Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 3787089, 10 pages
https://doi.org/10.1155/2017/3787089

Research Article

Why You Go Reveals Who You Know: Disclosing Social Relationship by Cooccurrence

Feng Yi,[1,2] Hong Li,[1] Hongtao Wang,[1,2] and Limin Sun[1]

[1]*Beijing Key Laboratory of IOT Information Security, Institute of Information Engineering, CAS, Beijing 100093, China*
[2]*School of Cyber Security, University of Chinese Academy of Sciences, 19 A Yuquan Rd, Shijingshan District, Beijing 100049, China*

Correspondence should be addressed to Hong Li; lihong@iie.ac.cn

Received 25 July 2017; Accepted 2 October 2017; Published 31 October 2017

Academic Editor: Chaokun Wang

The popularity of location-based services (LBS) and the ubiquity of sensor device have resulted in rich spatiotemporal data. A large number of human behaviors had been recorded including cooccurrence which refers to the phenomenon that two people have been to the same places at the same time. These data enable attackers to infer people's social relationship based on their cooccurrences and many attack models were proposed. However, current attack models still cannot effectively address the following two challenges: *How to distinguish cooccurrences between acquaintances and strangers? What kind of cooccurrence contributes to strong social strength?* In this paper, we present a novel social relationship attack model—the Mobility Intention-based Relationship Inference (MIRI) model—which can solve the above two issues. Firstly, we extract mobility intentions and adopt them to characterize cooccurrences. A classification model is trained for attacking social relationship. The experimental results on two real-world datasets demonstrate that the proposed MIRI model can properly differentiate cooccurrences by simultaneously considering spatial and temporal features. The comparison results also indicate that MIRI model significantly outperforms state-of-the-art social relationship attack models.

1. Introduction

Nowadays, along with the rapid advances in Internet of Things (IOT) devices [1], especially with the popularity of social network services (SNSs) [2], the volume of human spatiotemporal data increases tremendously. A large number of spatiotemporal datasets are developed for various researches and many applications. In the last decade, lots of attack models [3–6] and privacy protection strategies [7] are proposed to protect users' privacy before releasing datasets to the public. However, most existing studies focus on how to protect individual privacy. There are few researches about protecting the social relationship of two users.

On the one hand, it is quite clear that two people are often seen together who have intimate social relationship, such as friends, colleagues, or family members. This phenomenon is known as *cooccurrence* where two people have been to the same places at the same time [8]. The cooccurrence is very common in people's daily life. On the other hand, the abundant spatiotemporal datasets enable adversaries to analyze social relationship based on cooccurrence, and it leads to the problem of social relationship protection.

If the social relationship of two people which should not be known is inferred by criminals, it leads to crimes or even two people might get hurt in real world. In 2014, a criminal knew a madam is an accountant employed by the owner of private companies in China. The criminal pretended to be the owner and cheated $30,000. Along with the increasing amount of human spatiotemporal data, attacking social relationship based on *cooccurrence* has become a focal point in social relationship research.

In the past few years, many attack models based on cooccurrence are proposed. These attack models employed spatial features of cooccurrence like location entropy [9] to infer social relationship. The basic idea is that cooccurrences at nonpublic places imply strong social strength and cooccurrences at public locations contribute less to social strength. These attack models have two problems. First of all, as we will see later, these attack models can be prevented by inserting fake records. Second, the precision of these attack models

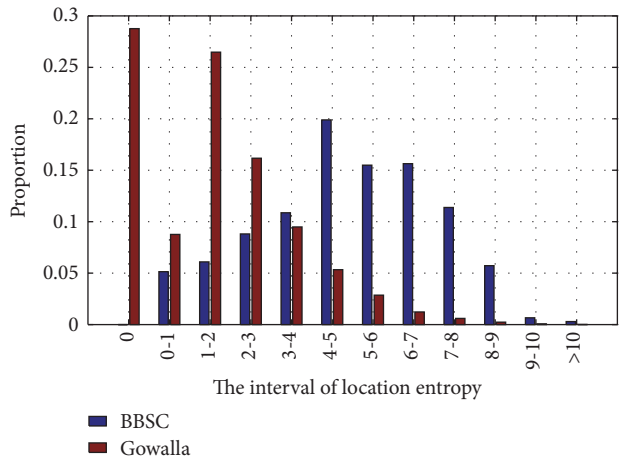


Figure 1: Distribution of location entropy in BBSC and Gowalla.

is not satisfied. Two people with an intimate relationship like friends may have cooccurred at public locations, such as shopping mall and cinema. It leads to an unreliable estimation of social strength.

What is more, existing models are based on self-reported data which involves explicit users' operations like Gowalla [8] and cell phone data [10]. In general, there are many cooccurrences belonging to acquaintances due to reporting motivation [11]. Hence, it is easy to distinguish cooccurrences of acquaintances from strangers in existing attack models. However, with the development of Internet of Things (IOT), rich passively collected spatiotemporal data are being produced by IOT's devices, such as the *Smart Card Data* (SCD) of public transport [12], the traffic surveillance [13], and bank notes [14]. The collection devices are deployed in public places and people have little awareness of them. Therefore, most cooccurrences are *coincidence* which means two strangers cooccurred.

What is more, all cooccurrences in passively collected spatiotemporal datasets took place at public locations. Figure 1 shows the distribution of location entropy in Beijing Bus Smart Card (BBSC) and Gowalla which are two real datasets used in our experiment. Locations with entropy less than 4 in Gowalla are over 90%, and 30% of the locations are checked by one user. In contrast, in BBSC dataset, more than 70% of the locations have entropy over 4. It means that colocations in BBSC are mainly public places. Hence, existing attack models can not be directly applied in the passively collected data. In summary, there are still two significant challenges in existing social relationship attack models: first, *how to distinguish cooccurrences of acquaintances from strangers* and, second, *what kind of cooccurrence contributes to strong social strength*.

