How Private Can I Be Among Public Users?

Elham Naghizade, James Bailey, Lars Kulik, Egemen Tanin
Computing and Information Systems Department
The University of Melbourne
Melbourne, Australia
{enaghi,baileyj,lkulik,etanin}@unimelb.edu.au

ABSTRACT

People are increasingly volunteering personal data. Services based on this data rely on a high number of participants and high data quality. Personal data is often seen as private and individuals are more likely to provide such data if they can choose its granularity, e.g., instead of an exact value, they may provide a range. Focusing on spatial crowdsourced data, this work aims to determine whether the common method of coarsening location data of privacy-conscious individuals is an effective approach if fine-grained location data has also been submitted by privacy-apathetic users. We propose a novel inference attack to refine the location of privacy-conscious individuals. Our experiments suggest that even with a dataset that is mostly populated with privacy-conscious users, our technique succeeds with high precision and recall.

Author Keywords

Privacy; Crowdsourced Trajectory Database; Matrix Factorization

ACM Classification Keywords

H.2.8 Database Management: Database Applications—*Data mining*

INTRODUCTION

The uptake in sensor-enabled smartphones and wearable devices enables individuals to monitor not only their environment, but also themselves on a 24/7 basis. Such rich datasets facilitate two types of applications: (i) personal analytics that informs people about their daily fitness, health and life choices, and (ii) public and social applications that benefit from people-centric sensing [3]. Volunteering/sharing their detailed data, individuals can monitor their health condition, compare themselves to others, estimate their exposure to pollution and learn about the traffic and road condition in a certain area. Two applications in that area are BikeNet [7] and Biketastic [12] that enable riders to monitor their personal progress and – using other riders' data – to avoid areas with high noise or pollution.

However, individuals have different perceptions of privacy, which has a direct impact on their data volunteering behaviour. Studies in the literature have stressed the importance of accommodating users' privacy preferences both as a means of encouraging them to share their data and as an effective privacy-preserving scheme [5].

A service benefits from fine-grained data and relying on only privacy-conscious users would adversely affect the quality of data analytics and services. On the other hand, even privacy-conscious individuals may share their data in return for receiving an improved service. As a result, some privacy-apathetic users may contribute detailed data, while privacy-conscious users may prefer to provide a range instead of a precise value or a cloaked region instead of an actual position. Our key question is *to what extent volunteering coarse-grained data can actually guarantee the desired level of privacy for privacy-conscious users in such scenario.*

With an increase in the amount of personal data being collected and analysed, and its privacy implications, many recent studies propose methods for preserving individuals' privacy such as k-anonymity, l-diversity and obfuscation techniques [9, 13]. Variations of these techniques enable users to opt for their desired level of privacy [10, 16]. Differential privacy [6] provides formal privacy guarantees by adding noise to the original data. However, the underlying assumption for current techniques is that the entries of the dataset, i.e., users' crowdsourced data, have the same granularity level.

This paper focuses on location data. Given a set of users who contribute their daily commute patterns with different resolutions respective to their privacy preference, we investigate if it is possible to use the more fine-grained trips of privacy-apathetic users in conjunction with the coarse-grained trips of the privacy-conscious users to refine the contributed data of the latter. Despite many efforts in the literature to address the privacy issues of sharing detailed location data, the privacy implications of having a multi-granular dataset has not been investigated. Furthermore, unlike various studies, we do not focus on identifying a participant but on whether or not the guaranteed level of privacy can be maintained.

One way of storing location data at different granularity levels is the use of a grid structure [10] since i) it provides a flexible as well as comprehensible means of facilitating users' preference specification ii) it is independent of the original trajectory and does not reflect any specific property of the data. We store the location data as a sequence of grid cell IDs.

We presume the adversary is any third party with access to the dataset but who exploits no other source of background knowledge, e.g., road network information. The adversary may extract multiple versions from the multi-granular grid-based dataset through generalization: A *complete* dataset that

UbiComp '15, September 7-11, 2015, Osaka, Japan.
Copyright 2015 © ACM 978-1-4503-3574-4/15/09...$15.00.
http://dx.doi.org/10.1145/2750858.2805836.

contains all users' location data at the lowest granularity, i.e., coarsest resolution, and several *incomplete* datasets with more fine-grained location data and in which the location data of privacy-conscious users for that specific level is not known.

