to the timestamps of the points. Let t be the whole recording period of the dataset. Each sub-trajectory $tr_{ij} \subseteq tr_i, 1 \leq j \leq k$ of trajectory tr_i contains a set of points $\{< p_m, t_m > |(j-1)\frac{t}{k} \leq t_m \leq j\frac{t}{k}\}$ that fall into the j-th time window. For each sub-trajectory, we assign a new random user identifier u_{ij}, thus effectively "cutting" the original trajectories into shorter sub-trajectories that cannot directly be linked together to the original trajectory. Figure 1 shows an example of a trajectory that is "cut" into multiple sub-trajectories with window size of 6 hours.

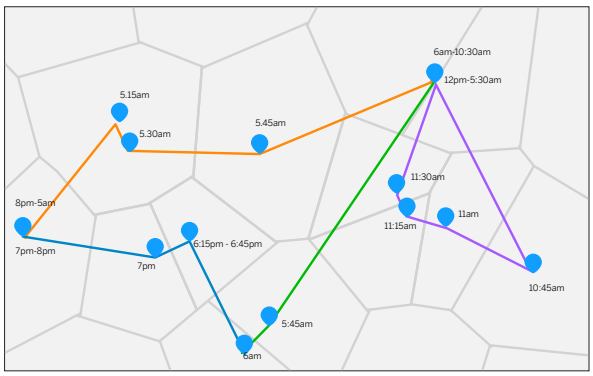


Figure 1: An example of a trajectory that is "cut" into several sub-trajectories with window size 6 hours, indicated by different colors.

The only input parameter for this method is the window size, i.e., the length of duration. The method is simple and efficient. It has a low computation complexity which is linear in the number of records. Therefore it is scalable on very large datasets.

4. DATASET DESCRIPTION AND PREPROCESSING

Our dataset contains one week of mobility data for 1.14 million people with 56 million records in total. Each record consists of one user identifier and one spatio-temporal point $<x_i, y_i, t_i>$. The true identities have been replaced by synthetic identities. The location of an individual was recorded every fifteen minutes. The spatial resolution of the data is equal to that of a fixed set of discrete locations rather than the exact locations of the users. The whole dataset contains about 1,700 unique locations.

Distances between two locations can be calculated by their Euclidian distance (2-norm): $||p_i - p_j||$. The smallest distance ($min||p_i - p_j||$) between any two points is about 0.11 km while the largest distance ($max||p_i - p_j||$) is about 49 km.

We preprocess the dataset in two steps. First, we filter the raw data to increase the data quality. Like any large, real-world data set, the original data contains some noise which should be removed to arrive at a more meaningful analysis. We found that the original dataset contained some duplicate records as well as many "singleton" users with only one location throughout the whole week. We filter out the duplicate records from the dataset and remove records for all users with only one record location to improve the efficiency of uniqueness computation. The filtered dataset contains 0.63 million users. Second, we extract trajectories from the records, i.e., we extract the ordered set of spatio-temporal points for each user. Each user u_i corresponds to exactly once trajectory tr_i. Table 1 shows an overview of the preprocessed dataset.

number of locations	$\sim$ 1,700
number of timestamps	672
number of trajectories	633,798
records in the original data	$\sim$ 56,000,000
records in the cleaned data	$\sim$ 43,000,000
users in the original data	$\sim$ 1,100,000
users in the cleaned data	$\sim$ 630,000
minimum number of points in one trajectory	3
maximum number of points in one trajectory	672
average number of points in one trajectory	56

Table 1: Statistics of dataset

5. EXPERIMENTAL ANALYSIS

We estimate the average number of points needed to uniquely identify the trajectory of an individual by calculating the uniqueness on the dataset both before and after anonymization.

