# Privacy-preserving Cross-domain Location Recommendation

CHEN GAO, CHAO HUANG, YUE YU, HUANDONG WANG, YONG LI, and DEPENG JIN, Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, China

Cross-domain recommendation is a typical solution for data sparsity and cold start issue in thefi eld of location recommendation. Specifically, data of an *auxiliary* domain is leveraged to improve the recommendation of the *target* domain. There is a typical scenario that two interaction domains (location based check-in service, for example) combine data to perform the cross-domain location recommendation task. Existing approaches are based on the assumption that the interaction data from the auxiliary domain can be directly shared across domains. However, such an assumption is not reasonable, since in the real world those domains may be operated by different companies. Therefore, directly sharing raw data may violate business privacy policy and increase the risk of privacy leakage since the user-location interaction records are very sensitive.

In this paper, we propose a framework named privacy-preserving cross-domain location recommendation which works in two stages. First, for the interaction data from the auxiliary domain, we adopt a differential privacy based protection mechanism to hide the real locations of each user to meet the criterion of differential privacy. Then we share the protected user-location interaction to the target domain. Second, we develop a new method of Confidence-aware Collective Matrix Factorization (CCMF) to effectively exploit the transferred interaction data. To verify its efficacy, we collect two real-world datasets suitable for the task. Extensive experiments demonstrate that our proposed framework achieves the best performance compared with the state-of-the-art baseline methods. We further demonstrate that our method can alleviate the data sparsity issue significantly while protecting users' location privacy.

CCS Concepts: • **Information systems → Recommender systems**; **Location based services**; • **Security and privacy → Privacy-preserving protocols**;

Additional Key Words and Phrases: Cross-domain Location Recommendation; Differential Privacy; Matrix Factorization

**ACM Reference Format:**
Chen Gao, Chao Huang, Yue Yu, Huandong Wang, Yong Li, and Depeng Jin. 2019. Privacy-preserving Cross-domain Location Recommendation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 11 (March 2019), 21 pages. https://doi.org/10.1145/3314398

## 1 INTRODUCTION

Location recommendation has wide applications in various services, such as GPS navigation, location-aware social network, etc [15, 25, 50]. The definitions of location in existing researches are various, including an individual POI (Point-of-Interest) [28, 29, 57, 58], venue [36, 56] or even POI sequence [12], among which the individual POI is the most widely used one. To achieve the goal of recommending a new location (*i.e.* individual POI) to a user that has never visited before, most of existing recommender systems adopt collaborative-filtering (CF) [41], *i.e.*,

Authors' address: Chen Gao, gc16@mails.tsinghua.edu.cn; Chao Huang, c-huang15@mails.tsinghua.edu.cn; Yue Yu, yue-yu15@mails.tsinghua.edu.cn; Huandong Wang, whd14@mails.tsinghua.edu.cn; Yong Li, liyong07@tsinghua.edu.cn; Depeng Jin, jindp@tsinghua.edu.cn, Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China.


2474-9567/2019/3-ART11 $15.00
https://doi.org/10.1145/3314398

utilizing collected records of users' historically visited locations (*a.k.a.* interaction data), to learn users' interests and locations' features. However, such method usually suffers from the data sparsity issue [27]. Specifically, it cannot precisely infer a user's interests if she visited only few locations. It is the same with inferring locations' features when a location has been visited by just a few users. Sometimes, it will be even worse if there exist cold-start users or locations [59]. To address such data sparsity issue, a typical solution is borrowing data from auxiliary domain (collected from another LBSN service provider, for example) to assist in learning users' interests and locations' features, and improve the recommendation accuracy. Such solution is named cross-domain location recommendation [10, 22].

Cross-domain location recommendation can be divided into three components, data sharing, model building, and recommendation presenting. However, there exist severe risks of leaking user privacy in such systems, especially when sharing data from the auxiliary domain to the target domain. Specifically, in most scenarios, these two domains are typically operated by two companies. Sharing data cross companies becomes a great concern recently, especially after the enactment of GDPR [1] (short for General Data Protection Regulation) in 2018. Furthermore, it has been demonstrated that users' identifiers can be de-anonymized with high probability from their trajectory data in recent studies [39, 40, 48, 49].

Therefore, to perform the task of utilizing interaction data across domains for location recommendation, approaches that avoid directly sharing data are more practical. However, on one hand, it has not been studied yet in the researchfi eld of cross-domain recommendation, to the best of our knowledge. On the other hand, existing researches on privacy-preserving recommendation are only suitable for single-domain scenario[2], focusing on preserving privacy in *collecting data from a single domain*, *model training* or *presenting recommendation results to users*. Actually, in this paper, we are focusing on potential risk of leaking privacy in *data sharing*, which is a specific component in cross-domain scenario. To address it, we propose a novel two-stage framework which applies privacy-preserving mechanism in thefi rst component (*i.e.*, data sharing). Specifically, at thefi rst stage, we apply a widely used privacy criterion, differential privacy [13], to our task and design a semantic-aware obfuscation method. Specifically, this method adds noise to raw data based on differential privacy, in order to hide the real locations for the auxiliary interaction data during data sharing. In other words, we obtain obfuscated interaction data that is protected under the differential privacy criterion. Then this protected data can be shared to the target domain. At the second stage, we propose a novel matrix factorization method whichfi rst calculates a confidence matrix with transferred auxiliary interaction matrix, and then combines this matrix with two interaction matrices of two domains to perform a collective matrix factorization for recommendation. This method of confidence-aware collective matrix factorization not only effectively extracts signal from transferred obfuscated interaction data, but also balances the influence of the auxiliary and target domains.

To summarize, the main contributions of this paper are as follows.

- We present a new framework for privacy-preserving cross-domain location recommendation. In this framework, interaction data from the auxiliary domain is protected via our obfuscation mechanisms, which meets the criterion of differential privacy, before shared to the target domain. With this criterion, the knowledge the attacker gains after obtaining the transferred data is bounded, no matter what prior knowledge he has.
- We propose a novel recommendation solution named CCMF to resolve the key challenges of leveraging the obfuscated interaction data. The design distills useful signals from interaction data, and appropriately combines them with the interaction data of the target domain.

[1] https://eudgpr.org

[2] Note that in single-domain recommendation, there are also three components: data collection (rather than data sharing), model training and results presenting. Therefore, risk of privacy leakage in data sharing has never been concerned in privacy-preserving single-domain recommendation methods.

- Experimental results on two real-world datasets demonstrate that our method can help improve recommendation in the target domain by 2.05%-111.76% while protecting the interaction data in the auxiliary domain. Further studies verify the efficacy of this method on recommendation task for sparse locations.

The remainder of this paper is as follows. Wefi rst discuss the related works in Section 2. We formulate the research problem and present the system overview in Section 3. We then elaborate our proposed method in Section 4 and conduct experiments on Section 5. Lastly, we further conclude this work and discuss the future work in Section 6.

## 2 RELATED WORK

**Cross-domain Location Recommendation.** Cross-domain recommendation aims to leverage data collected from multiple domains to alleviate data sparsity issue, which is well summarized and classified in two surveys [10, 22]. Li *et al.* [22] classified existing researches to three categories: system-domain, data-domain and temporal-domain, while Cremonesi *et al.* [10] identified four different cross-domain scenarios as no-overlap, user-overlap, item-overlap and full-overlap. As for the researchfi eld of location recommendation, cross-domain solutions recently become a hot topic [11, 16, 24, 54, 60]. Some works leveraged data from other interaction domains as the auxiliary data. Cuo *et al.* [11] studied the task of store site recommendation and combined shop-site interaction matrices from different domains (collected in multiple cities). Some other works leveraged data collected in the social domain, the online social network for example, as an auxiliary domain to improve location recommendation [16, 24, 54]. Li *et al.* [24] transferred the social relation data to the target domain and considered friends' visited locations as auxiliary information. Gao *et al.* [16] combined co-relation attributes in social network in social domain and geographical attributes in the interaction domain to build features and perform recommendation. Yang *et al.* [54] combined social domain and interaction domain in the latent space via setting constraints to users' and their friends' embeddings. In this work, we focus on a task of cross-domain location recommendation leveraging two interaction domains with overlapped locations. It falls into category of *domain-level* in [22] and category of *item-overlap* in [10]. Specifically, *domain-level* defined in [22] refers to that users or items can be related across domains via attributes or identifiers; *item-overlap* in [10] is defined as that the multiple domains have very same functionality and overlapped items. In our task, locations (*i.e.*, POIs) are the so-called *overlapped items* across two domains.

**Location Privacy Protection** User mobility traces contain sensitive information of individual users such as personal habits, residence, etc. To protect those private information extracted from users' trajectory data, a lot of approaches [2, 20, 43, 44, 52, 53] are proposed based on three widely used privacy models, $k$-anonymity [46], $l$-diversity [32], and $t$-closeness [26]. Among them, some researches [44, 52, 53] merged several points as one region in order to fuzz up real location, while some other studies [2, 20, 43] proposed approaches generating dummy points from the real points. Recently, researchers apply differential privacy to user-location interaction data [1, 2]. Andrés *et al.* [2] introduced geo-indistinguishability, which utilized the criteria of differential privacy to make sure user's exact location is unknown while maintaining enough utility. Gergely *et al.* [1] proposed a differential privacy based release scheme for aggregated statistics of user-location interaction data.

