



“When and Where Do You Want to Hide?” – Recommendation of Location Privacy Preferences with Local Differential Privacy

Maho Asada, Masatoshi Yoshikawa, and Yang Cao^(✉)

Kyoto University, Kyoto, Japan
asada@db.soc.i.kyoto-u.ac.jp
{yoshikawa,yang}@i.kyoto-u.ac.jp

Abstract. In recent years, it has become easy to obtain location information quite precisely. However, the acquisition of such information has risks such as individual identification and leakage of sensitive information, so it is necessary to protect the privacy of location information. For this purpose, people should know their location privacy preferences, that is, whether or not he/she can release location information at each place and time. However, it is not easy for each user to make such decisions and it is troublesome to set the privacy preference at each time. Therefore, we propose a method to recommend location privacy preferences for decision making. Comparing to existing method, our method can improve the accuracy of recommendation by using matrix factorization and preserve privacy strictly by local differential privacy, whereas the existing method does not achieve formal privacy guarantee. In addition, we found the best granularity of a location privacy preference, that is, how to express the information in location privacy protection. To evaluate and verify the utility of our method, we have integrated two existing datasets to create a rich information in term of user number. From the results of the evaluation using this dataset, we confirmed that our method can predict location privacy preferences accurately and that it provides a suitable method to define the location privacy preference.

Keywords: Privacy preference · Location data · Matrix factorization · Local differential privacy

1 Introduction

In recent years, due to the popularization of smartphones and the development of GPS positioning equipment, location information for people has been able to be obtained quite precisely and easily. Such data can be utilized in various fields such as marketing and urban planning. In addition, there are many applications

that do not function effectively without location information [9]. Because of such value, market maintenance to buy and sell it has started [9].

However, on the other hand, by publishing accurate location information, there are privacy risks associated with such as individuals being identified [3]. Due to such risks, privacy awareness regarding location information among people is very high. One of the most risky situations is when smartphones are used. This is because we are sending location information to them when using many applications [5].

In order to prevent privacy risks under such circumstances, it is necessary to anonymize or obfuscate location information. One of the countermeasures is a primitive one: turn off location information transmission manually when using a smartphone. There is also a method of applying a location privacy protection technique. Various techniques are used for protection, including k -anonymity [7, 14], differential privacy [4] [8], and encryption [16]. Fawaz [5] proposed a system that applies these privacy protection technologies to smartphones. This system controls the accuracy of the location information sent to each application. The user needs to input how accurate he/she wants to send the respective location information to each application.

In these countermeasures, to avoid a privacy risk by the disclosure of location information, the user has to decide location privacy preference for each location, that is, whether or not he/she publishes the location data at a certain place and time. However, we think that there is a problem in such a situation. What is the best privacy preference is unclear, and it may be different for each user. Therefore, individual users need to determine their location privacy preferences. However, most users find it difficult to determine these preferences themselves [12], and it is troublesome to set the privacy preference at each time.

Therefore, we need a system to recommend location privacy preferences for decision support when choosing a user's location information privacy preference and for the promotion of safe location information release. Recently, one such system was developed using the concept of item recommendation, which is used for online shopping. Item recommendation regards the combination of location and time as an item and whether or not to release location information as a rating of the item, and it predicts the rating of an unknown item using other users' data. Zhang [17] proposed a method of recommending by collaborative filtering.

We focus on the problems of existing location privacy preference recommendation methods and propose a recommendation method to solve these problems.

The contributions of this research are as follows.

1. **Clarifying the definition of location privacy preference:** In location privacy preservation, it is important to define where and when we want to preserve location privacy, which has various granularities. Although many location privacy protection methods have been proposed in the literature, none of them addresses the problem of how to set location privacy preference. Therefore, we generate recommendation models using various granularities for time information and compare their usefulness. From these results, we find

Table 1. A comparison of our method with related work [17].