To address the above two challenges, we propose a novel inference attack model called the Mobility Intention-based Relationship Inference (MIRI) model. We adopt mobility intention to analyze cooccurrence behaviors and exploit their contributions to social relationship. In general, human mobility is fundamentally driven by diverse mobility intentions, such as family party, shopping, and dining. If two people always cooccur for the same mobility intentions, they should be acquaintances with high probability. If two people cooccur for different mobility intentions, they are likely to be strangers and the cooccurrences should be coincidences. Moreover, it is obvious that social relationship of two people who often cooccur for shopping or entertainment is much closer compared to two people who only cooccur for commuting.

In the MIRI model, firstly, we obtain mobility intentions. Then, an Adaboost model is trained to map every cooccurrence to a mobility intention dyad. Finally, we train an SVM classifier based on mobility intention dyads to infer social relationship.

The major contributions of this paper are as follows:

(i) We analyze existing inference attack models and propose a method to protect social relationship.

(ii) We propose a novel social relationship attack model MIRI which can be applied in both self-reported and passively collected spatiotemporal data.

(iii) We conduct extensive experiments on two real spatiotemporal datasets: BBSC and Gowalla. With comparison of three state-of-the-art attack models—namely, Entropy-Based Model (EBM), Personal Global and Temporal (PGT) model, and Theme Aware social strength Inference approach (TAI)—the experiment results show that our proposed MIRI model is over 15% more precise than the baseline models.

The remaining of this paper is organized as follows. In Section 2 the problem is formally defined and current inference attack models are analyzed. Our proposed model is detailed in Section 3. In Section 4 extensive experiments are conducted. Finally, we conclude the paper with future directions in Section 5.

2. Problem Definition and Current Attack Models

2.1. Problem Definition. Given a spatiotemporal dataset of N users, a footprint of a user i is denoted by $o_k^i = (\mathrm{loc}_k^i, t_k^i)$ which states user i visited loc_k^i at time t_k^i. The subscript k denotes that o_k^i is the kth footprint of user i. The location loc_k^i is a geographic coordinate. The footprint history of user i is represented as a sequence of footprints: $O_i = (o_1^i, o_2^i, \ldots, o_n^i)$.

We say two users have a *cooccurrence* if they both visited a location at almost the same time, and the location of cooccurrence is called colocation.

Definition 1 (cooccurrence). The footprints $(\mathrm{loc}_r^i, t_r^i) \in O_i$ and $(\mathrm{loc}_s^j, t_s^j) \in O_j$ of users i and j can form a cooccurrence $c = (o_r^i, o_s^j)$ if they satisfy both spatial condition $\mathrm{dist}(\mathrm{loc}_r^i, \mathrm{loc}_s^j) < \delta$ and temporal condition $|t_r^i - t_s^j| < \tau$. δ and τ are distance and time thresholds, respectively, which is empirically decided by various application systems.

Let $C_{ij} = \{c_1, c_2, \ldots, c_w\}$ denote the set of w cooccurrences of users i and j. The formal definition of social relationship attack problem from spatiotemporal data is as follows.

Definition 2 (social relationship attack). Given a cooccurrence set C_{ij} of users i and j, the problem of social relationship attack is to infer whether they are acquaintances or strangers and how close their relationship is.

In this work, we use mobility intention to attack social relationship from spatiotemporal data. Formally, mobility intention can be defined by the following claim.

Definition 3 (mobility intention). The mobility intention refers to a common cause which can explain why a user appeared in location loc at time t.

2.2. Current Attack Models. In last decade, there are many research works which focus on attacking social relationship by cooccurrences. Most attack models are based on spatial features including location entropy and number of cooccurrences and so forth. The relation between social relationship and cooccurrence was first studied by Crandall et al. [15]. They found that the probability of a social tie increases sharply as the number of cooccurrences increases among Flickr users. The works [10, 16] obtained similar conclusions from cell phone data. Nonetheless, cooccurrences at different locations do not contribute equally to social relationship. Some other spatial features of cooccurrence were adopted to decide the contribution to social relationship.

Location entropy is a widespread used spatial feature. It takes into account both the number of users who are observed at the location and the relative proportions of their footprints. Let loc_k be a location and U_k be the set of all users who have footprints at loc_k. Let O_k be the set of footprints at loc_k and $O_{i,k}$ be the set of u_i's footprints at loc_k. The probability that a randomly picked footprint from O_k belongs to user u_i is $P_k(u_i) = |O_{i,k}|/|O_k|$ which is the total fraction of all footprints at location loc_k that are of user u_i. If we define this event as a random variable, then its uncertainty is given by the *location entropy*

$$H\left(\mathrm{loc}_k\right) = -\sum_{u_i \in U_k} P_k\left(u_i\right) \log P_k\left(u_i\right). \tag{1}$$

Location entropy with high value indicates many users have footprints at the location with equal proportion. Many public places, such as plazas, famous scenic spots, train stations, and stadiums, are popular to many visitors and have high value of location entropy. Conversely it will have low entropy if the distribution of footprints at a location is heavily concentrated on a few users. The private places, such as houses which are specific to a few people, have low value of location entropy. From definition of location entropy (1), it is clear that locations with high entropy usually have more cooccurrences than locations with low entropy.

Cranshaw et al. [9] firstly used location entropy to assign different contributions to cooccurrences. Their experiment results show that the precision of social relationship attack is greatly improved. Since then, many state-of-the-art inference models are based on location entropy. In [8], Pham et al. considered both location entropy and diversity of locations in cooccurrences. They proposed EBM model which is a linear regression model to attack social relationship of two users:

$$s_{ij} = \alpha D_{ij} + \beta F_{ij} + \gamma. \tag{2}$$

s_{ij} is social strength of u_i and u_j and α, β, and γ are optimal parameters. D_{ij} is a measurement for diversity of colocations and F_{ij} is closely related to entropy of colocations. Wang et al. [17] considered not only location entropy but also another two additional factors: individual mobility pattern and time gaps between two continuous cooccurrences. The result of the proposed model PGT is the product of the above three weights:

$$s_{ij} = \max\left\{w_{ij}^p\right\} \times \sum\left(w_{ij}^g \times w_{ij}^t\right). \tag{3}$$

w_{ij}^g is decided by location entropy and it plays a key role in the proposed attack model. Zhou et al. [18] proposed a TAI model which considers cooccurrence distribution on locations.