**Privacy-preserving Recommendation.** Recommender system, a kind of personalized service, is closely related to user's information/profile, such as gender, age and historical behaviors. In general, traditional recommender systems include three components of data collection, model training and recommendation results presenting. There are works studying to attack these three components (data collection [35], model training [21], results presenting [5]) in recommender systems, respectively, to infer users' private information, such as behavioral records. Therefore, existing researches on privacy-preserving recommendation can be categorized according to which component the privacy protection mechanism is applied to. Some works [19, 23, 31, 37] adopted protection

mechanism when collecting data. Li *et al.* [23] transformed the raw trajectory dataset into a bipartite graph, and then extracted association matrix to inject carefully calibrated noise to meet differential privacy. Bin *et al.* [31] incorporated both interest-functionality interactions and users' privacy preferences to perform personalized App recommendations. Polat *et al.* [37] presented a scheme for the binary ratings-based top-N recommendation on horizontally partitioned data, in which two parties own disjoint sets of users' ratings for the same items, while preserving data owners' privacy. Ho *et al.* [19] apply differential privacy to generate region quadtree based on raw trajectory data. Some other works introduced privacy-preserving mechanism in training models [8, 33]. Chen *et al.* [8] presented a MF based model which splits latent user vectors into local and global parts to protect user privacy in the task of point-of-interest recommendation. Mcsherry *et al.* [33] introduced differential privacy during the training of MF model via adding noise to latent factors. Besides, some works applied protection mechanism tofi nal recommendation results [3, 38]. Riboni *et al.* [38] proposed the use of differential privacy to extract statistics about users' preferences and then provided recommendation from those statistics. Berlioz *et al.* [3] applied perturbation to MF's output, the recommendation results, to meet the criterion of differential privacy.

However, these works are not appropriate for our scenario. Specifically, different with single-domain recommendation, in cross-domain scenario, there is a new component of *data sharing*. To avoid the potential attack from the adversary staff hired by the target domain (*i.e.*, company), it is essential to employ protection mechanism on shared data which is ignored by existing works since they focus on single-domain scenario. If we merge databases (*i.e.*, considering multiple domains as a single domain) and apply existing recommendation methods on the target domain, then there must be a staff of the target domain can access the raw data from the auxiliary domain. To completely eliminate this possibility, in this work, we apply protection mechanism in data sharing.

## 3 PROBLEM DEFINITION AND SYSTEM OVERVIEW

### 3.1 Problem Definition

*3.1.1 Cross-domain Location Recommendation.* Location recommendation is defined to recommend new locations to users. In this paper, if not otherwise specified, we refer location recommendation as to recommend an individual POI, which is the most widely setting in literature [28, 29, 57, 58], to a user that has not visited there before. That is, the task of cross-domain location recommendation is defined as to utilize auxiliary domain's user-location (*i.e.*, user-POI) interaction data to help improving the performance of location recommendation (*i.e.*, POI recommendation) in the target domain.

In the *target domain*, where $M^t$ and $N^t$ denote the number of users and locations, respectively, we have a user-location matrix $\mathbf{Y}^t \in \mathbb{R}^{M \times N}$ with a binary value at each entry defined as follows,

$$y_{ul}^t = \begin{cases} 1, & \text{if } u \text{ has visited } l; \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Similarly, in the *auxiliary domain*, we have another binary user-location matrix $\mathbf{Y}^a \in \mathbb{R}^{K \times L}$, where $M^a$ and $N^a$ are the number of users and locations. Different from single-domain location recommendation that only leverages data from the target domain, $\mathbf{Y}^t$, cross-domain location recommendation considers both $\mathbf{Y}^t$ and $\mathbf{Y}^a$ to learn a predictive function estimating the likelihood that a user $u$ will visit a location $l$ that has never been visited before in the target domain. Note that the overlapped locations across two matrices serve as the *bridge* to transfer knowledge to the target domain, while users are different.

*3.1.2 Privacy-preserving Cross-domain Location Recommendation.* As mentioned in the introduction, direct sharing makes the raw data of interaction collected from the auxiliary domain accessible in the target domain. It causes high potential of privacy leakage since domains are operated by different companies ( *i.e.*, the staff of the target domain may become an adversary attacker). Then the task of privacy-preserving cross-domain location

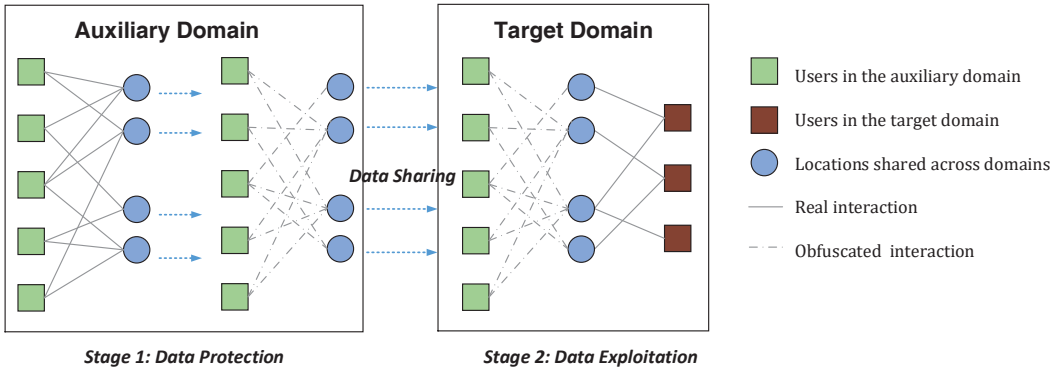



Fig. 1. Illustration of our solution for privacy-preserving cross-domain location recommendation.

recommendation can be defined as improving recommendation quality in the target domain with the help of auxiliary domain, while not sharing raw interaction data. It can be formulated as learning the predictive function from $\mathbf{Y}^t$ and $\mathbf{Y}^a$, given a limitation that $\mathbf{Y}^a$ cannot be directly shared to the target domain.

## 3.2 System Overview

To avoid the risk of leaking privacy due to the direct sharing of auxiliary domain's interaction data, we propose a novel design of the framework for privacy-preserving cross-domain location recommendation. Distinct from the typical problem settings of cross-domain location recommendation, we do not directly share the user interaction data (the user-location matrix $\mathbf{Y}^a$ in the auxiliary domain). Instead, we propose a solution that relies on a protection mechanism which makes the auxiliary domain's real (raw) interaction data not available in the target domain. Specifically, the framework of our solution can be divided into two stages, as illustrated in Figure 1.

At the *first stage* shown in Figure 1, we apply data protection mechanism in the auxiliary domain. That is, we perform a function $\mathcal{K}$ on $\mathbf{Y}^a$ to obtain obfuscated interaction data: $\tilde{\mathbf{Y}}^a = \mathcal{K}(\mathbf{Y}^a)$.

**Input:** Original user-location interaction data in the auxiliary domain $\mathbf{Y}^a$.

**Output:** Obfuscated user-location interaction data in the auxiliary domain $\tilde{\mathbf{Y}}^a$.

At the *second stage* illustrated in Figure 1, we extract signals from the obfuscated auxiliary data $\tilde{\mathbf{Y}}^a$ to help improving recommendation quality in the target domain:

**Input:** The user-location interaction data in the target domain $\mathbf{Y}^t$, and the obfuscated user-location interaction data $\tilde{\mathbf{Y}}^a$ from the auxiliary domain.

**Output:** A predictive model to estimate the likelihood that a user $u$ will visit a location $i$, which was not visited by her previously in the target domain.

Specifically, taking $u$, $l$, and transferred obfuscated trajectory $\tilde{Y}^a$, the model has to predict,

$$\hat{y}^t_{ul} = f(u, l), \tag{2}$$

where $\hat{y}^t_{ul} \in [0, 1]$ denotes the probability of interaction between user $u$ and location $l$. With the predictive model, we score the locations not visited before for a given user $u$, and select the top-ranked (*i.e.,* with higher visiting probability) locations as the recommendation results for $u$.

## 4 METHOD

Our proposed framework is featured with two special designs corresponding to the two stages in the task of privacy-preserving cross-domain location recommendation:

- **Differential-privacy based Data Obfuscation (*Data Protection* in Fig. 1)**. To protect interaction data of the auxiliary domain, we adopt obfuscation mechanism to add noise. With the power of differential privacy, we propose a generalized version of geo-indistinguishability, making sure users' exact locations are still unknown even if attacker can access to the *obfuscated noisy data*.
- **Confidence-aware Matrix Factorization based Recommendation (*Data Exploitation* in Fig. 1)**. To effectively distill useful signals from obfuscated auxiliary data, we calculate a confidence matrix, of which the elements stand for the possibility that the corresponding elements in transferred matrix are real interaction records. We then combine it with interaction matrices of two domains to perform a collective matrix factorization task for recommendation.

### 4.1 Data Protection

*4.1.1 Generalized Geo-Indistinguishability.* To protect transferred interaction data, we add noise to raw data. Thefi rst work of differential privacy [13] generated noise based on Hamming distance. For location data that is made up of a series of coordinates, Euclidean distance based noise is adopted to provide *indistinguishablity*, so as to meet the criterion of differential privacy while preserving enough utility [2]. For our task of location recommendation, in order to combine semantics of POI with location information, we propose the generalized version of geo-indistinguishability with an improved distance metric. Before we dive into the definition, wefirst introduce some notations.

Given a set of points of interest $\mathcal{T}$, where each POI is denoted as a tuple $t = (x, y, c)$ containing spatial coordinates, $(x, y)$, and POI category ID, $c$, respectively. The set of possible obfuscated values is $\mathcal{P}$. Then the mechanism $\mathcal{K}$ reports randomly selected value $p \in \mathcal{P}$ based on real location $t$, with the probability of $\Pr(\mathcal{K}(t)(p))$.

For an adversary model, we model the attacker's side information as $\pi(t)$, which is the probability of visiting a specific place $t$. According to Bayes' rule, a posterior distribution can be defined as $\sigma = \frac{\mathcal{K}(t)(p)\pi(t)}{\sum_{t'} \mathcal{K}(t')(p)\pi(t')}$. Here we assume that the attacker know the whole mechanism.

The key for generalized geo-indistinguishability is the definition of distance metric. Instead of choosing Hamming distance in [13] or Euclidean distance in [2], here we propose a new distance metric, called Semantic-Euclidean distance, denoted as $d_S(t, t')$.

Definition4.1. *(Semantic-Euclidean Distance). A distance metric $d_S$ using both semantics and location information is defined on all $t, t' \in \mathcal{T}$:*

$$d_S(t, t') = \begin{cases} d_{euc}\left((t.x, t.y), (t'.x.t'.y)\right) & t.c = t'.c; \\ \infty & t.c \neq t'.c, \end{cases} \tag{3}$$

which allows two secrets to be distinguishable. The intention here is that we only apply obfuscation within the same category and leave those locations from different categories distinguishable, which preserves the semantics of each POI to the category level while protecting location privacy at the same time.