	Method	How to preserve privacy
Related work [17]	Collaborative filtering (inaccurate for a large amount of data)	Add noise that is not strict mathematically
Our method	Matrix Factorization (accurate for a large amount of data)	Add noise based on local differential privacy

the best granularity that will produce trade-offs between the density of spatial data for the recommendation and the consideration of time.

2. **Applying matrix facrorization to location privacy preference recommendation:** Because location privacy preferences are very sensitive, the system must recommend them accurately. Collaborative filtering, which was used in the method by Zhang [17], experiences problems that when the number of users and products increase; accurate prediction cannot be achieved, and only the nature of either the user or product can be considered well. Therefore, we propose a method to improve by utilizing matrix factorization. As a result of experiments, we confirm that we can predict accurate evaluation values with a probability of 90% for large amount of data.
3. **Recommendation with local differential privacy:** Matrix factorization is involved in privacy risk, because each user needs to send their data to the recommendation system [2,6]. Location privacy preferences encompass location information of the users at a certain times and whether the information is sensitive for him/her. However, a location privacy preference recommendation that achieves highly accurate privacy protection has not been proposed so far. Actually, the method by Zhang [17] did not preserve privacy in a strict mathematical sense. Therefore, we propose a recommendation method that realizes it with local differential privacy, which refers to the method by Shin [13]. A comparison of our method with related work is shown in Table 1. We confirm that our method maintain precision that is the same as that achieved a method without privacy protection.
4. **Generating a location information privacy preference dataset:** A challenge in experiments for testing the performance of our methods is that no appropriate location privacy preference dataset is available in literature. The only available real-world dataset of location privacy preference [11] has few users. Such data is not suitable for the evaluation of the method using matrix factorization [10]. In addition, we need bulk data in the evaluation because there are many users of recommendation in the real world. Therefore, we created an artificial dataset that combines such the location privacy preference dataset and a trajectory dataset with a large number of users.

This paper is organized as follows: We describe the knowledge necessary for realizing our goal in Sect. 2. Then, we describe our method in Sect. 3 and evaluate and discuss about our method in Sect. 4.

2 Preliminaries

2.1 Matrix Factorization

Matrix factorization is one of the most popular methods used for item recommendation, which predicts the ratings of unknown items. This is an extension of collaborative filtering to improve the accuracy for the large amount of data by dimensionality reduction.

We consider the situation in which m users rate any item in n items. We express each user's rating of each item by $\mathcal{M} \subset \{1, \dots, m\} \times \{1, \dots, n\}$, and the number of ratings as $M = |\mathcal{M}|$, for the user i 's rating of item j . Matrix factorization predicts the ratings of unknown items given $\{r_{ij} : (i, j) \in \mathcal{M}\}$. To make a prediction, we consider a ratings matrix $R = m \times n$, a user matrix $U = d \times m$, and an item matrix $V = d \times n$. The matrices satisfy the formula: $R \approx U^T V$.

In matrix factorization, the user i 's element, i.e. the i -th column of U , is expressed by $u_i \in \mathbb{R}^d$, $1 \leq i \leq m$, and the item j 's element, i.e. the j -th column of V , is expressed by $v_j \in \mathbb{R}^d$, $1 \leq j \leq n$, which are learned from known ratings. The user i 's rating of item j is obtained by the inner product of u_i^T and v_j .

In learning, we obtain the matrices U and V , which minimize the following:

$$\frac{1}{M} \sum_{(i,j) \in \mathcal{M}} (r_{ij} - u_i^T v_j)^2 + \lambda_u \sum_{i=1}^m \|u_i\|^2 + \lambda_v \sum_{j=1}^n \|v_j\|^2 \quad (1)$$

λ_u and λ_v are positive variables for regularization.

