Inferring Friendship from Check-in Data of Location-Based Social Networks

Ran Cheng[‡*], Jun Pang[*†], Yang Zhang[†]

[*]Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg
[†]Faculty of Science, Technology and Communication, University of Luxembourg, Luxembourg
[‡]Department of Computer Science and Technology, Shandong University, China

***Abstract*—With the ubiquity of GPS-enabled devices and location-based social network services, research on human mobility becomes quantitatively achievable. Understanding it could lead to appealing applications such as city planning and epidemiology. In this paper, we focus on predicting whether two individuals are friends based on their mobility information. Intuitively, friends tend to visit similar places, thus the number of their co-occurrences should be a strong indicator of their friendship. Besides, the visiting time interval between two users also has an effect on friendship prediction. By exploiting machine learning techniques, we construct two friendship prediction models based on mobility information. The first model focuses on predicting friendship of two individuals with only one of their co-occurred places' information. The second model proposes a solution for predicting friendship of two individuals based on all their co-occurred places. Experimental results show that both of our models outperform the state-of-the-art solutions.**

I. Introduction

Mobility is one of the most common human behaviors, understanding it can result in many appealing applications, such as urban planning, public transportation system design, epidemiology, etc. It is evident that social relationships can affect human mobility, for example, friends tend to visit similar places or one visits some places recommended by his friends. On the other hand, human mobility also has influence on social connections, e.g., two people are more likely to become friends if their mobility profile are similar.

In the past, obtaining people's mobility information is considered as an obstacle for related study. Researchers have recruited a group of people to monitor their GPS-enabled devices [1], [2], [3] or conducted questionnaires [2]. These methods always end up with a biased dataset because of the limited number of people or an imprecise dataset considering people's memory pattern [2]. With the development of GPS-enabled devices, such as smart phones and tablets, people begin to share more of their mobility information on their social networks. Moreover, a new type of social network services has emerged, namely Location-based social networks (LBSNs). In LBSNs, a user can share his location information (called check-in) to get some reductions and engage in social games. Popular LBSNs include *Yelp*, *Instagram* and *Foursquare*.

Researchers have utilized the check-in data from LBSNs to understand human mobility [4], [5], [6], [7], [8], [9]. The research can be roughly partitioned into two directions, one is using social information to model human mobility, the other is using users' check-ins to analyze their social relationships. In this paper, we focus on the latter, concretely, we aim to use two users' mobility information to predict whether they are friends. Friendship prediction has a lot of applications, such as friends recommendation in social networks and targeted marketing.

Intuitively, friends tend to visit same places due to similar interests. This is known as social homophily [10]. Friends may visit same places together or separately. The former can refer to friends hanging out together while the latter may be an evidence of place recommendation. If two people visit many same places, it may indicate that they are probably friends. Similarly, if the number of visits for two people together to places is large, it is also a good indication that they are friends. On the other hand, the visiting time interval of two people can also have influence on their relationship. If two check-ins happen at roughly the same time, the corresponding users probably visit the place together with intention. If the check-in time interval is about a short time period (e.g., one or two months), these two visits can be considered to be linked because of place recommendations between friends. Based on these intuitions, we develop two models for friendship prediction.

Contributions. We tackle two sets of friendship prediction problems based on location information in this work. In the first one, we solve the problem of predicting whether two users are friends, given the check-in information at only one location that they both visit. We formalize the problem into a binary classification and apply machine learning technique to solve it. For any given two users and one of their co-occurred locations, we consider features that are related to check-in numbers, time intervals and location popularity. In the second problem, we aim to predict two users' relationship when all their check-in information are available. For each pair of users, we extract several features covering their co-occurrences and time difference on mobility to train our classifier. Through extensive experiments on a real-life LBSNs dataset, we have shown that our two solutions outperform the state-of-the-art solutions.

Organization. The rest of this paper is organized as follows. In Section II, we give a brief overview on related works. In Section III, we introduce some notions and the dataset.

Section IV presents the first friendship prediction model and our experiment results. The demonstration of the second friendship prediction model is given in Section V. Section VI concludes the paper with some future work.

II. Related Work

Many works aiming at predicting friendship from spatial-temporal information have been published during the last several years. Li et al. [1] extracted users' visiting trajectories and stay points from location information and represented the set of stay points as a hierarchical graph where each layer clusters the stay points into several spatial clusters divisively. The pair of users who share similar spatial clusters on a lower layer has stronger similarity, and similarity is used to indicate friendships between them. Along this direction, Chen et al. [8], [9], [11] have used trajectory pattern to represent user mobility profiles and proposed several metrics to measure the similarity among user mobility profiles with a tool support [12].