Now we can introduce the probabilistic model of geo-indistinguishability using these notations.

Definition4.2. *(Generalized Geo-Indistinguishability[3]). A mechanism $\mathcal{K}$ satisfies generalized $\epsilon$-geo-indistinguishability iff for all $t, t' \in \mathcal{T}$:*

$$d_P(\mathcal{K}(t), \mathcal{K}(t')) \leq \ \epsilon d_S(t, t'), \tag{4}$$

*where $d_P$ stands for the multiplicative distance between two distributions $\sigma_1$, $\sigma_2$ on some set $S$ as $d_P(\sigma_1, \sigma_2) = \sup_{S \subseteq S} |ln(\frac{\sigma_1(S)}{\sigma_2(S)})|$.*

Similar to Planar Laplace Mechanism introduced in [2], we canfi rst achieve generalized geo-indistinguishability by mapping each true POI $t$ to a randomly drawn point $p$ in the infinite continuous space $\mathcal{P}$ according to the probability density function, which is formulated as follows,

$$D_\epsilon(t)(p) = \frac{\epsilon^2}{2\pi} e^{-\epsilon d_S(t,p)}, \tag{5}$$

where the space $\mathcal{P}$ is constructed by taking Cartesian product between planar Cartesian space and the set of category id.

Note that although the semantic-Euclidean distance metric guarantees that the obfuscated location shares the same category id with the original POI, the obfuscated point $p$ may not be a valid POI, since x,y axis of space $\mathcal{P}$ is infinite and continuous. Thus, remapping is adopted here by transforming $p$ to the nearest POI tuple using KD-Tree whilefi xing the value of c, and then discretization and truncation are achieved.

**Prop. 1** claims that the remapping for achieving discretization and truncation preserves generalized $\epsilon$-geo-indistinguishability with a proof using the notation of differential privacy.

Proposition1. *(Privacy-Preserving Remapping). Let $\mathcal{K} : \mathcal{T} \rightarrow \mathcal{P}$ be a randomized mechanism mapping a POI tuple $t$ in the discrete set $\mathcal{T}$ to another tuple $p$ in the continuous space $\mathcal{P}$, which preserves generalized $\epsilon$-geo-indistinguishability. Let $\mathcal{R} : \mathcal{P} \rightarrow \mathcal{T}$ be a deterministic remapping which remaps a tuple $p$ to the nearest POI tuple $z$ having the same POI category as $p$. Then $\mathcal{R} \circ \mathcal{K} : \mathcal{T} \rightarrow \mathcal{T}$ still preserves generalized $\epsilon$-geo-indistinguishability.*

Proof. Consider two POI tuples in $\mathcal{T}$, namely $t_1$ and $t_2$. Let $S = \{p \in \mathcal{P} : \mathcal{R}(p) = z\}$, then we have:

$$\begin{aligned} \Pr[\mathcal{R}(\mathcal{K}(t_1)) = z] &= \Pr[\mathcal{K}(t_1) \in S] \\ &\leq e^{\epsilon d_S(t_1, t_2)} \Pr[\mathcal{K}(t_2) \in S] \\ &= e^{\epsilon d_S(t_1, t_2)} \Pr[\mathcal{R}(\mathcal{K}(t_2)) = z] \end{aligned} \tag{6}$$

□

The above algorithm of location obfuscation can be summarized as **Alg. 1**. In **Alg. 1**, wefi rst sampled two random variables from two uniform distributions, namely $\theta$ and $p$, and therefore derived the obfuscated distance $r$ via Lambert W function so that a planar Laplacian noise can be sampled efficiently in a polar system. Adding the noise to the original POI leads to an obfuscated location on the continuous plane. To transform it to a valid POI, wefi nd its nearest neighbor in the KD-Tree which contains all POIs within the same category. The last step is also known as privacy-preserving remapping as we mentioned earlier.

Here we state a characterization of this mechanism, which provides some insights about what privacy guarantee that our mechanism can provide by comparing prior and posterior of a certain user learned by the adversary. By using Bayes' rule, we can derive the following characterization with the help of **(4)**:

$$\frac{\Pr(t|S)}{\Pr(t'|S)} = \frac{\Pr(\mathcal{K}(t)(S))\,\pi(t)}{\Pr(\mathcal{K}(t')(S))\pi(t')} \leq e^{\epsilon r} \frac{\pi(t)}{\pi(t')}, \forall r > 0, S \in \mathcal{T}, \forall t, t' : d_S(t, t') \leq r \tag{7}$$

[3]The original definition of geo-indistinguishability is also compatible with this generalized version, if we discard all information about semantic and treat all locations as the same category.

As **(7)** shows, the observation has limited effect to improve the knowledge of the adversary when compared with side information, since the ratio between them is bounded by the maximum distance in the possible location set. Therefore, little information gain can be obtained with the observation under our mechanism.

**Algorithm 1** Data Obfuscation

**Input**: $t$, $\epsilon$, KD-Tree ▷ $t$: original POI tuple, $\epsilon$: privacy level
**Output:** $\hat{t}$ ▷ $\hat{t}$: obfuscated POI tuple

$\theta \leftarrow \boldsymbol{U}[0, 2\pi]$
$p \leftarrow \boldsymbol{U}[0, 1]$
$r \leftarrow -\frac{1}{\epsilon}(\boldsymbol{W_{-1}}(\frac{p-1}{e}) + 1)$ ▷ $\boldsymbol{W_{-1}}$: Lambert W function(-1 branch)
$< \hat{t}.x, \hat{t}.y > \leftarrow < t.x, t.y > + < r\cos\theta, r\sin\theta >$ ▷ transform to Cartesian system
$\hat{t} \leftarrow$ KD-Tree.query($\hat{t}$) ▷ search nearest POI within same category

## 4.2 Data Exploitation (Recommendation)

Since the transferred auxiliary data is obfuscated in the stage of *Data Protection*, it cannot accurately represent users' preferences towards different locations. Therefore it is challenging to utilize the *noisy* data to help improving recommendation performance in the target domain. Besides, there is another challenge about how to balance influence of two domains' data. To address them, wefi rst calculate confidence for the transferred matrix to discriminate its element to be *possibly real* or *possibly fake*. We then combine this confidence matrix with interaction matrices to perform a collective matrix factorization task for recommendation. This stage, *Data Exploitation*, is featured with two designs: 1) confidence matrix helps us tofi lter noise and keep useful signals reserved in transferred matrix; and; 2) collective matrix factorization can effectively balance two domains' influence with joint learning.

*4.2.1 Inferring Confidence Matrix.* For the transferred user-location interaction matrix, directly utilizing it to train a recommendation model does not work since each observed visited POI is the output of the protection mechanism $\mathcal{K}$, which may not represent the real visited POI. That is, some observed POIs in the transferred visiting matrix are likely to be unobserved in the real visiting matrix and vise visa. To address this challenge, we introduce a confidence matrix to help understanding how *confident* each element (no matter observed or unobserved) is in the transferred interaction matrix.

With regard to our scenario and obfuscation mechanism, using posterior distribution which correlates with Semantic-Euclidean distance between real and obfuscated points is a natural choice, since neighboring locations always share similar interaction patterns. Here, the neighborhood is defined on the semantic-location space. In order not to treat those obfuscated locations as real ones (*i.e.*, false-positive records) and not to mistakenly ignore potential real locations (*i.e.*, false-negative records), we obtain a confidence matrix $C^a$ to measure to what extent we could rely on the obfuscated records.

With a large amount of POIs within each category in our target area, it is inefficient to calculate the normalization factor (the denominator) in the formula of posterior mentioned in 4.2.1. Thus, only a sample of locations with highest posterior is taken into consideration. Without loss of generality, we assume the prior for each POI is equally-likely, and therefore only the transfer probability derived from our mechanism $\mathcal{K}$ matters. Recall that the probability of perturbing a real POI to another decreases exponentially with the semantic-euclidean distance, it's straight-forward to conclude that locations with highest posterior (or confidence) of being the real locations are those in the surrounding of the obfuscated location. Therefore, only a sample of $m$ locations are assumed to have non-zero values of confidence. After normalization, the confidence for user $u$ to have a real visiting record at

location $l$ can be derived as follows,

$$c_{ul} = \frac{e^{-\epsilon d_S(t_l, t')}}{\sum_{i=1}^{m} e^{-\epsilon d_S(t_i, t')}}, \quad l = 1, 2, ..., m, \quad t_l \in \mathcal{L}_m, \tag{8}$$

where $\mathcal{L}_m$ denotes the set of $m$-nearest locations surrounding the obfuscated location $t'$ in the semantic-location space. Here we omit the normalization constant in the PDF (Probabilistic Distribution Function) of planar Laplacian noise for simplicity. Note that for those locations beyond $\mathcal{L}_m$, we set their corresponding confidence to zero, indicating that the real location is unlikely to be among them. Since each user may check in several places in the auxiliary domain and every obfuscated location can derive a confidence vector for every POI, some locations can have several confidence at the same time. Here we use the maximum value of each location among all possible confidence values, given the fact that every location can only be perturbed to another location.

*4.2.2 Apply Confidence to Matrix Factorization.* We utilize the model of matrix factorization (MF), which is frequently used in thefi eld of location recommendation [28, 29]. The objective function of basic MF for a single user-location interaction matrix $\mathbf{Y}$ can be formulated as:

$$\min_{\mathbf{p}_*, \mathbf{q}_*} \sum_{u=1}^{M} \sum_{l=1}^{N} (y_{ul} - \mathbf{p}_u{}^T \mathbf{q}_l)^2 + \lambda_P ||\mathbf{P}||_2^2 + \lambda_Q ||\mathbf{Q}||_2^2, \tag{9}$$

where $\mathbf{p}_u$ and $\mathbf{q}_l$ represent the latent vector of user $u$ and location $l$ respectively, which are column vectors of latent matrix $\mathbf{P}$ and $\mathbf{Q}$. In addition, $\lambda_P$ and $\lambda_Q$ are the $L_2$ regularizer for user and location matrices, respectively.