A typical approach to infer the unknown values in incomplete datasets is employing matrix factorization (MF). MF is an unsupervised learning method that has been successfully utilized in the recommender systems to provide suggestions, e.g., movies, books, points of interest, etc., to users whose rating for a certain item is not known [8].

Our experiments show MF is largely unsuccessful at inferring the unknown values of privacy-conscious users because MF does not fully exploit the available information of privacy-apathetic users that can be derived from their fine-grained data by generalization. We propose an MF-based approach that uses the data available for both granularities for privacy-apathetic users: it learns a transition matrix that maps users' coarse-grained locations to their fine-grained locations. This transition matrix is then applied to the coarse-grained data of privacy-conscious users to predict their fine-grained location. This makes our method a supervised approach contrary to the unsupervised classical MF.

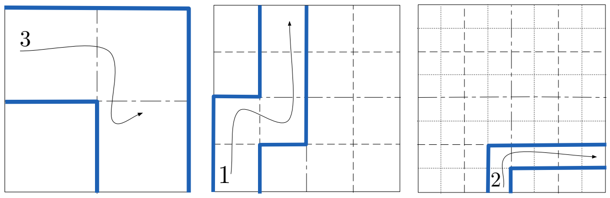
Our inference attack manages to refine the footprint of privacy-conscious users with a high precision and recall. We are not trying to discourage the practice of data volunteering but to highlight the shortcomings of current user-specified privacy settings and the risks of contributing data when privacy-apathetic users are involved.

PROBLEM STATEMENT

Let $D = \{T_1, T_2, ..., T_n\}$ be a set of trajectories, T_i, belonging to m users ($m \leq n$). Each user has a privacy profile as (u_i, l), where $l_{min} \leq l \leq l_{max}$ and l corresponds to user's predetermined granularity level - higher levels of granularity relate to less private users. Hence, G_l represents a 2^{2l}-cell grid that results from l consecutive decompositions of space into four quadrants. A multi-granular dataset, $D = \{S_1, S_2, ..., S_n\}$ is a set of sequences that maps any trip, T_i to a specific grid; S_i is the sequence of grid cells, the size of which is determined by l, that maps to the data points in T_i.

Figure 1 shows an example of three users traversing the same region. Assume users are given the option to choose an l between 1 to 3, with 1 relating to the coarsest resolution and 3 for the finest resolution required. Figure 1b, 1c, and 1a show the grid structure, i.e., a G_l, that matches their desired privacy settings, and Figure 1d provides a snapshot of their corresponding grid cell sequence in the multi-granular dataset.

Definition 1: **An l-mapped version of the multi-granular dataset**, $D_l = \{S_{1l}, S_{2l}, ..., S_{nl}\}$, is an adaptation of D where all the sequences correspond to the same level of granularity□. For all the users whose specified granularity level is greater than (or equal) to l, the grid cells are mapped to their coarser level (or remain untouched). For those who prefer a granularity level smaller than l (private people), the fine-grained mapping is as follows: the grid cells that have not been traveled (0 in their coarse-grained data) are assigned to zero and the remaining entries are marked as unknown. *Given any D_l, we aim to estimate the unknown values in D_l, i.e., the location of private users, e.g., u_1 and u_3 in Figure 1.*



(a) A private user. (b) A semi-private user. (c) A public user.

u_i	G_l	Sequence
1	16 cells	(0,0),(0,1),(1,1),(1,2),(1,3)
2	64 cells	(3,0),(3,1),(4,1),(5,1),(6,1),(7,1)
3	4 cels	(0,1),(1,1),(1,0)

(d) A multi-granular dataset of 3 users with different privacy preferences.