To estimate the uniqueness of a user's trajectory, we use a sampling-based approach similar to [5]. For each individual u_i, we randomly select m distinct spatio-temporal points among all the points of her trajectory t_i. We check through all other users based on the select points, and count the number of users having the same points. That is, we look for other users that have been at the same places at the same times. This number corresponds to the term $|\{tr_i|\{p_j|1 \leq j \leq n\} \subseteq tr_i\}| = 1$ in Equation 1. Let us denote the size of this set as s_i for trajectory t_i . Note that if $s_i = 1$, the user has successfully been re-identified. To quantify the uniqueness of the dataset, we compute the fraction of samples for which $s_i = 1$.

5.1 Experimental Results

Figure 2 shows the comparison of uniqueness before and after anonymization for different numbers m of points that are randomly chosen. The x-axis represents the m number of random points selected. The y-axis represents the estimated value of uniqueness. The uniqueness for the dataset before anonymization and for the dataset after anonymization is compared. We anonymize the dataset with three different time window sizes: 6 hours, 12 hours, and 24 hours. The uniqueness of the anonymized dataset decreases notably, especially when the duration is 6 hours. The uniqueness decreases by more than 0.2 from 0.6 to around 0.4, in the case of two random points.

Figure 3 shows the distribution of $\{s_i\}$, the number of individuals whose trajectories include the same spatio-temporal points randomly chosen for every individual. The anonymized data in this case is anonymized by 6 hours and the number randomly chosen points is 2. The x-axis is the number of users that have the same spatio-temporal points selected for the target user. The y-axis is the normalized count of individuals that have certain number of individuals that have the same spatio-temporal points. Note the large gap between the frequency for the uniquely identifiable individuals which reduces from 0.6 in the original data to 0.4 in the anonymized data.

Figure 4 (a) shows the same results as Figure 3 in the logarithmic scale and with larger x-axis range. Here, we

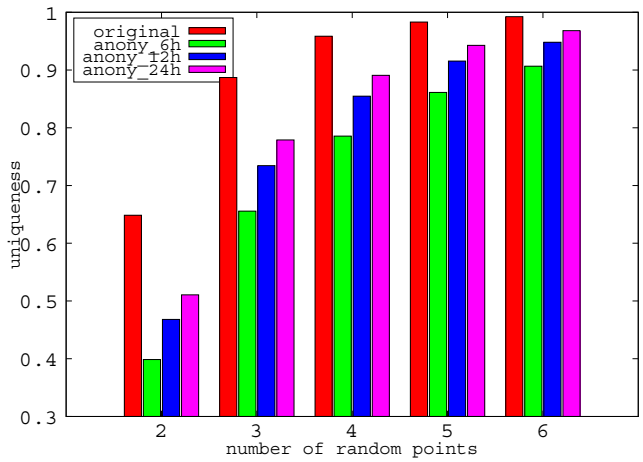


Figure 2: Comparison of uniqueness before and after anonymization. The x-axis represents the number of random points selected. The y-axis represents the estimated value of uniqueness. Lower values are better.