Eagle et al. [2] conducted a study to observe 94 students and faculty on their mobile phones for nine months. Through the analysis on the dataset, they found out that two people visiting the same place at roughly the same time is a strong indicator that they are friends. Particularly, the indicator becomes even stronger when the visits happen at non-working time and locations. Crandall et al. [5] have discovered the similar result, i.e., the larger the number of locations two people co-occurred at roughly the same time, the higher the probability that they are friends. In addition, they proposed a probabilistic model to predict friendship. However, the model does not fit the real life scenario since they made the assumption that each user only has one friend.

Cranshaw et al. [3] formalized the problem into a binary classification and extracted a large number of features including the spatial and temporal range of the set of co-locations, location diversity and specificity, and structural properties to train the friendship predictor. In addition, they propose a notion namely location entropy to characterize a location's popularity which we will use in our work. Similar to [3], Chang and Sun [6] also utilized machine learning classifier for friendship prediction. In their problem set, they only have one common location's information that two users have been to. In their solution, they only considered very simple features. We tackle the same problem in our first model. By considering more meaningful features, our model outperforms theirs significantly. Pham et al. [7] proposed an entropy-based model (EBM) to estimate social strength which also leads to friendship prediction. They extracted two factors for each pair of users to train their model. The result in [7] shows that EBM outperforms all the above mentioned models. We tackle the same problem in our second model. By considering time in a more general way, we are able to achieve better result than EBM. Some recent works on friendship prediction based on location information include [13], [14].

III. Preliminaries

In this section, we define some basic notions: *co-occurrence* and *co-location*, *location entropy* and *time interval sequence*, and introduce the check-in dataset that we use in experiments.

A. Notations

Given a set of users denoted by $\mathcal{U} = \{u_1, \ldots, u_n\}$, the check-in dataset $\mathcal{C}$ records which user appeared at what location at what time, in the form of $\langle u, t, \ell \rangle$ ($u \in \mathcal{U}, t \in \mathcal{T}$ and $\ell \in \mathcal{L}$) where $\mathcal{T} = \{t_1, \ldots, t_n\}$ represents a set of timestamps and $\mathcal{L} = \{\ell_1, \ldots, \ell_n\}$ represents a set of locations. To define locations, we partition the surface of the earth into a set of $s \times s$ grid-like cells which span s degrees of latitude and longitude. Each cell represents a location.

We use a sequence $C_u = (\langle u, t_1, \ell_1 \rangle, \ldots, \langle u, t_n, \ell_m \rangle)$ to denote all the check-ins made by user u where $t_i < t_{i+1}$ ($1 \leq i < m$). In the sequel, for the sake of simplicity, we use $c \in C_u$ to denote that c is a check-in of user u. Moreover, we use a sequence $C_u^\ell = (\langle u, t_1, \ell \rangle, \ldots, \langle u, t_n, \ell \rangle)$ to represent the set of all the check-ins conducted by user u at a certain location ℓ, where $t_i < t_{i+1}$ ($1 \leq i < m$).

B. Co-occurrence and Co-location

We say that two users have a *co-occurrence* at a location if they have both been to this location.

Definition 1 (Co-occurrences & co-locations): Given two users $u \in \mathcal{U}$ and $u' \in \mathcal{U}$ and their check-in sequences C_u and C'_u, we say that they co-occurred (or have a co-occurrence) at a location $\ell \in \mathcal{L}$ if

$$\exists \langle u, t, \ell_1 \rangle \in C_u, \langle u', t', \ell_2 \rangle \in C_{u'} \text{ such that } \ell_1 = \ell \wedge \ell_2 = \ell.$$

The location ℓ is called a *co-location* of users u and u'. The number of co-occurrences of users u and u' at the co-location ℓ, denoted by $|C^\ell_{u,u'}|$, is defined as the following.

$$|C^\ell_{u,u'}| = min(|C^\ell_u|, |C^\ell_{u'}|)$$

For example, suppose that u checked in at four locations ℓ_1, ℓ_2, ℓ_3 and ℓ_4, and user u' checked in at three locations ℓ_2, ℓ_4 and ℓ_6. We say that users u and u' co-occurred (or have co-occurrences) at ℓ_2 and ℓ_4. Here, ℓ_2 and ℓ_4 are called co-locations of u and u'. In addition, if u checked in at location ℓ_2 for 4 times and user u' for 6 times, we say that they have 4 co-occurrences at ℓ_2. Note that in our work, whether two users co-occurred at a location does not depend on the time when they visit the same location. This is different from many works in the literature [2], [3], [5], [7], in which they set a time interval parameter τ and give the definition that two users co-occurred at a location only if their visiting time interval was smaller than τ. In fact, by studying a LBSN dataset [4], we find out that nearly 30% of pairs of friends have a minimum check-in time interval larger than 30 days. Therefore, if we set τ to a relatively small value, a lot of useful information belonging to friends will be neglected. Instead, our definition of co-occurrences captures more information and, as a consequence, leads to better prediction results.