Figure 1: Applying a grid structure to accomodate participants' privacy specification.

$$\begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \end{array}\begin{array}{c} G_1 \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

(a) R_1 (coarse)

$$\begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \end{array}\begin{array}{c} G_2 \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & ? & ? & ? & ? & ? & ? & ? & ? & ? & ? & ? & ? \end{bmatrix} \end{array}$$

(b) R_2 (finer-grained)

Figure 2: A coarse-grained complete matrix vs a finer-grained matrix with missing values for a private user (user3). Columns correspond to the cell IDs of the grid.

Each D_l can be represented as an $n \times 4^l$ matrix, R_l. Figure 2 depicts two matrices, R_1 and R_2 that correspond to the sample dataset provided in Figure 1d. r_{ij} is set to 1 if the user i has passed the jth cell (based on z-curve indexing in this example) and if not, it is set to 0. Using such a representation, the aim of our inference attack is to predict the unknown values (question marks in Figure 2b) in D_l for each $l > l_{min}$. Conceptually, this process involves mapping the known location information of privacy-conscious users to its children in a hierarchical grid. For instance, for user 3 in Figure 1a the grid ID sequence is known for $l = 1$, and we wish to refine the trip by one level, i.e., $l^* = 2$ (we use the same notation throughout this paper to differentiate between coarse and fine granularity). Thus, for each cell in S_3, e.g., (0,1), we try to find the respective subcell(s) in G_2, $\{(0,2),(0,3),(1,2),(1,3)\}$ that is (are) most likely to have been traveled by user 3.

PROPOSED METHOD

Direct Factorization (DF)

One possible approach to infer unknown values in fine-grained matrices is to apply MF and decompose a matrix into two matrices. This approach has been successfully adopted in the field of recommender systems to provide users with suggestions that best suits them and hence, we use it as a baseline to evaluate the performance of our approach. MF uses the data of other users to discover the latent features that govern the interaction between users and items. For a matrix $R_{n \times m}$ that includes some unknown values, assume that k is the number of latent features. The aim of MF is to find two matrices $P_{n\times k}$ and $Q_{k\times m}$ that approximate R in the following way:

$$P \times Q = R' \approx R \quad where \quad r'_{ij} = \sum_{k=1}^{k} p_{ik} q_{kj}$$

The goal is to minimise the total error between the real known values and their respective predicted values, i.e., $\|R - P \times Q\|^2$. Further details on applications of matrix factorization in recommender systems can be found at [8]. We apply MF to the fine-grained matrix, i.e., R_{l^*}, in order to estimate the value of each cell for privacy-conscious users. In other words, we approximate P and Q using the available information of public users. The product of the obtained P and Q provides the estimate values for private users.

Granularity-based Factorization (GBF)

Direct factorization on the fine-grained data fails to use the complete coarse-grained data to improve its prediction. However, any coarse-grained matrix can be mapped to a fine-grained matrix using a transition matrix ($l^* > l$):

$$R_l \times \Phi = R_{l^*}$$

As a result, instead of factorizing R_{l^*} into any two matrices with an arbitrary k, we can decompose it to R_l and Φ. Note that when trying to minimise the estimation error, i.e., $\|R_{l^*} - R_l \times \Phi\|^2$, R_l remains unchanged and only Φ is modified.

Moreover, R_{l^*} can be divided to two parts; a complete part concerning individuals whose preferred level of granularity is equal or higher than l^*, $R^a_{l^*}$ and an incomplete part that involves users who have chosen lower levels of granularity, i.e., $R^c_{l^*}$. We now can learn a transition matrix, Φ, that best approximates $R^a_{l^*}$ using their coarse-grained matrix. This is done by a random initialization of Φ and then iteratively minimizing the distance $\|R^a_{l^*} - R^a_l \times \Phi\|$. Using Φ, we can apply it to the coarse matrix of privacy-conscious users to infer the unknown values in a finer-grained matrix of higher spatial resolution:

$$R^c_l \times \Phi \approx R^f_{l^*}$$

We use gradient descent to find the local minimum when total error is less than a predetermined threshold.