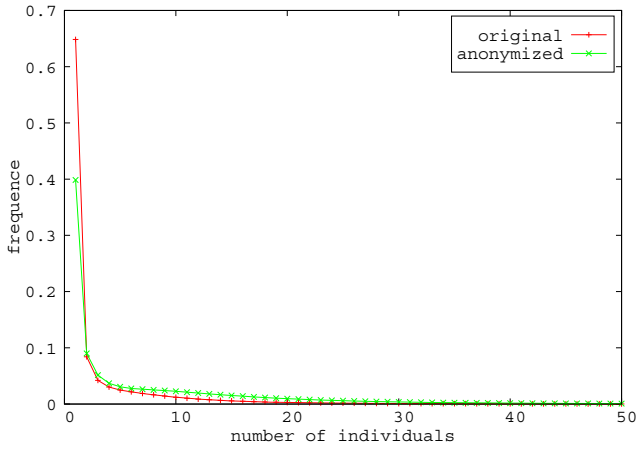


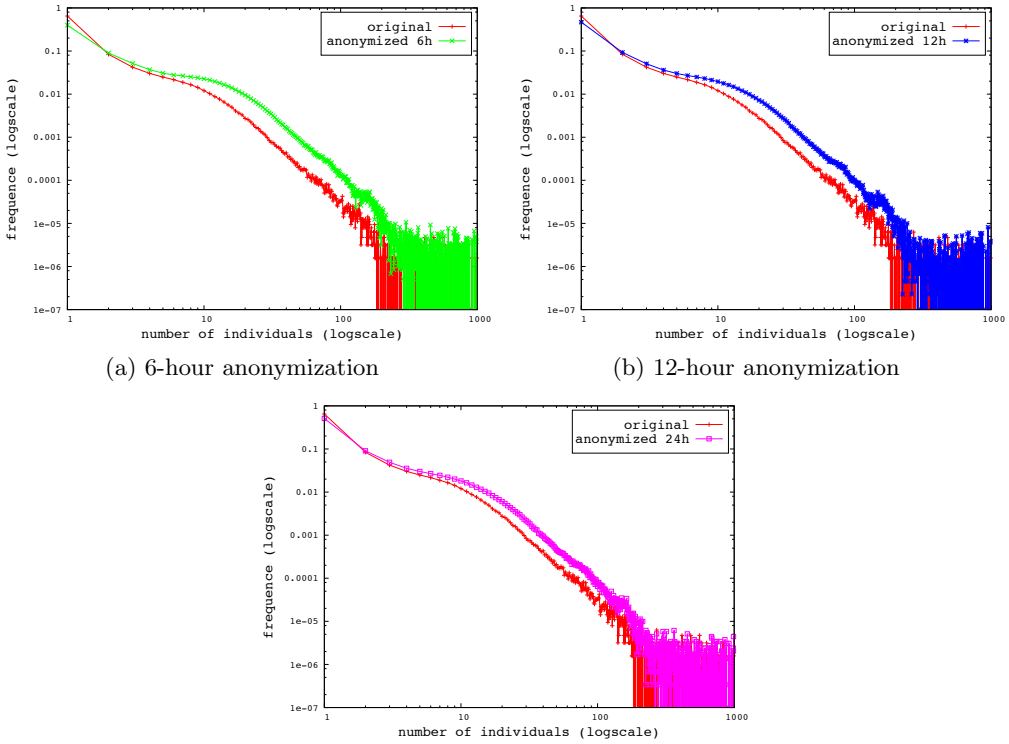
Figure 3: Distribution of the number of individuals that have been to the same places at the same times.

can see that there are more individuals who have at least two other individuals that have the same spatio-temporal points.

Figures 4 (b) and (c) compare the distribution of $\{s_i\}$ on the data before anonymization and the data after being anonymized by different window sizes. We can see that the decrease in the frequency for $s_i = 1$ is larger if the the duration of the time window is smaller, which meets our original intuition that shorter trajectories should be less unique. In other words, the shorter we "cut" the trajectories, the fewer individuals are uniquely identifiable.

6. DISCUSSION

The dataset before anonymization here refers to the original dataset which is only anonymized by having the original identifiers replaced by synthetic identifiers. The locations in the dataset are the locations of the antennas so the exact locations of individuals are already blurred to some extend.



(a) 6-hour anonymization

(b) 12-hour anonymization

(c) 24-hour anonymization

Figure 4: Distribution of the number of individuals that have been to the same places at the same times in a logarithmic scale for different time windows.

However, the assessment of uniqueness indicates that with two random points, more than 60 percent of the trajectories are unique. Therefore human mobility trajectories are highly re-identifiable and the privacy risk is high. However, it is possible to reduce the risks through our anonymization approach. The empirical experimental results show that our simple anonymization method reduces the uniqueness by over 30%.

From the data utility perspective, we intend to keep information loss low. Unlike most of the other methods that generalize or lower the resolution of the dataset spatially or temporally, our method keeps the original granularity on both dimensions. Consequently, the anonymized data can answer detailed queries that the coarse dataset cannot, e.g., how many individuals have travelled between two locations at a designated time.