U and V are obtained by updating using the following formulae:

$$u_i^t = u_i^{t-1} - \gamma_t \cdot \{\nabla_{u_i} \phi(U^{t-1}, V^{t-1}) + 2\lambda_u U_i^{t-1}\} \quad (2)$$

$$v_j^t = v_j^{t-1} - \gamma_t \cdot \{\nabla_{v_j} \phi(U^{t-1}, V^{t-1}) + 2\lambda_v V_j^{t-1}\} \quad (3)$$

γ_t is the learning rate at the t th iteration, and $\nabla_{u_i} \phi(U, V)$ and $\nabla_{v_j} \phi(U, V)$ are the gradients of u_i and v_j . They are obtained from derivative of (1) and expressed by the followings:

$$\nabla_{u_i} \phi(U, V) = -\frac{2}{M} \sum_{j:(i,j) \in \mathcal{M}} v_j (r_{ij} - u_i^T v_j) \quad (4)$$

$$\nabla_{v_j} \phi(U, V) = -\frac{2}{M} \sum_{i:(i,j) \in \mathcal{M}} u_i (r_{ij} - u_i^T v_j) \quad (5)$$

We predict the ratings of the unknown items by calculating U and V by these formulae.

2.2 Local Differential Privacy

We use local differential privacy to expand the matrix factorization into a form that satisfies privacy preservation. This approach is an extension of differential privacy [4], in which a trusted server adds noise to the data collected from the users. However, we assume that the server can not be trusted and use local differential privacy, in which users add noise to the data before sending the data to the server.

The idea behind local differential privacy is that for a certain user, regardless of whether or not the user has certain data, the statistical result should not change. The definition is given below:

Definition 1 (Local differential privacy). *We take $x \in N, x' \in N$. A mechanism M satisfies ϵ -local differential privacy if M satisfies the following:*

$$Pr[M(x) \in S] \leq \exp(\epsilon)Pr[M(x') \in S]$$

$\forall S \subseteq \text{Range}(M)$ is any output that M may generate. A randomized response [15] is used to realize local differential privacy, which decides the value to output based on the specified probability when inputting a certain value. Each user can add noise to the data according to their own privacy awareness, since he/she can decide the probability.

2.3 Definition of the Location Privacy Preference

The location privacy preference is defined by the following:

Definition 2 (Location privacy preference). *The location privacy preference $p_u(t, l)$, in which the user u wants to hide location information at time t in location l , is expressed by the following:*

$$p_u(t, l) = \begin{cases} 1 & (\text{Positive}) \\ 0 & (\text{Negative}) \end{cases}$$

1 (Positive) means that he/she can publish location information, and 0 (Negative) means that he/she does not want to publish location information.

The time t is expressed as a slot of time divided by a certain standard, and the division method varies depending on the reference time. Additionally, the location l is represented by a combination of geographic information and a category of place or either of these. Geographical information represents the latitude and longitude or certain fixed areas, and the category represents the property of a building located in that place such as a restaurant or a school.

There are various granularity regarding how to represent this information as mentioned in Sect. 1. As the granularity of information changes, the number of items in the recommendation and the degree of consideration of the nature of the time/location change, which influence the utility of the recommendation. However, it is not clear what kind of granularity is the best. Therefore, we confirm the best granularity, that is, the definition of best location information privacy preferences.

3 Recommendation Method

3.1 Framework

We propose a location privacy preference recommendation method that preserves privacy. When a user enters location privacy preferences for a certain number of places and time combinations, the method outputs location privacy preferences for a combination of unknown places and times. We assume the recommendation system exists on an untrusted server. We also assume an attacker who tries to extract the location and time of users' visit and their rating based on the output of the system. Our method aims for compatibility between high availability as a recommendation scheme and privacy protection. We realize the former by matrix factorization and the latter by local differential privacy.

First, we show a rough flow for the recommendation of the location privacy preference using the normal matrix factorization below:

1. The recommendation system sends the user matrix U and item matrix V to the user.
2. The gradients are calculated by using the user's data and step 1 and sent to the system.
3. The system updates U and V by the calculated gradients.

This operation is performed a number of times, and there is a risk that the user's data is leaked to the attacker. Therefore, we add noise to the data in step 2 above to avoid such risks. When the user calculate the gradients to update U and V , noise that satisfies local differential privacy is added to the information regarding "when and where the user visits" and "whether he/she wants to publish the location information." The gradients calculated based on the noise-added data are sent to the recommendation system We show the overview of our method in Fig. 1.