The problem of social relationship attack can not be solved by current location privacy protection techniques. Shahabi et al. [19] proposed a framework which can attack social relationship from privacy preserving spatiotemporal datasets. However, existing attack models are mainly based on location entropy. Furthermore, colocations with low entropy provide more information for social relationship attack. Hence, we can protect the social relationship by increasing the location entropy value of locations with low entropy. From (1), we know locations have high entropy if there are many users that visited the location with equal proportion. The locations with low entropy will have high value of entropy by inserting fake footprints on the colocations. For example, suppose there are five footprints for one user at a private location loc_k, and the entropy of loc_k is

$$H\left(\mathrm{loc}_k\right) = -1 \times \log 1 = 0. \tag{4}$$

If we add fifteen fake footprints with an equal allocation of three users, then the entropy of loc_k is

$$H^\star\left(\mathrm{loc}_k\right) = -4 \times \frac{1}{4} \log \frac{1}{4} = \log 4 = 2. \tag{5}$$

The entropy value of loc_k is much greater than the original value. Based on (2) and (3), the precision of EBM and PGT will decrease dramatically after modifying location entropy values. Therefore, we can protect social relationship by inserting fake footprints before releasing spatiotemporal datasets.

What is more, though the above inference models have shown how social relationship correlates to cooccurrence, the weights of these models have limited discernibility for considering only partial features of cooccurrence. The problem of distinguishing cooccurrences of acquaintances from strangers has not been studied in existing attack models. In contrast, as we will discuss later, our proposed MIRI model determines cooccurrence's contribution by corresponding mobility intentions which take into account various spatial and temporal features and overcomes the drawbacks in previous works.

3. Social Relationship Inference

3.1. Overview. As shown in Figure 2, MIRI model consists of three steps: (1) obtaining mobility intentions; (2) mapping every cooccurrence to a mobility intention dyad; (3) training an SVM classifier for attacking social relationship. After SVM classifier is trained, given cooccurrences of two people, then we can infer whether they are acquaintances or strangers.

We will explain how to obtain mobility intentions in Section 3.2. In Section 3.3, we train an Adaboost model with comprehensive feature engineering for mapping a footprint to a mobility intention. Then, the cooccurrence can be characterized by a mobility intention dyad. The problem of extracting cooccurrences from a spatiotemporal dataset is out of the scope of this paper, and we assume cooccurrences have been obtained before characterizing. In Section 3.4, a feature vector which is based on mobility intention dyads is constructed for attacking social relationship. If social strength between users in training dataset can be measured by continuous value like Katz score [8], we can train a linear regression model and can tell how close two users' relationship is. However, the training dataset of passively collected spatiotemporal data only tells if two users are acquaintances or not. Hence, in this paper, a binary SVM classifier is trained to infer whether two users are acquaintances or strangers. It is very easy to extend our binary classifier to linear regression model after obtaining continuous measurement in training dataset.

3.2. Mobility Intention Extraction. As mentioned before, mobility intentions can be used to infer social relationship. They are latent variables which can not directly be observed. First of all, we need to know how many and what kinds of mobility intentions hide in a spatiotemporal dataset. There are two ways to obtain mobility intentions: summarizing from auxiliary material and extracting from dataset.

First, we can obtain mobility intention from auxiliary materials, such as statistic materials or social network services (SNSs) provider. For instance, the 2009 National Household Travel Survey (NHTS) provides information of nation's inventory of daily travel including mobility intention (work, shopping, etc.) [20]. The Beijing Transport Institute release Beijing Transport Annual Report for public transport of Beijing to the public every year [21]. The annual reports provide the proportion of mobility intention in Beijing public transport. The proportion of seven mobility intentions in 2015 and 2016 are shown in Figure 3.

Many SNSs provide the information of location category. There were nine location categories in Gowalla dataset and five location categories in another popular check-in service Foursquare which include food, coffee, nightlife, fun, and shopping. Many researchers [22–24] consider location categories as activity categories, such as entertainment, food, and shopping. In general, the activity categories also can be considered as mobility intentions.

However, many spatiotemporal datasets have no auxiliary information. Furthermore, mobility intentions are not the same in different datasets. The second way to obtain mobility intention is extraction from spatiotemporal dataset. Many studies show that humans mobility follows simple reproducible patterns. The mobility patterns show a high degree of temporal and spatial regularity and can be considered as mobility intentions [14, 25]. For example, commuting which is a basic mobility pattern in many spatiotemporal dataset can be used to explain why a worker arrived at the work place around 9 a.m. on work days. Hence, we can use mobility patterns as mobility intentions.

In our former work [26], we use *Nonnegative Tensor Factorization (NTF)* to extract mobility patterns from a spatiotemporal dataset and consider them as mobility intentions. NTF is an effective tool for analyzing the interrelationship between spatial and temporal attributes for spatiotemporal dataset [27]. The `CANDECOMP/PARAFAC (CP)` decomposition algorithm [28] is a kind of widely used NTF and factorizes a tensor $\mathbf{Y}$ into a sum of component rank-one tensors in the following manner:

$$\mathbf{Y} \approx \sum_{r=1}^{R} \lambda_r \mathbf{Y}_r. \tag{6}$$

Every rank-one tensor $\mathbf{Y}_r$ is the outer product of three vectors

$$\mathbf{Y}_r = \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r. \tag{7}$$

$\mathbf{Y}_r$ is a mobility pattern and is considered as a mobility intention in our former work.

In order to extract mobility intentions, firstly, a three-dimensional tensor which is composed of `location-hour-day` $\mathcal{X} \in \mathbb{R}^{M \times H \times D}$ is constructed. We partition one day into h time bins with approximately equal time intervals. The element y_{r_i,t_j,d_k} of the three-way tensor $\mathcal{X} \in \mathbb{R}^{M \times H \times D}$ can be computed as

$$y_{r_i,t_j,d_k} = \frac{\text{Count}\left(r_i, t_j, d_k\right)}{\sum_{q=1}^{L} \text{Count}\left(r_q, t_j, d_k\right)}, \tag{8}$$

where r_i, t_j, and d_k are the index of the location, the time bin, and the day of month, respectively; L is the total number of locations; and $\text{Count}(r_i, t_j, d_k)$ is the number of users who appeared at location r_i at time t_j on d_kth days.