As mentioned before, in this paper we focus on a location-overlap scenario, where two domain's data are collected from two companies operated on a same city. Therefore, we assume POIs' inherent features keep steady across domains. Thus, it is intuitive to let the overlapped POIs share the same embeddings across two domains. Thus, we formulate the objective function to optimize in our CCMF (confidence-aware collective matrix factorization) model as:

$$\min_{\mathbf{p}^a_*, \mathbf{p}^t_*, \mathbf{q}_*} \sum_{u=1}^{M^a} \sum_{l=1}^{N} \omega_a c_{ul} (y^a_{ul} - \mathbf{p}^{a\,T}_u \mathbf{q}_l)^2 + \sum_{u=1}^{M^t} \sum_{l=1}^{N} \omega_t (y^t_{ul} - \mathbf{p}^{t\,T}_u \mathbf{q}_l)^2, \tag{10}$$

where $\omega_a$ and $\omega_t$ are coefficients controlling the influence of these two domains and we have $\omega_a + \omega_t = 1$. Here $\mathbf{p}^a_u$ and $\mathbf{p}^t_u$ denote the embeddings of user $u$ in two domains, respectively [4]; $\mathbf{q}_l$ denotes the shared embedding of the overlapped item $i$ across two domains; $\mathbf{y}^a_{ul}$ and $\mathbf{y}^t_{ul}$ denote the interaction in two domains, respectively. Note that we omit the $L_2$ regularization term for clarity. We can observe this is a joint objective function is made up of objective functions of two tasks on two domains.

*4.2.3 Training.* In our task of location recommendation, both two matrices need to be factorized are binary. We adopt *negative sampling*, a widely-used training manner for implicit matrices in existing researches [18, 34], to learn the latent matrices $\mathbf{P}^a$, $\mathbf{P}^t$ and $\mathbf{Q}$. We introduce negative sampling to stochastic gradient descent (SGD) , a widely generic solver for machine learning, to optimize our proposed CCMF method. Specifically, to construct a mini-batch, wefi rst sample a batch of historical user-location interaction pairs $(u, l^a)$ and $(u, l^t)$ on two domains. For each $(u, l^a)$, we then adopt the negative sampling technique, to randomly select unobserved locations in the auxiliary domain $\{l'^a_1, l'^a_2, ..., l'^a_n\}$ for user u with a sampling ratio of $n$. It is the same with target domain's pair $(u, l^t)$. With the two constructed mini-batches, we take a gradient step to minimize the objective function. Here we give the updating rule in SGD as follows.

[4]The users of two domains are always not overlapped, which means user 1 in target domain and auxiliary domain are two totally different users.

For auxiliary domain's mini-batch we have:

$$\begin{cases} \epsilon^a_{ul} = 2(y^a_{ul} - \mathbf{q}_l \cdot \mathbf{p}^a_u), \\ \mathbf{q}_l \leftarrow \mathbf{q}_l + \mu(\omega_a \epsilon^a_{ul} \mathbf{p}^a_u - \lambda_Q \mathbf{q}_l), \\ \mathbf{p}^a_u \leftarrow \mathbf{p}^a_u + \mu(\omega_a \epsilon^a_{ul} \mathbf{q}_l - \lambda_P \mathbf{p}^a_u); \end{cases} \tag{11}$$

and for target domain' mini-batch, similarly, we have:

$$\begin{cases} \epsilon^t_{ul} = 2(y^t_{ul} - \mathbf{q}_l \cdot \mathbf{p}^t_u), \\ \mathbf{q}_l \leftarrow \mathbf{q}_l + \mu(\omega_t \epsilon^t_{ul} \mathbf{p}^t_u - \lambda_Q \mathbf{q}_l), \\ \mathbf{p}^t_u \leftarrow \mathbf{p}^t_u + \mu(\omega_t \epsilon^t_{ul} \mathbf{q}_i - \lambda_P \mathbf{p}^t_u), \end{cases} \tag{12}$$

where $\mu$ and $\lambda$ denote learning rate and regularization terms, respectively.

In conclusion, to exploit the transferred obfuscated interaction data, wefi rst calculate a confidence matrix to extract useful signals, and we then apply collective matrix factorization to combine this confidence matrix and two interaction matrices to perform the recommendation task.

## 5 EVALUATION

In this section, we conduct extensive experiments on two real-world datasets to answer the following three research questions:

- **RQ1:** How does our proposed CCMF method perform in the task of top-K recommendation compared with single-domain methods and cross-domain methods that leak user privacy? Can our proposed method alleviate data sparsity issue? What is the relationship between the recommendation performance with obfuscated and non-obfuscated data?
- **RQ2:** What is the relationship between the recommendation accuracy and privacy bound in our proposed solution?
- **RQ3:** How do the key hyper-parameters, building manner of simulated cross-domain dataset, and density of area affect our proposed solution's performance?

In what follows, wefi rst describe the experimental settings, and then answer the above three research questions.

### 5.1 Experimental Settings

*5.1.1 Datasets and Evaluation Protocol.* We experiment with two real-world POI check-in datasets which both contain interaction data from two domains.

- **Wechat-Foursquare**. Wechat[5] is the biggest online social network service in China. Users can check-in with Wechat mobile App, which is known as Moment. We collect users' check-in records in the two largest cities in China, Beijing and Shanghai, both within the time period from 2017-06-01 to 2018-05-31. Each check-in record includes an anonymous user identifier, a POI identifier and timestamp. Foursquare[6] is a famous world-wide check-in service. We utilize the released dataset in [55] including long-term (about 18 months from April 2012 to September 2013) global-scale check-in data collected from Foursquare. Each record in this dataset contains an anonymous user identifier, a POI identifier and timestamp. We carefully match this dataset with Wechat dataset as follows. First, we obtain POI attributes (coordinate, name, address and category) via querying POI identifier through Foursquare API[7]. Then we obtain POI identifier in Wechat dataset via querying these POI attributes

[5]https://weixin.qq.com
[6]https://www.foursquare.com
[7]https://developer.foursquare.com

Table 1. Statistics of Dataset

| Dataset | City | Location# | Target Domain | | | Auxiliary Domain | | |
|---|---|---|---|---|---|---|---|---|
| | | | User# | Record# | Sparsity | User# | Record# | Sparsity |
| SimuWechat | Beijing | 19106 | 3000 | 6326 | 99.989% | 7000 | 90781 | 99.932% |
| | Shanghai | 18891 | 3000 | 5389 | 99.990% | 7000 | 85467 | 99.935% |
| Wechat-Foursquare | Beijing | 3187 | 899 | 7556 | 99.736% | 10368 | 62609 | 99.811% |
| | Shanghai | 4068 | 1063 | 10814 | 99.750% | 17680 | 111032 | 99.846% |

through Wechat Map API[8]. Last, we match these POI identifiers across Foursquare dataset and Wechat dataset mentioned above to obtain cross-domain interaction matrices. We release this precious dataset[9] and we believe it will benefit the community.

- **SimuWechat** Since there is no public dataset for the task, to build another dataset other than Wechat-Foursquare, we utilize above mentioned Wechat dataset to simulate a cross-domain scenario. Wefi rst sort the users using the number of interactions in a descending way. Users within the top 70% along with their interactions are considered as the auxiliary domain, while the others are used as the target domain, which simulates the data sparsity problem in the target domain. We name this dataset as SimuWechat since it is a simulated cross-domain dataset. Such operations of building simulated dataset is similar as [10].

The statistics of two utilized datasets are summarized in Table 1. In the evaluation stage, given a user in the testing set, each algorithm ranks a test location with 99 locations that the user has not visited before. We applied the widely used leave-one-out technique, a common setting in existing works [14, 18, 30], to obtain the training set and test set. Therefore for every user, there always exists a test location she has not visited before. We then adopted two popular metrics, *HR* and *NDCG*, to judge the performance of the ranking list:

- **HR@K:** *Hit Ratio* (HR) measures whether the test location is contained by the top-K location ranking list (1 for yes and 0 for no).
- **NDCG@K:** *Normalized Discounted Cumulative Gain* (NDCG) complements HR by assigning higher scores to the hits at higher positions of the ranking list.

*5.1.2 Baselines.* We compared the performance of our proposed CCMF with four baselines, which can be divided into two groups based on whether it models single-domain or cross-domain data.

For thefi rst group, we use the state-of-the-art matrix factorization method for single domain.

**SMF [34]** : Matrix factorization is a competitive recommendation method for single domain. Mnih *et al.* [34] proposed a matrix factorization method for implicit data with the help of negative sampling technique. We name this method as SMF (short for **S**ingle-domain **MF**)

The second group contains four methods for cross-domain interaction data.

**CMF [45]** Collective matrix factorization is proposed to factorize multiple matrices simultaneously. It can be applied directly to factorize transferred noisy auxiliary matrix and the target interaction matrix, with sharing embeddings of locations across two domains. Here we denote the collective matrix factorization without considering confidence as CMF.

**WC-CCMF** In the stage of data protection, there is a degeneralized version of our CCMF to ignore category of location. We name this as WC-CCMF (**W**ithout **C**ategory **CCMF**).

**WC-CMF** Similar as CMF, after we conduct the protection mechanism ignoring category of locations we name the method factorizing transferred auxiliary and the target interaction matrices as WC-CMF.

[8]https://lbs.qq.com
[9]https://github.com/FIBLAB

**RN-CMF** Despite our specially designed obfuscation mechanism, a *random-choice* noise can also be adopted to protect privacy. Specifically, for each record in the raw auxiliary data, we randomly select one from the whole location sets with the same category. Then we perform collective matrix factorization on transferred data and target data directly, and name this method as RN-CMF. Note that because the obfuscated location is sampled from the uniform distribution, RN-CMF represents a special case where transferred records is obfuscated with the largest amount of noise.