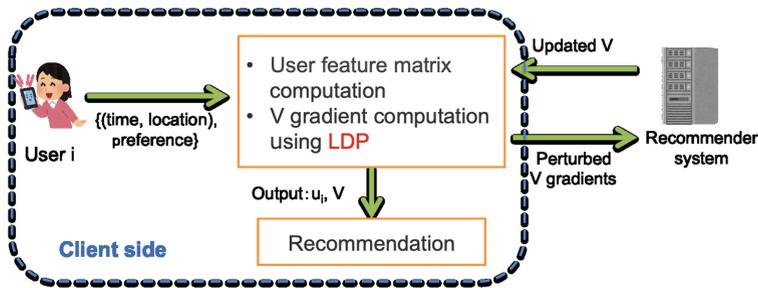


Fig. 1. Overview of our method.

3.2 Addition of Noise

We preserve privacy for the information, that is, when and where the user visits and whether he/she wants to publish their location information. In privacy protection process, we refer the method by Shin [13].

First, we will describe how to add noise to the information regarding the time and location of the user's visits. Let y_{ij} be a value indicating whether or not user i has visited place j , which is 1 if he/she has visited the place and 0 otherwise. The following equation holds: $\sum_{(i,j) \in \mathcal{M}} (r_{ij} - u_i^T v_j)^2 = \sum_{i=1}^n \sum_{j=1}^m y_{ij} (r_{ij} - u_i^T v_j)^2$. Therefore, Eq. (5) can be transformed as follows: $\nabla_{v_j} \phi(U, V) = -\frac{2}{n} \sum_{i:(i,j) \in \mathcal{M}} y_{ij} u_i (r_{ij} - u_i^T v_j)$. To protect information regarding the time and location of the user's visits from privacy attacks, we should add noise to a vector $Y_i = (y_{ij})_{1 \leq j \leq m}$. We use a randomized response, and the value y_{ij}^* is obtained by adding noise to y_{ij} as follows.

$$y_{ij}^* = \begin{cases} 0, & \text{with probability } p/2 \\ 1, & \text{with probability } p/2 \\ y_{ij}, & \text{with probability } 1 - p \end{cases}$$

Next, we describe a method of privacy protection for information on whether to disclose location information for a certain place and time combination. We add noise η_{ijl} , which is based on a Laplace distribution, to the value $g_{ij} = (g_{ijl})_{1 \leq l \leq d} = -2u_i (r_{ij} - u_i^T v_j)$. The noise-added value g_{ijl}^* is expressed as follows: $g_{ijl}^* = g_{ijl} + \eta_{ijl}$.

Each user adds noise to his/her own data in the above way, and the noise-added gradients, $\{(y_{ij}^* g_{ij1}^*, \dots, g_{ijd}^*) : j = 1, \dots, m\}$, are sent to the server. By repeating the operation k times, updating the value of the matrix using the slope calculated using the data with noise added, we find the matrix for predicting the evaluation value.

4 Evaluation

4.1 Overview

In this section, we describe the evaluation indices and points of view to consider when verifying the utilities of our method.

We evaluate the approximation between the true ratings value. We describe the details of these metrics in Sect. 4.3.

In the evaluations, we compare the utilities of the recommendation methods using normal matrix factorization and local differential privacy. In addition, we evaluate the method from the following three viewpoints:

1. What is the best location privacy preference definition?
2. How much impact does changes in the privacy preservation level make?
3. What impact does changes in the number of unknown evaluation values make?

More detailed results of the evaluation can be found in [1].

Table 2. A measure representing true or false values of the result.

		True rating	
		Positive	Negative
Predicted rating	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

4.2 Dataset

In the evaluation, we use artificial data combining the location privacy preference dataset and the position information dataset.