Then the tensor $\mathcal{X}$ is factorized into a linear combination of rank-one tensors $\mathbf{Y}_r$ through `CP` decomposition algorithm. After decomposition, we manually named the labels to summarize the mobility intention described by every rank-one tensor $\mathbf{Y}_r$. Here, every rank-one tensor $\mathbf{Y}_r$ is the outer product of three vectors $\mathbf{h}_r$, $\mathbf{l}_r$, and $\mathbf{d}_r$:

$$\mathbf{Y}_r = \mathbf{h}_r \circ \mathbf{l}_r \circ \mathbf{d}_r, \tag{9}$$

where $\mathbf{h}_r$, $\mathbf{l}_r$, and $\mathbf{d}_r$ can be considered inherent characteristics of mobility intention for *hour*, *location,* and *day*, respectively. What is more, these rank-one tensors can also be used to analyze spatiotemporal features which are applied to form mapping function in the following subsection.

We use m_i to denote the ith mobility intention and $\mathcal{M} = \{m_i \mid 1 \leqslant i \leqslant M\}$ to represent the set of M mobility intentions. After obtaining the mobility intentions, we can map every footprint to a mobility intention.

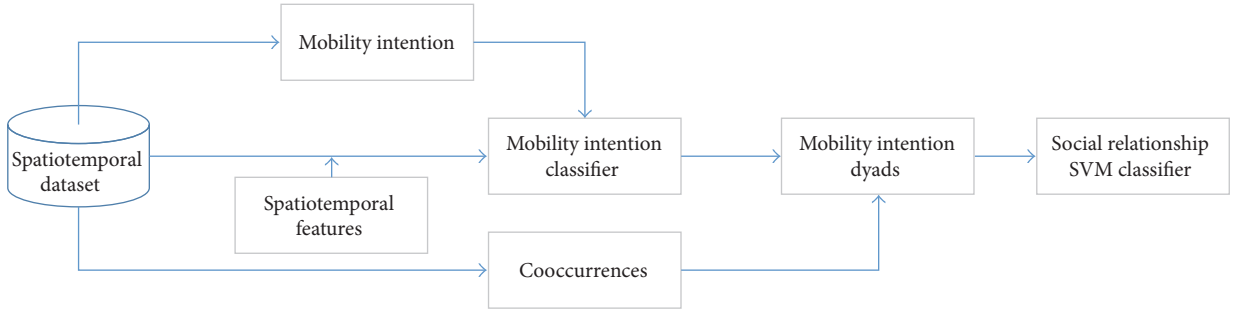


Figure 2: The framework of the proposed MIRI model.

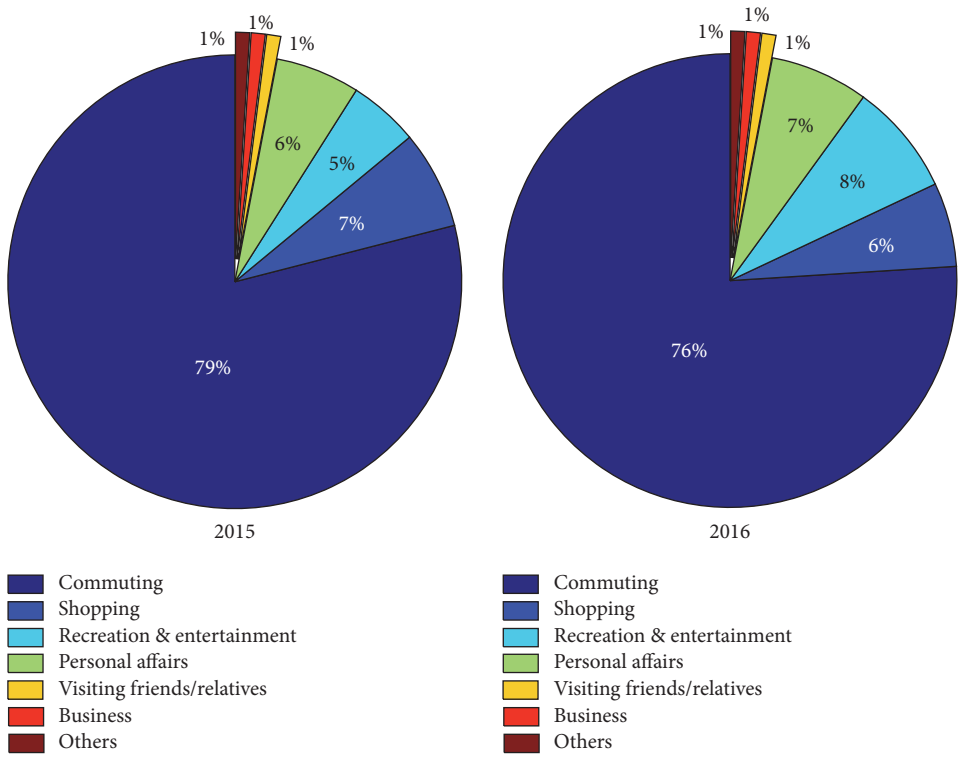


Figure 3: The proportion of mobility intention in Beijing public transport.

3.3. Inferring Mobility Intentions Dyads. In this subsection, we will detail how to map a cooccurrence $c = (o_r^i, o_s^j)$ to a mobility intention dyad. We consider every mobility intention $m_i \in \mathcal{M}$, $\leqslant i \leqslant M$, as one class. Each footprint $o_k^i = (\mathrm{loc}_k^i, t_k^i)$ corresponds to one mobility intention, in other words, belongs to a class. Then the mobility intention mapping can be considered as a multiclass classification problem.