The main steps of our granularity-based factorization are depicted in Figure 3. We first train our model with the complete matrices of privacy-apathetic users, i.e., R^a_l and $R^a_{l^*}$, and learn Φ. We then apply the learnt Φ to the complete coarse matrix of private users to predict their unknown information (grey cells in the fine-grained matrix).

EXPERIMENTS

We used the Porto taxi trajectory dataset [1] that contains the GPS trajectories of 442 taxis for a complete year in Porto. We randomly sampled 10000 trips with an average length of $\approx 4km$ that reside within a ($\approx 10km \times 10km$) area in the business district of Porto. Since the trips of same taxi drivers may have large overlaps that may lead to biased results, we initially categorize the drivers as privacy-conscious and privacy-apathetic users and then select their trips to feed to our model.

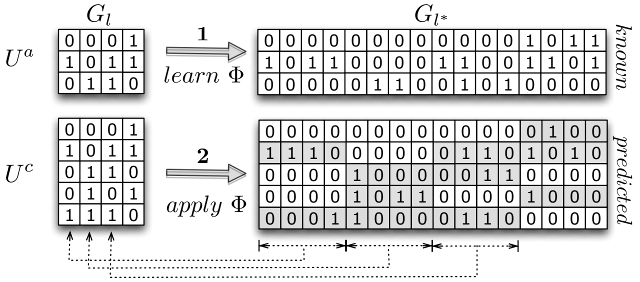


Figure 3: Major steps of GBF. Gray cells show the inferred values for the 1s in the coarse-grained matrix.

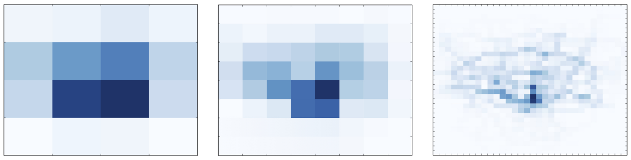
Resolution (l)	1	2	3	4	5
Area(km^2)	25	6.25	1.56	0.39	0.09
Occupancy rate	51%	18%	7.5%	3.4%	1.5%

Table 1: Summary of the trajectory dataset after being mapped to different granularity levels.

The GPS points of each trajectory are mapped to a sequence of grid cell IDs for each level of granularity. We vary l from 1 to 4 which reflects $5km \times 5km$ grid cells to $600m \times 600m$ for private users to demonstrate coarse spatial information since larger levels, i.e. smaller grid cells, are not private anymore. We aim to refine the private footprints up to $300m \times 300m$ grid cells that correspond to $l^* = 5$. Table 1 provides a summary of the dataset information and Figure 4 shows the density of grid cells for three different granularity levels.

As default setting for our experiments we assume that 90% are private users, i.e., the model is trained with 10% of the data. We set the learning rate, α to 0.0004 and the regularization parameter, β to 0.015. For DF, k is equal to the number of coarse grid cells (4^l). A five-fold cross validation was used for the provided results.

To evaluate the success of our inference attack, we used precision and recall. While recall reflects how much of the original trip has been retrieved, precision shows how much of the inferred trip is predicted correctly. More generally, we define the number of correctly predicted occupied cells as *True Positives (TP)* and the number of correctly predicted empty cells as *True Negatives (TN)*, the number of empty grid cells that have been falsely predicted to be occupied as *False Positives (FP)* and the number of grid cells that were actually occupied but predicted to be empty as *False Negatives (FN)*. We



(a) $l = 2$ (b) $l = 3$ (c) $l = 5$

Figure 4: The density of grid cells for different granularity levels within a $\approx 10km \times 10km$ area in Porto.

	Precision			Recall		
Private Users	90%	70%	50%	90%	70%	50%
DF	0.35	0.36	0.37	0.36	0.37	0.38
FBS	0.67	0.66	0.67	0.67	0.67	0.67
GBF	0.75	0.75	0.74	0.75	0.75	0.74

Table 2: Varying number of private users for $l = 2, l^* = 3$.