For the location privacy preference dataset, we use LocShare acquired from the data archive CRAWDAD [11]. This dataset was obtained from 20 users in London and St. Andrews over one week from April 23 to 29 in 2011, with privacy preference data for 413 places. This dataset has few users, so it is not suitable for the evaluation of our method using matrix factorization [10]. In addition, we need bulk data in the evaluation because there are many users of recommendation in the real world.

Therefore, we generated an artificial dataset by combining the location privacy preference dataset with the trajectory dataset Gowalla, which was acquired from the location information SNS in the U.S. This dataset includes check-in histories of various places from 319,063 users collected from November 2010 to June 2011. The total number of check-ins is 36,001,959, and the number of checked-in places is 2,844,145.

4.3 Metrics

We describe the metrics for verifying the utility of our method.

We measure how accurately the recommendation can predict the ratings. The predicted value in the recommendation can be classified based on the true evaluation value as in Table 2. For example, TP (True Positive) is the number of data that are truly Positive (can be released) and whose predicted values are also Positive. We calculate the number of TP, FP, TN, and FN results from the prediction result and measure the following two indices.

- False Positive Rate: The false positive rate is the percentage of false positives predicted relative to the number of negatives.

$$FPR = \frac{FP}{TN + FP}$$

This metric is an index for verifying whether the location information that the user wants to disclose is not erroneously disclosed. The lower the false positive rate, the higher the accuracy of the privacy protection.

- Recall: Recall is the proportion of data predicted to be positive out of the data that are actually positive.

$$Recall = \frac{TP}{TP + FN}$$

Recall is an index for verifying whether the location information that the user can publish is predicted to be positive, since if the released location information decreases, the benefit decreases. The higher the recall value is, the higher the utility of the recommendation.

4.4 Evaluation Process

We evaluate our method from the three viewpoints mentioned in Sect. 4.1, and we adopt each of the following methods.

1. We used 10-fold cross validation, that divides the users into training data and test data, in which 90% of the users are regarded as training data and 10% of the users are regarded as test data.
2. Among the user’s data included in the test data, we regard the known evaluation value as unknown according to the values of the *UnknownRate* mentioned later.
3. We predict the evaluation value by using the test data and the training data which have undergone conversion processing and calculate the metrics.
4. We repeat the above process 100 times and verify the average of the evaluation indices.

In the evaluation, we change the value of one of the following parameters: *time*, ϵ , *UnknownRate*. *time* is the length of the standard when dividing time into multiple slots. ϵ is privacy protection level when using local differential privacy. *UnknownRate* is the ratio of what is regarded as unknown.

4.5 Results

We describe the results of experiments to confirm the utility of the recommendation method.

The Best Location Privacy Preference Definition: When defining the location privacy preference, the best definition regarding the granularity of the information is not yet clear. In this experiment, we examine the influence of changing the granularity of time on the utility. We change the criterion for dividing time into multiple slots as follows: *time* = 2, 3, 4, 6, 8, 12, the other parameters are set as follows: $\epsilon = 0.01$ and the *UnknownRate* = 0.1. The results are shown in Fig. 2.

From these results, we confirm that the precision drops when the granularity is small, that is, when the criterion time is short. This is because the coarser the definition of the granularity is, the smaller the number of goods in the recommendation, and the matrix used for prediction becomes dense. On the other

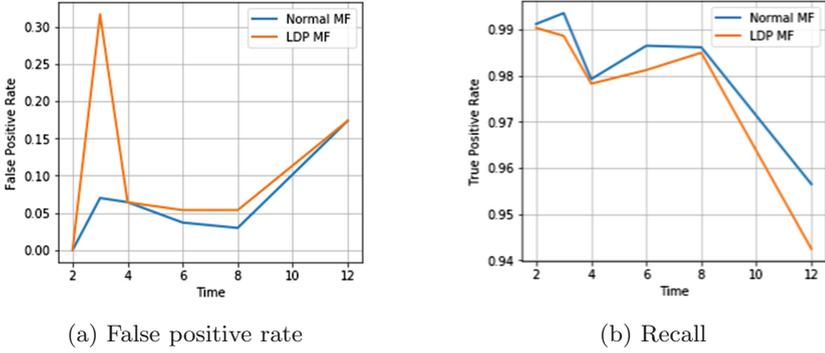


Fig. 2. Results when the time granularity is changed.