In order to acquire good performance of multiclass classification, we perform a comprehensive feature engineering and model training. After analyzing vectors of rank-one tensor in the last subsection and considering experiment results in feature engineering, we propose three kinds of exploited and distinguishable features: spatial features, hour features, and day features. Let $o_k^i = (\mathrm{loc}_k^i, t_k^i)$ and $o_{k-1}^i = (\mathrm{loc}_{k-1}^i, t_{k-1}^i)$ be kth and $(k-1)$th footprint of user i, respectively, and $o_{k'}^i = (\mathrm{loc}_k^i, t_{k'}^i)$ be the last footprint of user i which also occurred at location loc_k^i. Then, the features can be defined as follows:

(i) Spatial features

(a) *Location entropy* is used to measure the popularity of a location and its definition is

$$H\left(\mathrm{loc}_k^i\right) = -\sum_i p_i \ln p_i, \tag{10}$$

where p_i is the probability that a randomly picked check-in from all check-ins at location loc_k^i belongs to user u_i.

(b) *Location type* indicates the category of location loc_k^i such as bar and mall. Location type can be obtained from LBS's application program interface (API), such as Google Places API (https://developers.google.com/places/?hl=zh-cn) or sina weibo API (http://open.weibo.com/).

(c) *Distance* refers to the distance between location loc_k and last location loc_{k-1}.

(ii) Hour features

(a) *Hours* is the hour of time t_k^i in a day denoted by $\{0, \ldots, 23\}$.

(b) *Stay time* is time interval between two continuous footprints of a user: $t_k^i - t_{k-1}^i$.

(c) *Time span* is time interval between two footprints which occurred at the same location loc_k^i: $t_k^i - t_{k'}^i$.

(iii) Day features

(a) *Day of week* is weekday of t_k^i. It is denoted by $\{0, \ldots, 6\}$ which means Sunday to Saturday.

(b) *Day of month* is day in month of t_k^i. It is denoted by $\{1, \ldots, 31\}$.

(c) *Day type* refers to the category of t_k^i. There are three categories: workday, short breaking holidays, and long holidays.

Finally, with the above features, we train an Adaboost model [29] to map o_k^i to a mobility intention m_k^i. Then, we can characterize a cooccurrence $c = (o_r^i, o_s^j)$ with a mobility intention dyad (m_r^i, m_s^j) through the trained Adaboost model.

In the last section, we discuss how to protect social relationship by changing entropy of location. However, this method does not work here. The entropy of location is a parameter in our Adaboost model. The Adaboost model can achieve similar classification results with much less tweaking of parameters. Hence, the changing of location entropy will not seriously affect multiclassification results and it can not protect the social relationship.

3.4. Social Relationship Inference. So far, we can build the inference model which is based on the mobility dyads. By adopting the Adaboost model, the cooccurrence sequence C_{ij} of user i and user j can be characterized as the following sequence of mobility intention dyads:

$$\left\langle \left(m_1^i, m_1^j\right), \left(m_2^i, m_2^j\right), \ldots, \left(m_w^i, m_w^j\right) \right\rangle, \quad m_k^i, m_k^j \in \mathscr{M},\ 1 \leqslant k \leqslant w. \tag{11}$$

In order to identify the weight of each of the mobility intention dyads, we construct a *mobility intention vector* $\mathbf{m}$ with the following $q(q+1)/2$ elements:

$$\left(\text{Count}\left(m_1, m_1\right), \ldots, \text{Count}\left(m_M, m_M\right), \text{Count}\left(m_1, m_2\right), \ldots, \text{Count}\left(m_{M-1}, m_M\right)\right)^T, \tag{12}$$

where $\text{Count}(m_r, m_s)$ is the number of mobility intention dyads (m_r^i, m_s^j).

For example, if there are three mobility intentions m_1, m_2, and m_3 extracted from a spatiotemporal dataset. Given the mobility intention pairs of u_i and u_j as follows:

$$\langle (m_1, m_1), (m_1, m_1), (m_1, m_1), (m_1, m_2), (m_2, m_1), (m_1, m_3), (m_2, m_3) \rangle \tag{13}$$

the mobility intention vector $\mathbf{m}_{ij}$ is

$$\mathbf{m}_{ij} = (3, 0, 0, 2, 1, 1)^T. \tag{14}$$

From the feature vector $\mathbf{m}_{ij}$, we see that the two users cooccurred three times for the same mobility intention m_1, cooccurred twice with one for m_1 and the other for m_2, and cooccurred once with left different mobility intention dyads.

Then, we can adopt virtually any existing binary classifier algorithm to distinguish acquaintances from strangers. In this paper, we use SVM as our binary classifier. After training is finished, given a pair of users and their cooccurrences, we first map them to mobility intentions dyads through the Adaboost model. Then we construct the mobility intention vector. Finally, we can infer whether the two users are acquaintances or not by the trained classifier.

4. Experiment and Analysis

In this section, we first describe experimental settings including the dataset and experimental environment. Next, baselines and evaluation metric which are used to evaluate the performance are discussed. Finally, we compare MIRI model with several state-of-the-art models to demonstrate its effectiveness.

4.1. Settings. We use two real-world datasets in our experiment: Gowalla [2] and BBSC. The Gowalla dataset is a publicly available check-in dataset which was collected from February 2009 to October 2010. It is a self-reported dataset and consists of two different sets. The first one is spatiotemporal data with 6,442,890 check-ins from 107,092 users, and the format of every check-in is as follows: <user ID, latitude, longitude, timestamp, location ID>. The other one is a social graph of friendships among users and has 950,327 friend pairs which serve as the ground truth. We mine cooccurrences from the first set and form the first experimental data collection with corresponding relationship information of user pairs in the second set. The BBSC collects prepaid smart card records for public transportation in Beijing, China, and is a passively collected dataset. The format of each card record is as follows: <card ID, line ID, bus station, swap card time>. The geographic coordinate of each bus station can be obtained from Google Places API. We obtained a dataset with 275,951,094 bus transaction records with about 16,161,460 cards from October 1, 2014, to October 31, 2014. We identified 412 card users and 2,796 friend pairs among these card users, and the cooccurrences of these 412 card users and corresponding relationship are formed to the second experimental data collection.