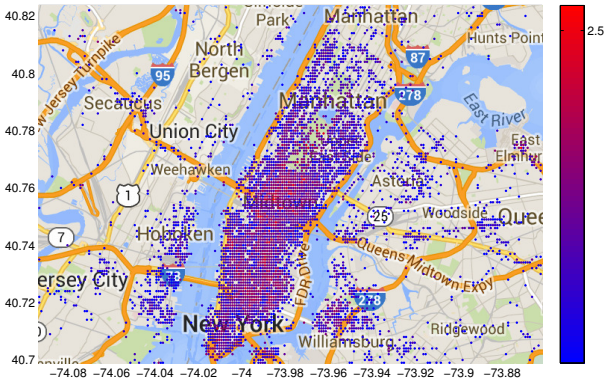


Fig. 1. The heat map of location entropy in New York.

C. Location Entropy

Location entropy, as introduced in [3], is a metric to quantify the popularity of a location. Intuitively, a location is popular if it has been visited by a large number of people, while a location is not popular (i.e., private) if it has a limited number of visitors.

Definition 2 (Location entropy): Given a location $\ell \in \mathcal{L}$, we can calculate its location entropy as follows:

$$H_\ell = -\sum_{u \in \mathcal{U}} p_\ell(u) \cdot \log p_\ell(u), \text{ where}$$

$$p_\ell(u) = \frac{\mid \{\langle u, t, \ell' \rangle \in C_u \mid \ell' = \ell\} \mid}{\mid \cup_{u' \in \mathcal{U}} \{\langle u', t', \ell'' \rangle \in C_{u'} \mid \ell'' = \ell\} \mid}$$

A large value of location entropy indicates a popular location (e.g., train stations and shopping malls), and a small value of location entropy indicates a private location (e.g., private offices and homes). Figure 1 presents the heat map of New York with respect to location entropy. Every point refers to a location in a shape of grid-like cell, and each cell spans 0.001° of latitude and longitude. From the heat map, we can see that the middle town, where lots of popular places such as The Empire State Building, Times Square, Rockefeller Center, MOMA have high location entropies. On the other hand, the residential areas, e.g., up Manhattan are less popular. All the facts above conform with the effect of location entropy.

D. Time Interval Sequence

Given two different users u and u' and one location ℓ that they co-occurred, user u has a sequence $C_u^\ell = (\langle u, t_1, \ell \rangle, \ldots, \langle u, t_m, \ell \rangle)$ denoting all the check-ins of user u at location ℓ. Similarly, user u' has a sequence $C_{u'}^\ell = (\langle u', t'_1, \ell \rangle, \ldots, \langle u', t'_n, \ell \rangle)$ denoting all his check-ins at location ℓ. From C_u^ℓ and $C_{u'}^\ell$, we obtain their time sequences $T_u^\ell = (t_1, \ldots, t_m)$ and $T_{u'}^\ell = (t'_1, \ldots, t'_n)$. To measure the general check-in time interval between T_u^ℓ and $T_{u'}^\ell$, we define the notion *time interval sequence*.

Definition 3 (Time interval sequence): Given two users $u, u' \in \mathcal{U}$, a location $\ell \in \mathcal{L}$ that they co-occurred, and their time sequences $T_u^\ell = (t_1, \ldots, t_m)$ and $T_{u'}^\ell = (t'_1, \ldots, t'_n)$ at location ℓ, the time interval sequence is defined as $TIS_{u,u'}^\ell = (d_1^u, \ldots, d_m^u, d_1^{u'}, \ldots, d_n^{u'})$, where

$$d_i^u = min(\{|t_i - t'_1|, \ldots, |t_i - t'_n|\})(1 \le i \le m), \text{ and}$$

$$d_j^{u'} = min(\{|t'_j - t_1|, \ldots, |t'_j - t_m|\})(1 \le j \le n).$$

The intuition of this definition is to find each check-in time a matching check-in time which is closest to it and belongs to the other user. Each pair of such kind of matches indicates a possible co-occurrence of two users. That is to say, the time interval sequence takes all possible co-occurrences' time intervals into consideration, which is considered to be more accurate to indicate the general check-in time interval between two users at the same location. For example, given two users u and u' and one of their co-locations ℓ. Their time sequences at location ℓ are T_u^ℓ = (2, 14, 67, 89, 135, 136) and $T_{u'}^\ell$ = (2, 3, 56), respectively. We calculate their time interval sequence to be $TIS_{u,u'}^\ell$ = (0, 11, 11, 33, 79, 80, 0, 1, 11).