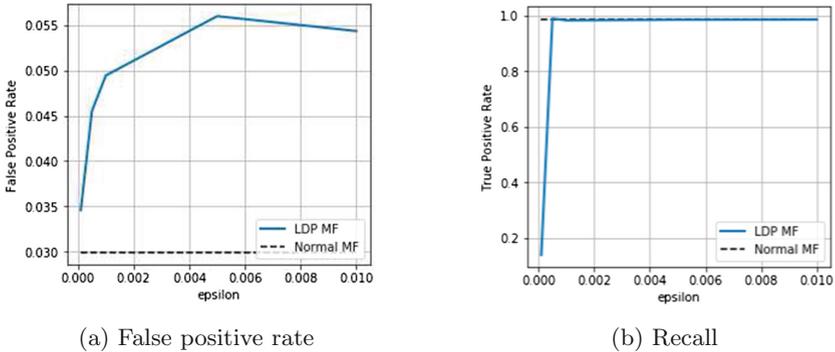


Fig. 3. Results when the ϵ is changed.

hand, however, we confirm that the accuracy drops even if the granularity is too large. Therefore, in defining the location privacy preference, we should choose the criterion with the highest utility. In this evaluation, the best criterion is 8 h.

Impact of Changes in the Privacy Preservation Level: The strength of the privacy protection can be adjusted with the value ϵ in local differential privacy. The smaller the value of ϵ , the more privacy is protected. On the other hand, there is a risk that the utility of the recommendation decreases as the added noise become large. Therefore, we examine the influence of changing the privacy protection level on the utility.

In this evaluation, we change the privacy protection level as follows: $\epsilon = 0.0001, 0.0003, 0.001, 0.005, 0.01$, the other parameters are set as follows: $time = 6$ and the $UnknownRate = 0.1$. The results are shown in Fig. 3.

From the results, we confirm that a normal recommendation is more useful in general, and as the value of ϵ increases, the usefulness increases. On the other hand, however, the change in the usefulness due to the change in the value of ϵ is small for $\epsilon = 0.003$. Larger values do not have a significant effect on the usefulness. We should select the maximum parameter that can maintain

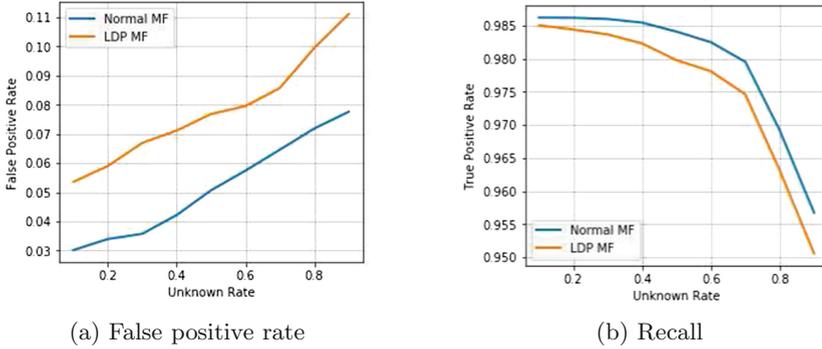


Fig. 4. Results when the *Unknown Rate* is changed.

prediction accuracy, since a stronger privacy protection level is achieved for a smaller value of ϵ . Therefore, in this evaluation, the best privacy protection level is $\epsilon = 0.001$.

Impact of Changes in the Number of Unknown Evaluation Values:

We verify how much each user should know his/her privacy preference for an accurate prediction. In the evaluation, we regard a certain number of evaluated data as unevaluated in generating the model and verify the influence of the number of unknown evaluation values on the utility. We change the parameter for the percentage of unevaluated data as follows: *UnknownRate* = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. The smaller the *UnknownRate*, the higher the number of evaluation values is. The other parameters are set as follows: *time* = 6 and $\epsilon = 0.01$. The results are shown in Figs. 4.