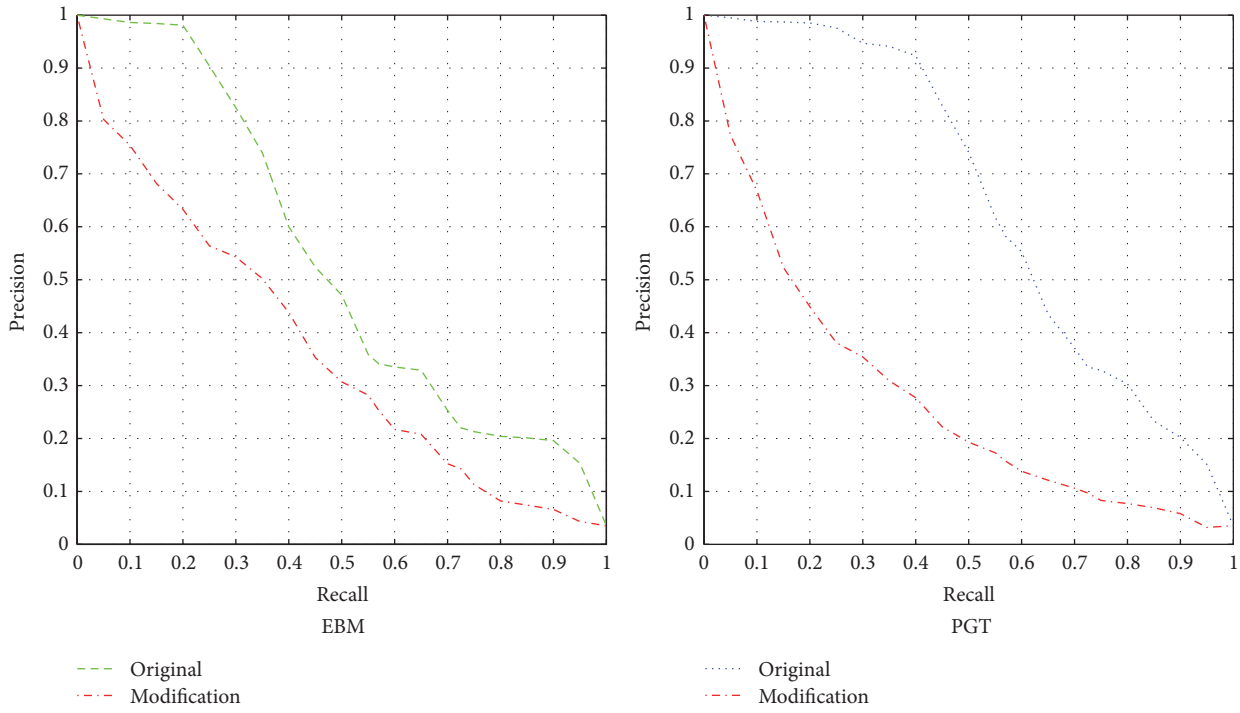


Figure 4: Comparison of original Gowalla dataset and its modification.

Every experimental data collection is divided into two subsets: training set and test set. We construct mobility intention-based relationship attack model and other baseline models on the training set and test the performance of all attack models on the test set. In order to verify the effectiveness of tensor decomposition, we extract mobility intentions from BBSC and Gowalla.

4.2. Methodology. The precision-recall curve is used to measure the precision of our model and make comparison with other baseline models. Let TR denote the set of ground truth friend pairs in the test set and MR be the set of friend pairs reported by a social relationship inference model. The precision and recall are defined as

$$\text{Precision} = \frac{|\text{TR} \cap \text{MR}|}{|\text{MR}|}, \qquad \text{Recall} = \frac{|\text{TR} \cap \text{MR}|}{|\text{TR}|}. \tag{15}$$

Three baseline models are chosen for performance comparison: EBM [8], PGT [17], and TAI [18]. EBM is a state-of-the-art social relationship inference model. It is an Entropy-Based Model and two major factors are considered: location diversity which is measured by Renyi entropy and weighted frequency which is based on location entropy. PGT is an extension of EBM by considering two additional factors: personal factor which indicates an individual user's probability to visit a certain location and temporal factor which considers the time gaps between consecutive cooccurrences. TAI is a probability model which is based on LDA and its performance is close with numbers of "topic themes" and spatiotemporal windows. The topic themes are similar to the mobility intention of our proposed MIRI.

4.3. Results

4.3.1. Protecting Social Relationship by Increasing Location Entropy. In Section 2.2, we analyze current social relationship attack models and conclude that social relationship can be protected by increasing location entropy value. We will use some experiments to illustrate the proposed protection method in this subsection.

The location entropy values of locations with location entropy less than 3 are modified to 3 in Gowalla dataset. Then EBM model and PGT model are trained and tested. Figure 4 shows the performance comparison of two inference models on original dataset and modified dataset. After location entropy modification, the performance of both attack models decreases dramatically. The prediction precisions when recall is greater than 0.2 and less than 0.8 reduce by 35% to 89% and 120% to 296% for EBM model and PGT model, respectively. These results show that increasing location entropy is an efficient method to protect social relationship.

4.3.2. Performance Comparison with Baseline Models. In Figure 5, the precision-recall curves of baseline models and MIRI on two datasets illustrate that MIRI performs the best among all comparison models on two datasets. In detail, for the Gowalla dataset in Figure 5(a), MIRI outperforms PGT by 5% to 20% in precision for considering comprehensive spatiotemporal features. For BBSC dataset in Figure 5(b), the