	$l = 3, l^* = 4$				$l = 4, l^* = 5$			
	P	R	F1	MAP	P	R	F1	MAP
FBS	0.64	0.65	0.77	0.71	0.60	0.62	0.72	0.66
GBF	0.75	0.75	0.84	0.82	0.64	0.65	0.71	0.67

Table 3: Varying resolution ($l^* - l = 1$) for 90% private users.

compute precision and recall as follows:

$$Recall = TP/(TP+FN), \quad Precision = TP/(TP+FP).$$

We also use the F-measure (F1) and mean average precision (MAP) metrics popular in information retrieval.

We implemented a technique that randomly selects one to four random subcell(s) for each known coarse-grained cell but its performance was not competitive and not reported. We also developed a frequency-based sampling technique (FBS) that uses the available fine-grained data of public people to retrieve the frequency of through each subcell and estimates the private user locations as the probability of passing them.

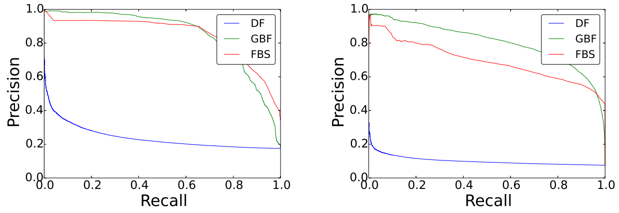
Table 2 and Figure 5 compares the performance of DF, FBS and GBF for varying numbers of private users and varying granularity levels, respectively. DF as an unsupervised approach is not successful in predicting the locations of private users, especially when the majority of users are private. GBF's performance remains largely the same, even with an increase in the number of private people,. When l is sufficiently small, FBS has a comparable performance to GBF (Figure 5a). However, for a denser resolution, e.g., $l = 2,3$, GBF outperforms FBS because larger ls results in an increase in the number of latent features, i.e., k being considered. However, if k is too large ($l = 4$), GBF's performance declines to levels comparable with FBS. Table 4 shows that GBF succeeds to refine major parts of private trips when $l^* - l = 2$, i.e., spatial information is 16 times more refined.

RELATED WORK

Sweeney et al. [13] laid the foundation of privacy-preserving data publication by proposing k-anonymity in 2002. l-diversity built on this idea and made any sensitive attribute indistinguishable from $l-1$ other attributes in [9]. Users, however, may differ in the way they perceive privacy and sensitive contexts [4, 14]. As a result, [16] proposed an adaptive method to guarantee for each user their required level of privacy while minimizing the distortion of the original data.

	$l = 1, l^* = 3$				$l = 2, l^* = 4$			
	P	R	F1	MAP	P	R	F1	MAP
FBS	0.56	0.57	0.73	0.60	0.44	0.46	0.58	0.48
GBF	0.59	0.59	0.73	0.62	0.54	0.55	0.66	0.59

Table 4: Varying resolution ($l^* - l = 2$) for 90% private users.



(a) $l = 1, l^* = 2$

(b) $l = 2, l^* = 3$

Figure 5: Precision-recall curves for varying spatial resolutions when 90% users are private.

Many adaptive obfuscation techniques were proposed that intend for protecting the sensitive data rather than users' identity while accommodating users' privacy preferences [11, 15, 17]. New Casper [10] proposes a grid structure to obfuscate location information according to users' privacy profiles. We applied a similar structure as our privacy preserving foundation. Also in [2], the reported area is enlarged as long as it meets the privacy specification of users. This enlargement is evaluated against the linkability of location information and what can be inferred from user's movement history.

The above-mentioned studies focus on the data provided by a privacy-conscious user, e.g., what has been previously volunteered, what has been shared with other applications, how frequently data is reported, etc., to preserve privacy. The privacy implications of available fine-grained data in the database have not been studied. We apply a modified MF since it has been successfully utilized to estimate missing values, e.g., for recommender systems [8]. [18] provides an example of successfully deploying MF techniques to decompose a user-location-activity tensor ausing an external source of data to profile users. Our approach, however, modifies MF to make use of available coarse-grained information.

CONCLUSION

Our method, called GBF, can refine the coarse-grained data of privacy-conscious users using the more fine-grained data of privacy-apathetic users. Our inference attack manages to refine private data with high resolution and recall, even when the majority of users are privacy-conscious without any background knowledge of the road network. This highlights the vulnerability of current approaches that attempt to provide user-specified privacy, urging the privacy community to find a new solution for this vulnerability.