*5.1.3 Parameter Settings.* To obtain the optimal hyper-parameter setting, we build validation set for each user similar with building test set. We adopt vanilla SGD to optimize all the methods, and for each method, we tune its learning rate $\mu$ in [0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1] and regularization term $\lambda_P, \lambda_Q$ in [0.001, 0.002, 0.005, 0.1, 0.2, 0.5, 1, 1.5, 2, 3, 5, 10] to report the best performance. For cross-domain methods, we carefully tune the weight of each domain, $\omega_a$ and $\omega_t$ respectively, from 0 to 1 with a step size 0.1. Wefi x the sampling ratio to 4 and size of mini-batch to 128.

## 5.2 Performance Comparison (RQ1)

Wefi rst compare the top-$K$ performance of our proposed CCMF and baseline methods. We study the performance with setting $K$ from 1 to 10. We choose a relatively small $K$ since we rank the test item in a list with 100 locations, as mentioned above. In Figure 2, we present the top-K recommendation performance for the two utilized real-world datasets. We compare our proposed CCMF method with one single-domain baselines and four cross-domain ones. We also plot the performance of cross-domain recommendation without applying any protection mechanism. That is, we apply collective matrix factorization on raw interaction data, named **RAW-CMF**. From these results, we have the following observations:

- **Our proposed CCMF significantly improves recommendation performance in the target domain.** For two cities' data in both SimuWechat and Wechat-Foursquare datasets, our proposed CCMF achieves the best performance in most metrics. Note that for all top-K setting, our proposed CCMF method achieves significant performance gain (12.69%-33.10% for HR and 18.71%-30.11% for NDCG on SimuWechat-Beijing, 9.88%-51.43% for HR and 25.30%-51.40% for NDCG on SimuWechat-Shanghai, 10.13%-87.50% for HR and 31.11%-87.50% for NDCG on Wechat-Foursquare-Beijing, and 7.88%-93.0% for HR and 30.67%-93.65% for NDCG on Wechat-Foursquare-Shanghai.), which demonstrates its efficacy. Only for SimuWechat dataset in Beijing, CCMF achieve similar or slightly worse performance compared with its simplified version, WC-CCMF, which means category information brings no gain in this scenario. Even so, it still outperforms other baseline methods significantly. On the other hand, cross-domain methods directly using protected auxiliary interaction data, CMF, RN-CMF and WC-CMF, achieve poor recommendation performance in all datasets on two adopted metrics.This phenomenon can be easily explained that the protection mechanisms harm the utility of interaction data of the auxiliary domain while protecting location privacy. Among these three methods, RN-CMF, which perturbs the auxiliary interaction matrix via random choice achieves the worst performance, because the obfuscated matrix is the most noisy one.
- **No matter what kind of protection mechanism is adopted, models considering confidence always achieve better recommendation.** Specifically, CCMF and WC-CCMF significantly outperform CMF and WC-CMF, respectively. The quantitative result of gain can be found in 2. This verifies the effect of our specially designed confidence technique.
- **The single domain recommendation method, SMF, achieve the worst recommendation performance, out of all the methods.** It shows that it is essential to borrow auxiliary data to help improving recommendation in our utilized real-world datasets.
- **Sacrifice of utility is minor and acceptable for our proposed CCMF to preserve privacy.** Compared with RAW-CMF on unobfuscated data, CCMF achieves a similar or even better performance while protecting

Table 2. The gain on HR and NDCG for different datasets.

| Dataset | City | Metric | Gain of WC-CCMF v.s.WC-CMF | Gain of CCMF v.s. CMF |
|---|---|---|---|---|
| Wechat-Foursquare | Shanghai | HR | 7.88%-93.66% | 13.61%-111.76% |
| | | NDCG | 23.88%-71.43% | 43.09%-139.04% |
| | Beijing | HR | 2.05%-87.50% | 10.13%-84.79% |
| | | NDCG | 23.32%-87.47% | 31.10%-84.78% |
| SimuWechat | Shanghai | HR | 10.90%-57.09% | 9.89%-51.44% |
| | | NDCG | 28.50%-51.70% | 25.30%-51.45% |
| | Beijing | HR | 10.90%-29.19% | 12.52%-30.12% |
| | | NDCG | 17.81%-29.18% | 18.84%-30.12% |

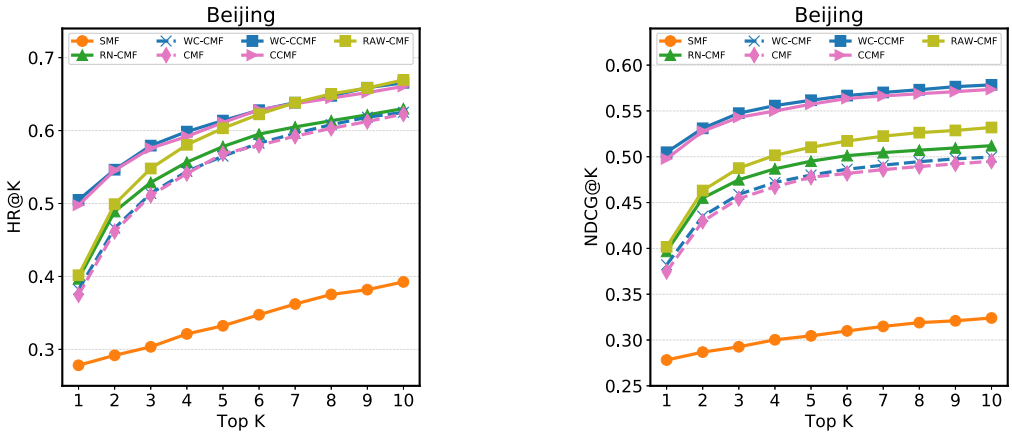



Fig. 2. Top-K recommendation performance comparison on SimuWechat dataset in Beijing(K is set from 1 to 10)

user privacy. This is a interesting and surprisingfi nding that our method with obfuscated data can outperform baseline methods with non-obfuscated data. It can be explained from the perspective of learning. Actually, in recommender systems, positive sample (*i.e.*, observed interaction) is always sparse due to the data-sparsity issue. Recently, some studies [6, 51] demonstrated that selecting some unobserved interactions similar to observed ones as fake positive samples can significantly enhance the learning of recommendation model and further improve the recommendation performance. In our model, some unobserved locations near observed ones are regarded as positive samples. Such operation benefits the model learning since neighboring locations are very similar from the perspective of collaborativefi ltering (a user is very likely to visit a location near her visited locations).

Since the main purpose of cross-domain recommendation is to alleviate the data sparsity and cold start problem, we further study our proposed method's recommendation performance for those sparse locations. Specifically, we apply the same evaluation protocol with above experiments, *leave-one-out*. For each location, its performance is defined as the average of HR@10 and NDCG@10 when it is in the test set. Here we divide locations into two groups: sparse (has been visited less than 5 users) and non-sparse (has been visited no less than 5 users). And we conduct experiments on the proposed CCMF and CMF and present the performance gain compared with SMF in Table 3 and Table 4. From the tables we can observe that for both datasets, our proposed CCMF can efficiently improve recommendation performance for both sparse and non-sparse locations. Note that in

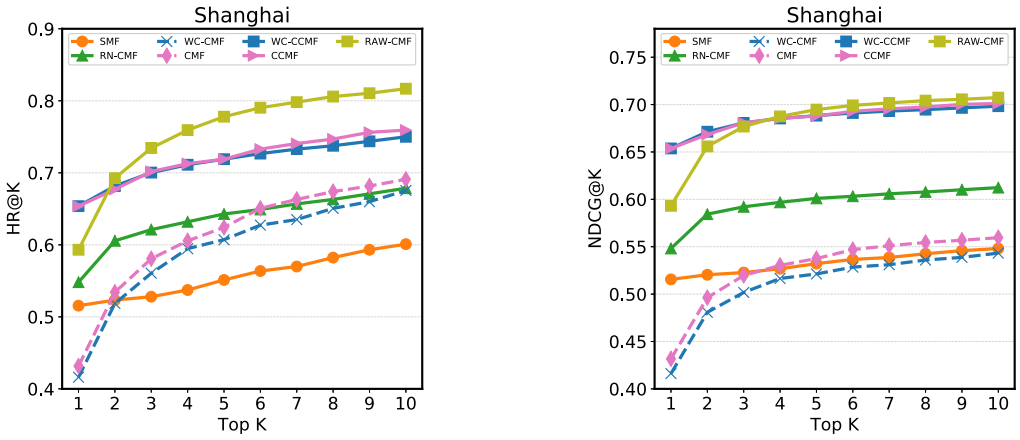



Fig. 3. Top-K recommendation performance comparison on SimuWechat dataset in Shanghai (K is set from 1 to 10)

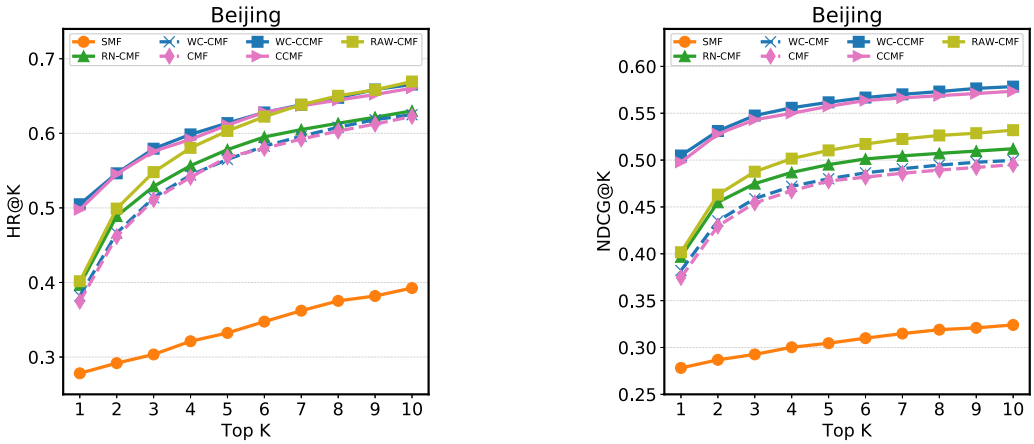



Fig. 4. Top-K recommendation performance comparison on Wechat-Foursquare dataset in Beijing(K is set from 1 to 10)

Table 3. The comparison between gain of HR@10 and NDCG@10 on sparse and non-sparse locations in SimuWechat dataset, where $\epsilon$=2.

| City | **Beijing** | | | | **Shanghai** | | | |
|---|---|---|---|---|---|---|---|---|
| Metric | **HR@10** | | **NDCG@10** | | **HR@10** | | **NDCG@10** | |
| Location type | Sparse | Non-sparse | Sparse | Non-sparse | Sparse | Non-sparse | Sparse | Non-sparse |
| CMF | -14.81% | -18.57% | -23.04% | -4.32% | -1.45% | 8.24% | -4.28% | 5.23% |
| Our CCMF | **17.66%** | **18.33%** | **12.69%** | **38.59%** | **41.12%** | **23.40%** | **42.47%** | **30.61%** |

Shanghai, target domain's interaction data is sparser, therefore we can observe a significant performance gain for sparse locations. For Wechat-Foursquare dataset, HR@10 gain for Shanghai's non-sparse locations is $-0.827\%$, which can be explained that those frequently visited locations are not so dependent on borrowing auxiliary data to learn their features, and then transferring data may bring noise instead of useful signals.