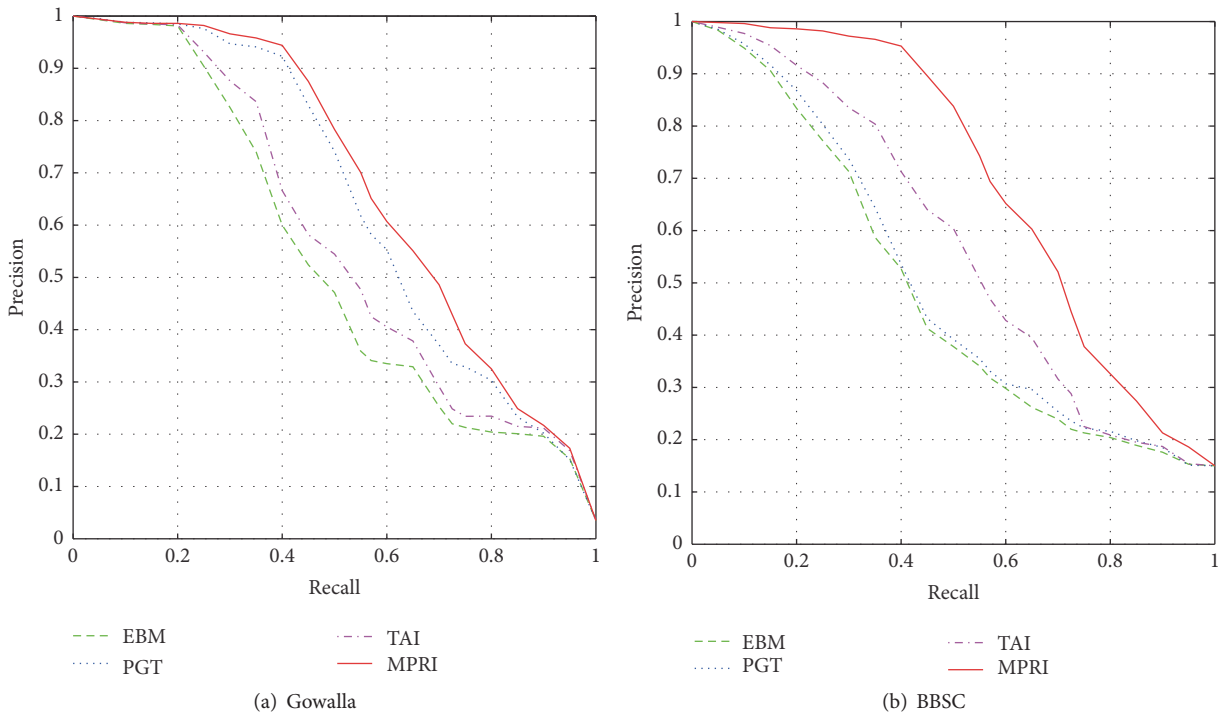


Figure 5: Comparison with the state-of-the-art models.

performance of EBM and MIRI on BBSC is not as good as that on Gowalla. There are two reasons for this. The first is that all cooccurrences of BBSC occurred at public places. The second is that there are more coincidences in BBSC than in Gowalla. Although the coincidence weights of EBM and PGT are very small, the amount of coincidence is large and it leads to unreliable estimation of social relationship. However, MIRI have stable performance on both datasets. The possible reason is that mobility intention is introduced and plays a key role in differentiating cooccurrences between acquaintances and strangers. Our experiments demonstrates that the use of mobility intention improves the precision of social relationship attack.

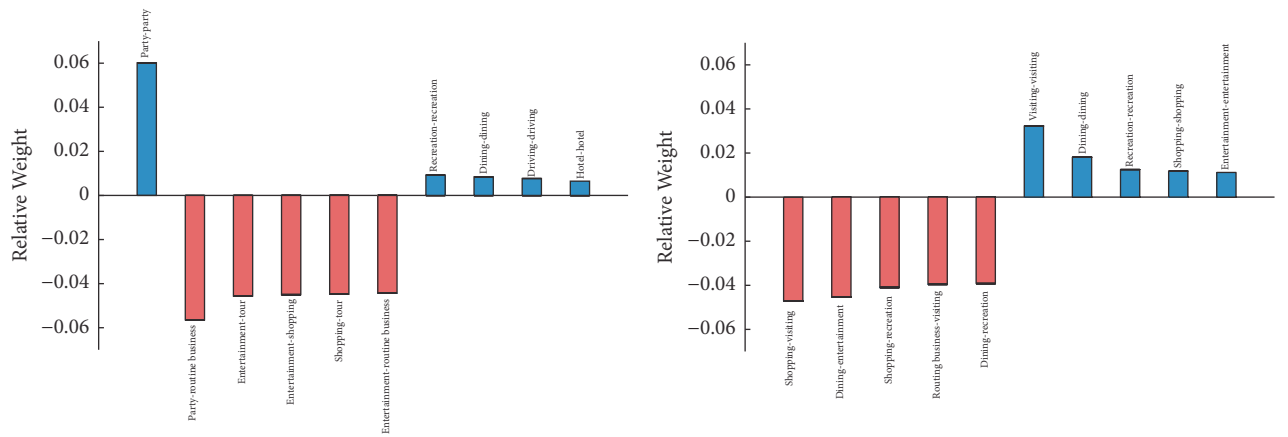
4.3.3. Contribution of Mobility Intention Dyads. After tensor decomposition, we extract ten and seven mobility intentions from Gowalla dataset and BBSC dataset, respectively. The seven mobility intentions from BBSC dataset are commuting, shopping, visiting relatives/friends, dinning, recreation, entertainment, and routing business. With comparison with the mobility intentions in Figure 3, the mobility intentions extracted from dataset are consistent with the mobility intentions in Beijing Transport Annual Report [21]. This result verifies the effectiveness of extracting mobility intentions form spatiotemporal dataset by tensor decomposition method.

The top 5 positive and negative weights of mobility intention dyads are illustrated in Figure 6. The weights presented in the figure are normalized by the total sum of the absolute value of all the weights to reflect a pair of mobility intentions' relative importance. From the figure, it is clear that negative weights are generally more important than positive weights and they are not quite different. This provides evidence that two people are most likely strangers when they cooccurred for different mobility intentions. The positive weights are very different. For Gowalla dataset in Figure 6(a), the mobility intention pair *party-party* which means two people cooccurred for family party is a dominant indicator for acquaintances. On the contrary, there are no single dominant positive weights in BBSC, probably due to the fact that there are no private locations in BBSC. In general, the positive weights are very different which mean different contribution to social relationship both in Figure 6(a) and in Figure 6(b). This result shows that some cooccurrences with high positive value weights imply strong social strength.

5. Conclusions

In this paper, we discuss existing works about attacking social relationship by cooccurrences and propose a method by inserting fake footprints to prevent this kind of attack. We have proposed a novel social relationship attack model called MIRI, which considers the mobility intention dyads as features and adopts a classifier to infer whether two persons are acquaintances or not. Extensive experiments indicate that the proposed model significantly outperforms existing social relationship attack models and can be applied in both self-reported datasets and passively collected datasets.