In this work, we only focus on the location of the users, regardless of time. In future, we will explore temporal information to improve the efficacy of our inference attack. In addition, we will investigate how to generalise our attack to incorporate general hierarchical data and handle any probabilistic or noisy hierarchical mappings.

ACKNOWLEDGMENT

This research was partially supported under Australian Research Council's Discovery Projects funding scheme (project number DP110100757) and partially supported by the NICTA Victoria Research Laboratory.

REFERENCES

1. Porto taxi trajectory dataset. `http://www.geolink.pt/ecmlpkdd2015-challenge`, Accessed: 2015-06-22.
2. Agir, B., Papaioannou, T., Narendula, R., Aberer, K., and Hubaux, J.-P. User-side adaptive protection of location privacy in participatory sensing. *GeoInformatica 18*, 1 (2014), 165–191.
3. Campbell, A., Eisenman, S., Lane, N., Miluzzo, E., Peterson, R., Lu, H., Zheng, X., Musolesi, M., Fodor, K., and Ahn, G.-S. The rise of people-centric sensing. *Internet Computing, IEEE 12*, 4 (July 2008), 12–21.
4. Carrascal, J. P., Riederer, C., Erramilli, V., Cherubini, M., and de Oliveira, R. Your browsing behavior for a big mac: Economics of personal information online. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 (2013), 189–200.
5. Christin, D. Impenetrable obscurity vs. informed decisions: privacy solutions for participatory sensing. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on* (March 2010), 847–848.
6. Dwork, C. Differential privacy. In *Encyclopedia of Cryptography and Security*, H. van Tilborg and S. Jajodia, Eds. Springer US, 2011, 338–340.
7. Eisenman, S. B., Miluzzo, E., Lane, N. D., Peterson, R. A., Ahn, G.-S., and Campbell, A. T. Bikenet: A mobile sensing system for cyclist experience mapping. *ACM Trans. Sen. Netw. 6*, 1 (Jan. 2010), 6:1–6:39.
8. Koren, Y., Bell, R., and Volinsky, C. Matrix factorization techniques for recommender systems. *Computer*, 8 (2009), 30.
9. Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data 1*, 1 (Mar. 2007).
10. Mokbel, M. F., Chow, C.-Y., and Aref, W. G. The new casper: Query processing for location services without compromising privacy. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, VLDB '06, VLDB Endowment (2006), 763–774.
11. Rahman, F., Hoque, M. E., Kawsar, F. A., and Ahamed, S. I. User privacy protection in pervasive social networking applications using pco. *International Journal of Social Computing and Cyber-Physical Systems*, 3 (2012).
12. Reddy, S., Shilton, K., Denisov, G., Cenizal, C., Estrin, D., and Srivastava, M. Biketastic: Sensing and mapping for better biking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, ACM (New York, NY, USA, 2010), 1817–1820.
13. Sweeney, L. K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*, 05 (2002), 557–570.
14. Toch, E., Cranshaw, J., Drielsma, P. H., Tsai, J. Y., Kelley, P. G., Springfield, J., Cranor, L., Hong, J., and Sadeh, N. Empirical models of privacy in location sharing. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, ACM (New York, NY, USA, 2010), 129–138.
15. Wishart, R., Henricksen, K., and Indulska, J. Context privacy and obfuscation supported by dynamic context source discovery and processing in a context management system. In *Ubiquitous Intelligence and Computing*, vol. 4611. Springer Berlin Heidelberg, 2007, 929–940.
16. Xiao, X., and Tao, Y. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, ACM (New York, NY, USA, 2006), 229–240.
17. Xu, T., and Cai, Y. Feeling-based location privacy protection for location-based services. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, CCS '09 (2009), 348–357.
18. Zheng, V. W., Cao, B., Zheng, Y., Xie, X., and Yang, Q. Collaborative filtering meets mobile recommendation: A user-centered approach. In *AAAI 2010* (July 2010).