From these results, we confirm that the utility tends to decrease as the number of unevaluated data values increases. This is because the accuracy of the recommendation will be reduced if the training dataset is small. From these results, ideally, the user should know the location privacy preference nearly as much as possible, for about 70% of all products.

In all the evaluation, we compare the utility of the models using normal matrix factorization and local differential privacy. Since appreciable differences were not observed in devising the parameters, we confirm that the recommendation method can maintain an accuracy comparable to that of normal matrix factorization.

5 Conclusion

We propose a location privacy preference recommendation system that uses matrix factorization and achieves privacy protection by local differential privacy. We also confirm how to determine the best location privacy preference definition.

We evaluate our method using an artificial dataset from a location privacy preference dataset and a trajectory dataset. From its results, we confirm that our method can maintain the utility at a level that is the same as a method without privacy preservation. In addition, we confirm the best parameters, the granularity of the location privacy preference definition, the privacy protection level, and the number of rated items.

References

1. Asada, M., Yoshikawa, M., Cao, Y.: When and where do you want to hide? Recommendation of location privacy preferences with local differential privacy. arXiv preprint [arXiv:1904.10578](https://arxiv.org/abs/1904.10578) (2019)
2. Calandrino, J.A., Kilzer, A., Narayanan, A., Felten, E.W., Shmatikov, V.: “You might also like:” privacy risks of collaborative filtering. In: 2011 IEEE Symposium on Security and Privacy (SP), pp. 231–246. IEEE (2011)
3. De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: the privacy bounds of human mobility. *Sci. Rep.* **3**, 1376 (2013)
4. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79228-4_1
5. Fawaz, K., Shin, K.G.: Location privacy protection for smartphone users. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 239–250. ACM (2014)
6. Frey, D., Guerraoui, R., Kermarrec, A.M., Rault, A.: Collaborative filtering under a sybil attack: analysis of a privacy threat. In: Proceedings of the Eighth European Workshop on System Security, p. 5. ACM (2015)
7. Huo, Z., Meng, X., Hu, H., Huang, Y.: You can walk alone: trajectory privacy-preserving through significant stays protection. In: Lee, S., Peng, Z., Zhou, X., Moon, Y.-S., Unland, R., Yoo, J. (eds.) DASFAA 2012. LNCS, vol. 7238, pp. 351–366. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29038-1_26
8. Jiang, K., Shao, D., Bressan, S., Kister, T., Tan, K.L.: Publishing trajectories with differential privacy guarantees. In: SSDBM (2013)
9. Kanza, Y., Samet, H.: An online marketplace for geosocial data. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 10. ACM (2015)
10. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. *Computer* **8**, 30–37 (2009)
11. Parris, I., Abdesslem, F.B.: *Crowdad st_andrews/locshare* dataset (2011). https://crowdad.org/st_andrews/locshare/20111012/
12. Sadeh, N., et al.: Understanding and capturing people’s privacy policies in a mobile social networking application. *Pers. Ubiquitous Comput.* **13**(6), 401–412 (2009)
13. Shin, H., Kim, S., Shin, J., Xiao, X.: Privacy enhanced matrix factorization for recommendation with local differential privacy. *IEEE Trans. Knowl. Data Eng.* **30**, 1770–1782 (2018)
14. Sweeney, L.: k-anonymity: a model for protecting privacy. *Int. J. Uncertainty Fuzziness Knowl. Based Syst.* **10**(05), 557–570 (2002)
15. Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.* **60**(309), 63–69 (1965)

16. Wasef, A., Shen, X.S.: REP: location privacy for vanets using random encryption periods. *Mob. Netw. Appl.* **15**(1), 172–185 (2010)
17. Zhao, Y., Ye, J., Henderson, T.: Privacy-aware location privacy preference recommendations. In: Proceedings of the 11th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, MOBIQUITOUS 2014, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, pp. 120–129 (2014). <https://doi.org/10.4108/icst.mobiquitous.2014.258017>