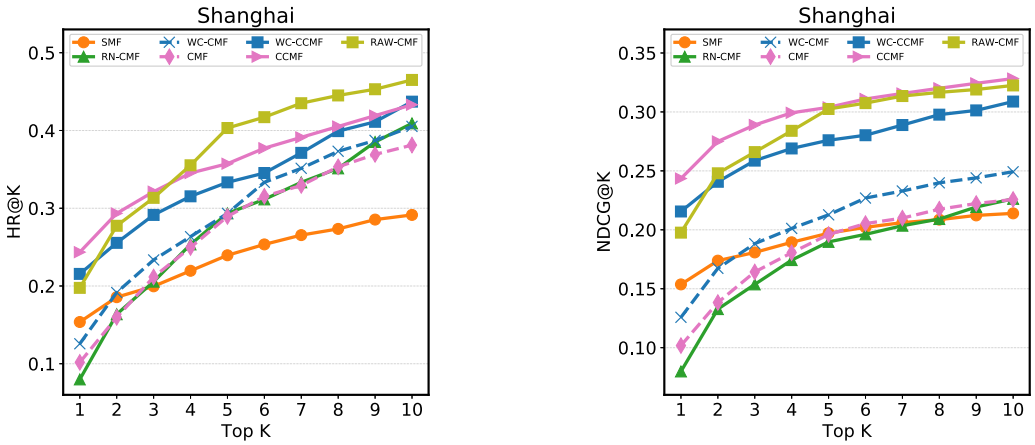



Fig. 5. Top-K recommendation performance comparison on Wechat-Foursquare dataset in Shanghai(K is set from 1 to 10)

Table 4. The comparison between gain of HR@10 and NDCG@10 on sparse and non-sparse locations in Wechat-Foursquare dataset, where $\epsilon$=2.

| **City** | **Beijing** | | | | **Shanghai** | | | |
|---|---|---|---|---|---|---|---|---|
| Metric | HR@10 | | NDCG@10 | | HR@10 | | NDCG@10 | |
| Location type | Sparse | Non-sparse | Sparse | Non-sparse | Sparse | Non-sparse | Sparse | Non-sparse |
| CMF | 3.60% | 6.02% | -1.55% | 5.62% | 2.68% | 13.07% | -2.01% | -1.38% |
| CCMF | 8.47% | 15.54% | 6.37% | 11.97% | 16.98% | -0.827% | 12.28% | 6.30% |

In summary, experiments on two datasets demonsrate our proposed CCMF method significantly outperform baseline the-state-of-the-art methods. In addition, CCMF can achieve similar or even better performance than RAW-CMF on obfuscated data. Further studies show that CCMF can effectively alleviate data sparsity issue.

## 5.3 Utility and Privacy (RQ2)

In our proposed CCMF solution, the parameter $\epsilon$ controls the privacy bound of protected auxiliary interaction data. Obviously, there is a trade-off between the privacy bound and utility for the data-protection stage. Here we use the expected quality loss to measure the intensity of noise and study how the noise affects recommendation performance. The expected quality loss can be formulated as below, where we only focus on the loss of location privacy:

$$QL(\mathcal{K}, \pi, d_{euc}) = \sum_{t,z \in \mathcal{T}} \pi(t)\mathcal{K}(t)(z)d_{euc}(t,z), \tag{13}$$

where $d_{euc}$ denotes Euclidean distance metric. According to [7] , when using planar Laplacian mechanism (PL for short) due to the symmetry in $\mathbb{R}^2$ the Euclidean quality loss of PL is independent from the prior $\pi$ and $QL(PL, \pi, d_{euc}) = 2/\epsilon$. The intention here is that we model the average obfuscation distance as the quality loss, since distance is an important feature in location based service.

We choose $\epsilon$ from {0.4, 2, 4, 20}, corresponding to expected quality loss of {5km, 1km, 500m, 100m} respectively. For experiment, we use both SimuWechat and Wechat-Foursquare dataset in Beijing as an example, since Beijing is a larger city than Shanghai. As depicted in **Fig. 6**, we report how the amount of noise hurts our model's performance when compared to the baselines. From these results, the following observations can be obtained:

- **Models using cross-domain data always significantly outperform the single-domain baseline, namely SMF.** Although the noise varies in a large range, the performances of CMF-like models are much better than SMF, which indicates the robustness of CMF-like models for learning cross-domain features.
- **With confidence enhancement, the models show afl atter trend in the degradation of performance when the noise gets larger, which indicates the improvement of robustness given by the confidence.** Our specially designed confidence inference technique helps extract more useful information and therefore helps achieve better performance for CCMF and WC-CCMF, compared with CMF and WC-CMF which directly utilize noisy auxiliary data. Also the CCMF and WC-CCMF model enjoy a higher gain in performance when the amount of noise is larger, which indicates that the confidence plays a more important role.
- **The semantics of the locations help improve the performance after taking confidence into consideration in most scenarios.** However, for real-world cross domain scenarios, under two extreme cases of noise when $\epsilon$=0.2 and 20, corresponding to quality loss of 10km and 100m, the improvement is not significant. For small amount of noise, WC-CCMF performs even better than CCMF. The reason is quite reasonable when considering the pattern of interactions in real life. For those locations sharing the same function within a small region, they are always in a competing manner. Users may have special preferences for only one or two of them. However, for those locations from different categories, users may have interactions with many of them because they are all in walking distance and provide quite different services. Therefore, constraining the obfuscation in the same category may introduce more noise to the inference of user preference. For high level of privacy protection, the real locations can be far away from the obfuscated ones, therefore even the obfuscated location shares the semantics, users are unlikely to visit them as frequently as the real one. These are cases where distance plays an equal or even more important role than semantics. However, this amount of noise is unrealistic in real life since it either provides no privacy guarantee or no utility. Note that when the amount of noise is mild, our CCMF model performs better than WC-CCMF and is less influenced by the noise, which provides more options for real-world platforms to choose from without a significant drop in performance.

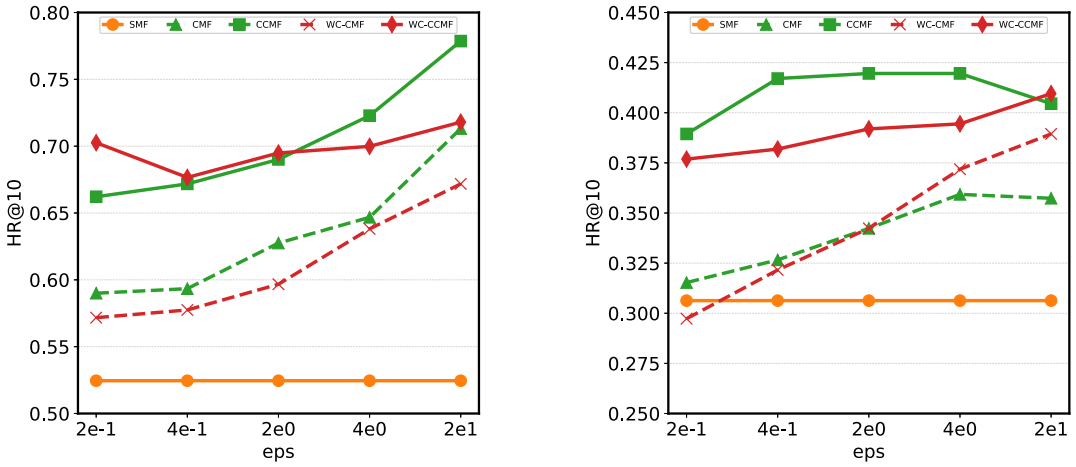
In summary, we study the trade-off between efficacy and privacy of our proposed framework. The experimental results show that the sacrifice of utility to preserve privacy is definitely acceptable.

### 5.4 Hyper-parameter Study (RQ3)

For our proposed CCMF method, there are some significant hyper-parameters closely related to the recommendation performance. Therefore, we focus on a key hyper-parameter, dimensionality *dim* and evaluate the performance on our utilized datasets with different settings.

We compare the performance of all methods in Figure 8 and Figure 9 *w.r.t* different dimensionality *dim* of the latent space. The results demonstrate that the optimal dimensionality in the latent space depends on both the dataset and method itself. For SimuWechat dataset, the optimal setting for most of the methods is 64. For Wechat-Foursquare dataset, dimensionality varies from 16 to 256 but brings slight changes on cross-domain methods' performance, while for single-domain methods, SMF, 256 is too large. This can be explained that SMF can only utilize interaction data from the the target domain and the amount of data is too small for training parameters in the latent space.