7. CONCLUSIONS

In this paper, we study location privacy. Specifically we study the re-identification risk of trajectories in human mobility data based on a large dataset of more than half a million individuals over a period of one week. We empirically assess how unique human trajectories are. We find that individuals are highly re-identifiable with only a few spatio-temporal points. Releasing such data will pose serious privacy risks. We propose a simple anonymization approach to modify the dataset by shortening the trajectories. Examining the uniqueness on the anonymized data, we conclude that anonymization techniques can help improve the privacy protection and reduce the risks of re-identification and information disclosure, although the anonymized data cannot provide full anonymity.

Acknowledgement

The research is partially funded by the Economic Development Board and the National Research Foundation of Singapore.

8. REFERENCES

[1] A. J. Blumberg and P. Eckersly. On locational privacy, and how to avoid losing it forever. 2009.

[2] J. Brickell and V. Shmatikov. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 70–78, New York, NY, USA, 2008. ACM.

[3] 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*, KDD '11, 2011.

[4] C.-Y. Chow and M. F. Mokbel. Trajectory privacy in location-based services and data publication. *SIGKDD Explor. Newsl.*, 13(1):19–29, 2011.

[5] 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.

[6] Y. De Mulder, G. Danezis, L. Batina, and B. Preneel. Identification via location-profiling in gsm networks. In *Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society*, WPES '08, pages 23–32, New York, NY, USA, 2008. ACM.

[7] J. Freudiger, R. Shokri, and J.-P. Hubaux. Evaluating the privacy risk of location-based services. In *Proceedings of the 15th International Conference on Financial Cryptography and Data Security*, FC'11, pages 31–46, Berlin, Heidelberg, 2012. Springer-Verlag.

[8] S. Gao, J. Ma, C. Sun, and X. Li. Balancing trajectory privacy and data utility using a personalized anonymization model. *J. Netw. Comput. Appl.*, 38:125–134, 2014.

[9] F. Giannotti and D. Pedreschi. *Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*. Springer Publishing Company, Incorporated, 1 edition, 2008.

[10] P. Golle and K. Partridge. On the anonymity of home/work location pairs. In *Proceedings of the 7th International Conference on Pervasive Computing*, Pervasive '09, pages 390–397, Berlin, Heidelberg, 2009. Springer-Verlag.

[11] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Achieving guaranteed anonymity in gps traces via uncertainty-aware path cloaking. *IEEE Trans. Mob. Comput.*, pages 1089–1107, 2010.

[12] P.-R. Lei, W.-C. Peng, I.-J. Su, and C.-P. Chang. Dummy-based schemes for protecting movement trajectories. *J. Inf. Sci. Eng.*, 28(2), 2012.

[13] T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 517–526, New York, NY, USA, 2009. ACM.

[14] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao. Privacy vulnerability of published anonymous mobility traces. In *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking*, MobiCom '10, pages 185–196, New York, NY, USA, 2010. ACM.

[15] M. E. Nergiz, M. Atzori, and Y. Saygin. Towards trajectory anonymization: A generalization-based approach. *Trans. Data Privacy*, 2(1):47–75, 2009.

[16] X. Pan, X. Meng, and J. Xu. Distortion-based anonymity for continuous queries in location-based mobile services. In *GIS*, pages 256–265, 2009.

[17] K. G. Shin, X. Ju, Z. Chen, and X. Hu. Privacy protection for users of location-based services. *IEEE Wireless Commun.*, 19:30–39, 2012.

[18] C. Song, Z. Qu, N. Blumm, and A.-L. BarabÃąsi. Limits of predictability in human mobility. *Science*, 327:1018–1021, 2010.

[19] L. Sweeney. K-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 2002.

[20] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Measuring serendipity: Connecting people, locations and interests in a mobile 3g network. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '09, pages 267–279, New York, NY, USA, 2009. ACM.

[21] 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*, MobiCom '11, pages 145–156, New York, NY, USA, 2011. ACM.

[22] H. Zang and J. C. Bolot. Mining call and mobility data to improve paging efficiency in cellular networks. In *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking*, MobiCom '07, pages 123–134, New York, NY, USA, 2007. ACM.