This work leads to an important future research. How can we protect our social relationship under MIRI attack? What kind of operations do we need before releasing spatiotemporal datasets? We believe that this work provides a necessary step towards addressing such questions.



(a) Gowalla (b) BBSC

Figure 6: Top 5 negative and positive mobility intention dyads for attacking social relationship.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 61572231), the Major R&D Plan of Beijing Municipal Science & Technology Commission (Grant no. Z161100002616032), the Security Detection and Supervision for High Level Bio-Safety Laboratory Control System (Grant no. CXJJ-16Z234), and the National Defense Basic Research Program of China (Grant no. JCKY2016602B001).

References

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): a vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[2] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1082–1090, ACM, August 2011.

[3] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proceedings of the IEEE Symposium on Security and Privacy (SP '08)*, pp. 111–125, IEEE, Oakland, Calif, USA, May 2008.

[4] Narayanan Arvind and Shmatikov Vitaly, "De-anonymizing social networks," in *Proceedings of the In 30th IEEE Symposium on Security and Privacy (S&P 2009)*, pp. 173–187, Oakland , California, USA, 2009.

[5] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: Using social networks as a side-channel," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS 2012*, pp. 628–637, usa, October 2012.

[6] J. Su, A. Shukla, S. Goel, and A. Narayanan, "De-anonymizing Web Browsing Data with Social Networks," in *Proceedings of the the 26th International Conference*, pp. 1261–1269, Perth, Australia, April 2017.

[7] C. Dwork, "Differential privacy: a survey of results," in *Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25–29, 2008. Proceedings*, vol. 4978 of *Lecture Notes in Computer Science*, pp. 1–19, Springer, Berlin, Germany, 2008.

[8] H. Pham, C. Shahabi, and Y. Liu, "EBM - An entropy-based model to infer social strength from spatiotemporal data," in *Proceedings of the 2013 ACM SIGMOD Conference on Management of Data, SIGMOD 2013*, pp. 265–276, usa, June 2013.

[9] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*, pp. 119–128, ACM, September 2010.

[10] N. Eagle, A. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 106, no. 36, pp. 15274–15278, 2009.

[11] E. Malmi, T. M. T. Do, and D. Gatica-Perez, "Checking in or checked in: Comparing large-scale manual and automatic location disclosure patterns," in *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia, MUM 2012*, deu, December 2012.

[12] M. Itoh, D. Yokoyama, M. Toyoda, Y. Tomita, S. Kawamura, and M. Kitsuregawa, "Visual fusion of mega-city big data: An application to traffic and tweets data analysis of Metro passengers," in *Proceedings of the 2nd IEEE International Conference on Big Data, IEEE Big Data 2014*, pp. 431–440, usa, October 2014.

[13] R. Cucchiara, "Multimedia surveillance systems," in *Proceedings of the 3rd ACM International Workshop on Video Surveillance and Sensor Networks, VSSN 2005*, pp. 3–10, sgp.

[14] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[15] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring social ties from geographic coincidences," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 107, no. 52, pp. 22436–22441, 2010.

[16] M. Yu, W. Si, G. Song, Z. Li, and J. Yen, "Who were you talking to - Mining interpersonal relationships from cellphone network data," in *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014*, pp. 485–490, chn, August 2014.

[17] H. Wang, Z. Li, and W.-C. Lee, "PGT: Measuring Mobility Relationship Using Personal, Global and Temporal Factors," in *Proceedings of the 14th IEEE International Conference on Data Mining, ICDM 2014*, pp. 570–579, chn, December 2014.

[18] N. Zhou, X. Zhang, and S. Wang, "Theme-aware social strength inference from spatiotemporal data," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 8485, pp. 498–509, 2014.

[19] C. Shahabi, L. Fan, L. Nocera, L. Xiong, and M. Li, "Privacy-preserving inference of social relationships from location data: A vision paper," in *Proceedings of the 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS 2015*, usa, November 2015.

[20] Santos Adelia, McGuckin Nancy, Yukiko Nakamoto Hikari, Danielle Gray, and Liss Susan, "Summary of travel trends: 2009 national household travel survey," Federal Highway Administration, National Household Travel Survey, U.S. Department of Transportation, 2009, http://nhts.ornl.gov/2009/pub/stt.pdf.

[21] Yongyan Quan, Jifu Guo, Xian Li, and Gengchen Wang, "Beijing transport annual report," The Beijing Transport, 2017, http://www.bjtrc.org.cn/JGJS.aspx?id=5.2Menu=GZCG.

[22] C. Yu, Y. Liu, D. Yao et al., "Modeling user activity patterns for next-place prediction," *IEEE Systems Journal*, vol. PP, no. 99, 2015.

[23] A. Likhyani, D. Padmanabhan, S. Bedathur, and S. Mehta, "Inferring and exploiting categories for next location prediction," in *Proceedings of the 24th International Conference on World Wide Web, WWW 2015*, pp. 65-66, ita, May 2015.

[24] J. Ye, Z. Zhu, and H. Cheng, "What's your next move: User activity prediction in location-based social networks," in *Proceedings of the 13th SIAM International Conference on Data Mining, SMD 2013*, pp. 171–179, May 2013.

[25] S. Hasan, X. Zhan, and S. V. Ukkusuri, "Understanding urban human activity and mobility patterns using large-scale location-based data from online social media," in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing (UrbComp '13)*, pp. 6:1–6:8, ACM, August 2013.

[26] F. Yi, H. Li, H. Wang, H. Wen, and L. Sun, "Mobility Intention-Based Relationship Inference from Spatiotemporal Data," in *Wireless Algorithms, Systems, and Applications*, vol. 10251 of *Lecture Notes in Computer Science*, pp. 871–876, Springer International Publishing, Cham, 2017.

[27] Z. Fan, X. Song, and R. Shibasaki, "CitySpectrum: A non-negative tensor factorization approach," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp 2014*, pp. 213–233, usa, September 2014.

[28] H. A. L. Kiers, "A three-step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity," *Journal of Chemometrics*, vol. 12, no. 3, pp. 155–171, 1998.

[29] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research (JMLR)*, vol. 4, no. 6, pp. 933–969, 2004.