To further evaluate our model, as there is no other public dataset available, we utilize the dense WeChat dataset to simulate cross-domain scenario, similar as [10]. Specifically, in [10], there is an operation to manually control density of data of a domain, which is also very similar with our operation. This is can be explained by two reasons. First, in real-world applications, cross-domain recommendation is performed to assist in building recommendation service for a target domain suffering from data sparsity issue. In other words, data in the target




(a) SimuWechat dataset

(b) Wechat-Foursquare dataset

Fig. 6. Top-10 recommendation performance with different $\epsilon$ in Beijing City

domain is always sparser. Second, with such experimental setting, cross-domain methods can outperform sing-domain ones significantly, which makes cross-domain solutions meaningful. Nevertheless, to make sure there is no bias in our operation, we add experiments with random splitting. Specifically, different from SimuWechat, we split users to two domains randomly and name this dataset as SimuWechat-Random. Then we follow the same experimental settings in Section 5.2 and compare the top-$K$ performance of our proposed and baseline methods, using HR@K as the metric. We present the experimental results in Figure 7, from which we can have very similar observations with SimuWechat dataset. First, our proposed CCMF still significantly outperforms all the baseline methods by 6.6%-24.4% on Shanghai and 6.5%-27.5% on Beijing. Second, the performance gain keeps steady with setting $K$ from 1 to 10. Lastly, our proposed CCMF still achieves similar or even better performance than Raw-CMF which cannot protect user privacy. In conclusion, our proposed method still works when adopting another rule to simulate cross-domain scenario.

We further analyze the impact of POI density. Although we utilize the dataset collected from two large cities in China, the utilized Foursquare-Wechat dataset is very sparse because only POIs recorded by both Foursquare and Wechat are reserved. Furthermore, during the pre-processing of the dataset, a lot of POIs arefi ltered out due to no records in the specific time period or no name for use of matching. The density offi nally utilized dataset of Foursquare-Wechat is about 0.18 POI/km$^2$ in Beijing and 0.63 POI/km$^2$ in Shanghai, while the SimuWechat dataset is multiple times denser in both two cities. In fact, experimental results on both Foursquare-Wechat and SimuWechat are very similar, demonstrate the efficacy of our proposed method, no matter in dense of sparse areas. Nevertheless, we choose a suburban district of Shanghai, Minhang District, to study the performance. Among all positive samples in the test set, wefi rstfi lter out those POIs lying within this district. Then we evaluate the performance of all models based on these POIs using average HR@1 and HR@10. Here the calculation of average HR is conducted by two steps. First, for each positive sample, wefi rst calculate a user-based average HR among users who have a real interaction with this positive sample. Then, a sample-based average is calculated among all filtered-out positive POIs. Then we evaluate the performance of all models based on these POIs using average HR@1 and HR@10. For HR@1, models considering confidence performs much better than those which do not.

CCMF and WC-CCMF achieve 0.18779 and 0.17352 of HR@1 respectively, outperforming their corresponding no-confidence models by 47.12% and 24.20%. For HR@10, CCMF and WC-CCMF achieve 0.4460 and 0.4412 of HR@10 respectively, outperforming their corresponding no-confidence models by 5.28% and 6.78%. These results further demonstrate that our method can work well in sparse areas.

In summary, further studies on the impact of hyper-parameters demonstrate that our method is not so sensitive to hyper-parameter, which means it is convenient to apply it to a new dataset. In addition, experiments on another simulated cross-domain dataset demonstrate the robustness of our proposed method.

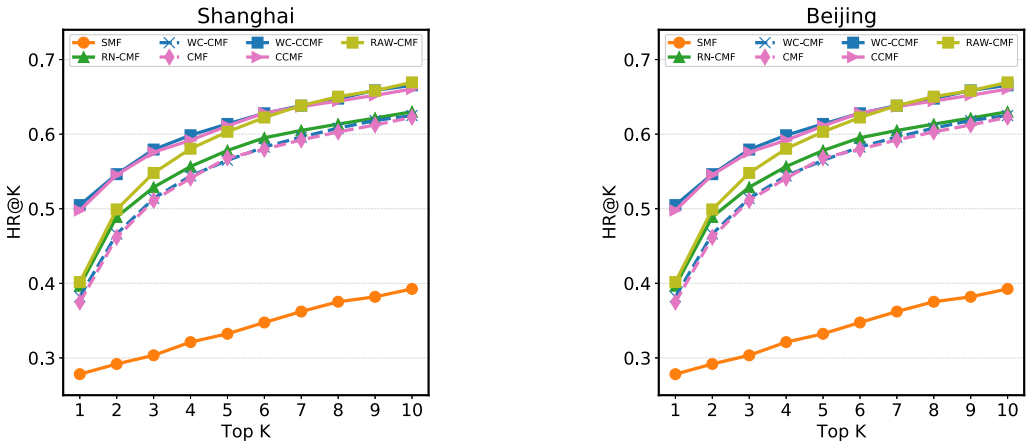



Fig. 7. Top-K recommendation performance comparison on SimuWechat-Random dataset in Shanghai and Beijing(K is set from 1 to 10)

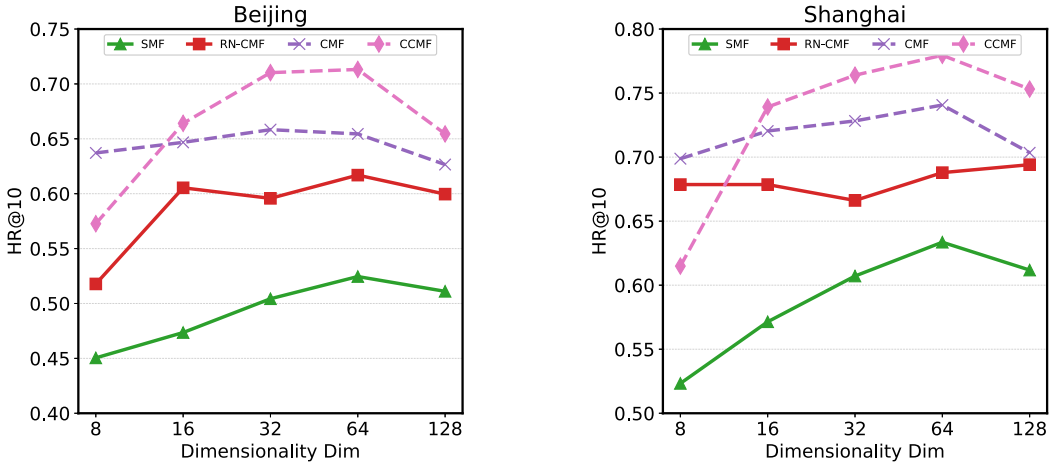



Fig. 8. Performance with different dimensionality *dim* on SimuWechat dataset

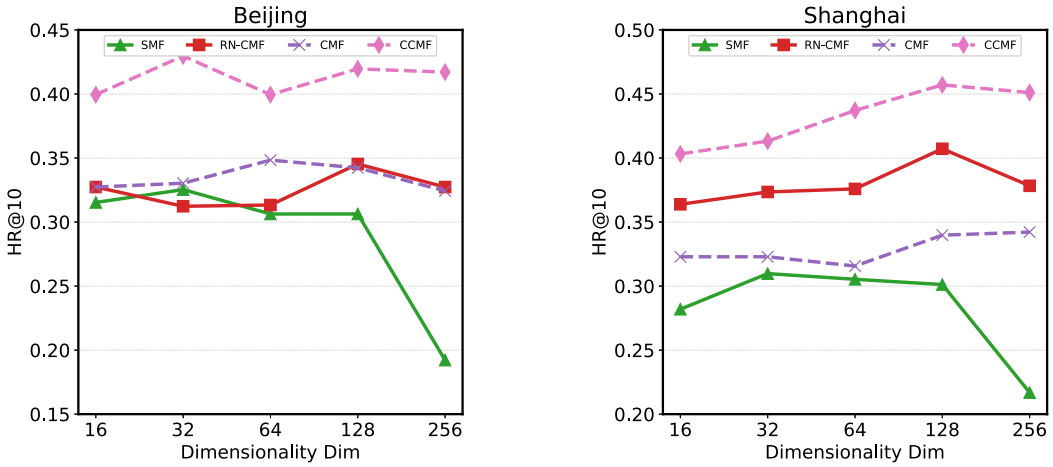



Fig. 9. Performance with different dimensionality *dim* on Wechat-Foursquare dataset

## 6 DISCUSSION AND CONCLUSION

In this paper, we focus on the risk of privacy leakage of data sharing in cross-domain location recommendation and propose a two-stage framework . Specifically, wefi rst apply obfuscation mechanism to raw data via adding noise to meet the criterion of differential privacy, and then perform a confidence-aware collective matrix factorization in the target domain to exploit the transferred obfuscated interaction matrix. Extensive experiments on real-world datasets demonstrate the efficacy of our proposed method in improving recommendation while protecting location privacy. Further studies show that our proposed method can also effectively alleviate data sparsity issue. Our proposed semantic-euclidean distance is demonstrated to be a better solution than euclidean distance to preserve utility of interaction data. Besides, wefi nd there is an interesting trade-off between utility and privacy.

It is worth mentioning that our work has strong characteristic of practicality and applicability in the real world since privacy preserving ubiquitous system has aroused more and more concerns [4, 42, 47]. In addition, in this work, we rely on matrix factorization, a widely used technique in recommender system, to design our framework for privacy-preserving cross-domain location recommendation. Actually, our novel design can be applied to other techniques such as neural network based recommendation [9, 14, 17].

Despite the novelty of our work, there are some points about that work we plan to address in the future. First, we plan to evaluate an online A/B test, which is one of the most important and significant future improvements. Second, we will try to collect more cross-domain datasets from various areas for further study. Lastly, we will study utilizing more types of auxiliary domains, such as social domain, in the future work.

## ACKNOWLEDGMENTS

This work was supported in part by The National Key Research and Development Program of China under grant 2017YFE0112300, the National Nature Science Foundation of China under 61861136003, 61621091 and 61673237, Beijing National Research Center for Information Science and Technology under 20031887521, and research fund of Tsinghua University - Tencent Joint Laboratory for Internet Innovation Technology.

## REFERENCES

[1] Gergely Acs and Claude Castelluccia. 2014. A case study: Privacy preserving release of spatio-temporal density in paris. In *SIGKDD*. 1679–1688.

[2] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *CCS*. 901–914.

[3] Arnaud Berlioz, Arik Friedman, Mohamed Ali Kaafar, Roksana Boreli, and Shlomo Berkovsky. 2015. Applying differential privacy to matrix factorization. In *RecSys*. 107–114.

[4] Ciaran Bryce, Marnix AC Dekker, Sandro Etalle, Daniel Le Métayer, Frédéric Le Mouël, Marine Minier, Joël Moret-Bailly, and Stéphane Ubéda. 2007. Ubiquitous privacy protection. In *UbiComp Workshop*.

[5] Joseph A Calandrino, Ann Kilzer, Arvind Narayanan, Edward W Felten, and Vitaly Shmatikov. 2011. " You Might Also Like:" Privacy Risks of Collaborative Filtering. In *IEEE S&P*. 231–246.

[6] Dong-Kyu Chae, Jin-Soo Kang, Sang-Wook Kim, and Jung-Tae Lee. 2018. CFGAN: A Generic Collaborative Filtering Framework based on Generative Adversarial Networks. In *CIKM*. 137–146.

[7] Konstantinos Chatzikokolakis, Ehab Elsalamouny, and Catuscia Palamidessi. 2017. Efficient utility improvement for location privacy. (2017), 308–328.

[8] Chaochao Chen, Ziqi Liu, Peilin Zhao, Zhou Jun, and Li Xiaolong. 2018. Privacy Preserving Point-of-Interest Recommendation Using Decentralized Matrix Factorization. In *AAAI*.

[9] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *DLRS*. 7–10.

[10] Paolo Cremonesi, Antonio Tripodi, and Roberto Turrin. 2011. Cross-domain recommender systems. In *ICDM Workshops*. 496–503.

[11] Bin Cuo, Jing Li, Vincent W Zheng, Zhu Wang, and Zhiwen Yu. 2018. CityTransfer: Transferring Inter-and Intra-City Knowledge for Chain Store Site Recommendation based on Multi-Source Urban Data. *IMWUT* 1, 4 (2018), 135.

[12] Yerach Doytsher, Ben Galon, and Yaron Kanza. 2011. Storing routes in socio-spatial networks and supporting social-based route recommendation. In *SIGSPATIAL Workshop on Location-Based Social Networks*. 49–56.

[13] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *TAMC*. 1–19.

[14] Chen Gao, Xiangnan He, Dahua Gan, Xiangning Chen, Fuli Feng, Yong Li, Tat-Seng Chua, and Depeng Jin. 2019. Neural Multi-Task Recommendation from Multi-Behavior Data. In *ICDE*.

[15] Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. 2013. Exploring temporal effects for location recommendation on location-based social networks. In *RecSys*. 93–100.

[16] Huiji Gao, Jiliang Tang, and Huan Liu. 2012. gSCorr: modeling geo-social correlations for new check-ins on location-based social networks. In *CIKM*. 1582–1586.

[17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborativefi ltering. In *WWW*. 173–182.

[18] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *SIGIR*. 549–558.

[19] Shen-Shyang Ho and Shuhua Ruan. 2011. Differential privacy for location pattern mining. In *SIGSPATIAL Workshop on Security and Privacy in GIS and LBS*. 17–24.

[20] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. 2007. Preserving privacy in gps traces via uncertainty-aware path cloaking. In *CCS*. 161–171.

[21] Jingyu Hua, Chang Xia, and Sheng Zhong. 2015. Differentially Private Matrix Factorization.. In *IJCAI*. 1763–1770.

[22] Bin Li. 2011. Cross-domain collaborativefi ltering: A brief survey. In *ICTAI*. 1085–1086.

[23] Chao Li, Balaji Palanisamy, and James Joshi. 2017. Differentially private trajectory analysis for points-of-interest recommendation. In *BigData Congress*. 49–56.

[24] Huayu Li, Yong Ge, Richang Hong, and Hengshu Zhu. 2016. Point-of-interest recommendations: Learning potential check-ins from friends. In *SIGKDD*. 975–984.

[25] Jing Li, Bin Guo, Zhu Wang, Mingyang Li, and Zhiwen Yu. 2016. Where to place the next outlet? harnessing cross-space urban data for multi-scale chain store recommendation. In *UbiComp*. 149–152.

[26] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*. 106–115.

[27] Xutao Li, Gao Cong, Xiao-Li Li, Tuan-Anh Nguyen Pham, and Shonali Krishnaswamy. 2015. Rank-geofm: A ranking based geographical factorization method for point of interest recommendation. In *SIGIR*. 433–442.

[28] Defu Lian, Cong Zhao, Xing Xie, Guangzhong Sun, Enhong Chen, and Yong Rui. 2014. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *SIGKDD*. 831–840.

[29] Defu Lian, Kai Zheng, Yong Ge, Longbing Cao, Enhong Chen, and Xing Xie. 2018. GeoMF++: Scalable Location Recommendation via Joint Geographical Modeling and Matrix Factorization. *TOIS* 36, 3 (2018), 33.

[30] Kwan Hui Lim, Jeffrey Chan, Christopher Leckie, and Shanika Karunasekera. 2015. Personalized Tour Recommendation Based on User Interests and Points of Interest Visit Durations. In *IJCAI*, Vol. 15. 1778–1784.

[31] Bin Liu, Deguang Kong, Lei Cen, Neil Zhenqiang Gong, Hongxia Jin, and Hui Xiong. 2015. Personalized mobile app recommendation: Reconciling app functionality and user privacy preference. In *WSDM*. 315–324.

[32] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. 2006. l-Diversity: Privacy Beyond k-Anonymity. In *ICDE*. 24.
[33] Frank McSherry and Ilya Mironov. 2009. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *SIGKDD*. 627–636.
[34] Andriy Mnih and Ruslan R Salakhutdinov. 2008. Probabilistic matrix factorization. In *NIPS*. 1257–1264.
[35] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *IEEE S&P*. 111–125.
[36] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. 2012. A random walk around the city: New venue recommendation in location-based social networks. In *SOCIALCOM-PASSAT*. 144–153.
[37] Huseyin Polat and Wenliang Du. 2005. Privacy-preserving top-n recommendation on horizontally partitioned data. In *WI*. 725–731.
[38] Daniele Riboni and Claudio Bettini. 2012. Private context-aware recommendation of points of interest: An initial investigation. In *PERCOM Workshops*. 584–589.
[39] Christopher Riederer, Yunsung Kim, Augustin Chaintreau, Nitish Korula, and Silvio Lattanzi. 2016. Linking users across domains with location data: Theory and validation. In *WWW*. 707–719.
[40] Luca Rossi and Mirco Musolesi. 2014. It's the way you check-in: identifying users in location-based social networks. In *WOSN*. 215–226.
[41] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborativefi ltering recommendation algorithms. In *WWW*. 285–295.
[42] Peter Andrew Shaw, Mateusz Andrzej Mikusz, Petteri Tapio Nurmi, and Nigel Andrew Justin Davies. 2018. Tacita-A Privacy Preserving Public Display Personalisation Service. *IMWUT* (2018).
[43] Reza Shokri, George Theodorakopoulos, Carmela Troncoso, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. 2012. Protecting location privacy: optimal strategy against localization attacks. In *CCS*. 617–627.
[44] Reza Shokri, Carmela Troncoso, Claudia Diaz, Julien Freudiger, and Jean-Pierre Hubaux. 2010. Unraveling an old cloak: k-anonymity for location privacy. In *CCS*. 115–118.
[45] Ajit P Singh and Geoffrey J Gordon. 2008. Relational learning via collective matrix factorization. In *SIGKDD*. 650–658.
[46] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.
[47] Zhen Tu, Runtong Li, Yong Li, Gang Wang, Di Wu, Pan Hui, Li Su, and Depeng Jin. 2018. Your apps give you away: distinguishing mobile users by their app usagefingerprints. *IMWUT* 2, 3 (2018), 138.
[48] Huandong Wang, Chen Gao, Yong Li, Gang Wang, Depeng Jin, and Jingbo Sun. 2018. De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice. In *NDSS*.
[49] Huandong Wang, Chen Gao, Yong Li, Zhi-Li Zhang, and Depeng Jin. 2017. From Fingerprint to Footprint: Revealing Physical World Privacy Leakage by Cyberspace Cookie Logs. In *CIKM*. 1209–1218.
[50] Hao Wang, Manolis Terrovitis, and Nikos Mamoulis. 2013. Location recommendation in location-based social networks using user check-in data. In *SIGSPATIAL*. 374–383.
[51] Jun Wang, Lantao Yu, Weinan Zhang, Yu Gong, Yinghui Xu, Benyou Wang, Peng Zhang, and Dell Zhang. 2017. Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *SIGIR*. 515–524.
[52] Toby Xu and Ying Cai. 2008. Exploring historical location data for anonymity preservation in location-based services. In *INFOCOM*. 547–555.
[53] Toby Xu and Ying Cai. 2009. Feeling-based location privacy protection for location-based services. In *CCS*. 348–357.
[54] Carl Yang, Lanxiao Bai, Chao Zhang, Quan Yuan, and Jiawei Han. 2017. Bridging collaborativefi ltering and semi-supervised learning: a neural approach for poi recommendation. In *SIGKDD*. 1245–1254.
[55] Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. Participatory cultural mapping based on collective behavior data in location-based social networks. *TIST* 7, 3 (2016), 30.
[56] Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhu Wang. 2013. A sentiment-enhanced personalized location recommendation system. In *ACM HT*. 119–128.
[57] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun Lee. 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. In *SIGIR*. 325–334.
[58] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Time-aware point-of-interest recommendation. In *SIGIR*. 363–372.
[59] Yongfeng Zhang. 2015. Incorporating phrase-level sentiment analysis on textual reviews for personalized recommendation. In *WSDM*. 435–440.
[60] Vincent Wenchen Zheng, Bin Cao, Yu Zheng, Xing Xie, and Qiang Yang. 2010. Collaborative Filtering Meets Mobile Recommendation: A User-Centered Approach.. In *AAAI*, Vol. 10. 236–241.

Received August 2018; revised November 2018; accepted January 2019