E. Dataset

The check-in dataset we use in experiments is collected by [4] from Gowalla, a popular LBSN service back in 2011. This dataset consists of two parts. One is the check-in data composed of 6,442,890 check-ins from more than 100,000 users during the period from Feb. 2009 to Oct. 2010. The format of one check-in item is as follows:

$$\langle userID, timestamps, latitude, longitude, locationID \rangle.$$

The other part is a social graph of users which contains 196,591 users and 950,327 edges.

Since most check-ins are made in the territory of US, we only exploit the check-ins with latitudes between $25^\circ N$ and $50^\circ N$ and with longitudes between $65^\circ W$ and $125^\circ W$ in all our analysis and experiments. With the diversity of American users, the derived results will not lose generality. Concretely, we use 3,672,646 check-ins from 54,622 users. Each location cell represents a location. The cell size s is set to be 1°, 0.1°, 0.01° and 0.001°, respectively.

IV. Friendship Prediction Model $\mathcal{I}$

In this section, we introduce our first friendship prediction model. We first formally present the problem. Next, we prose our solutions. Evaluation results are presented in the end.

A. Problem Definition

To predict whether two users are friends, it is ideal to collect all their mobility information. However, in some circumstances, we may only obtain partial information of two users. The extreme case is we only have one co-occurrence of two users. To deal with this situation, we propose model $\mathcal{I}$. The problem is formally defined as follows:

Problem definition: Given $u, u' \in \mathcal{U}$, one of their co-location $\ell_0 \in L_{u,u'}$, and the check-in dataset C_{ℓ_0} ($C_{\ell_0} \subset \mathcal{C}$) at location ℓ_0 in the form of $\langle u, t, \ell_0 \rangle$ where $u \in \mathcal{U}$ and $t \in \mathcal{T}$, the problem is to predict whether u and u' are friends or not.

B. Model Description

The friendship prediction problem can be naturally formalized into a binary classification problem. If two users are friends, the label is 1. Otherwise, the label is 0. Therefore, we can utilize a machine learning classifier to solve the problem.

The key of building the prediction model is to decide what kind of features we should use. From the check-in dataset, we can obtain three kinds of information: the number of check-ins, check-in time, location characteristics. We consider all these information as features. Specifically, for each tuple $\langle u, u', \ell \rangle$, we extract the following seven features:

1) the number of check-ins conducted by all users at location ℓ, denoted by $|C_\ell|$
2) the number of check-ins conducted by user u at location ℓ, denoted by $|C_u^\ell|$
3) the number of check-ins conducted by user u' at location ℓ, denoted by $|C_{u'}^\ell|$
4) the maximum time interval of u and u' at location ℓ, denoted by $max(TIS_{u,u'}^\ell)$
5) the minimum time interval of u and u' at location ℓ, denoted by $min(TIS_{u,u'}^\ell)$
6) the average time interval of u and u' at location ℓ, denoted by $mean(TIS_{u,u'}^\ell)$
7) the location entropy of location ℓ, denoted by $H(\ell)$.

In following discussions, we give our intuition on considering the time and location features.

Time intervals. For two users co-occur at one location for one or several times, the check-in time interval of the co-occurrence(s) captures important information of their relationship. We consider the time interval through three aspects: the average, maximum and minimum time intervals. Here, the average check-in time interval measures the information of every possible co-occurrence of two users. If the average time interval is small, it indicates that every co-occurrence happens within a short time period, thus two users are highly likely to be friends. On the other hand, if the average interval is large, it does not necessarily mean that the possibility that two users are friends is small. Consider a situation when there is only one check-in match happened during a long time period such as 2 years, then the average time interval will be enlarged by this long period. Therefore, we consider the minimum interval as a feature. But it is not certain that a small minimum interval definitely infers a high possibility of friendship. The check-in match with a small time interval can also be a co-incidence. In the case, we also take into account the maximal time interval.

Location entropy. Considering location entropy (which is used to measure the popularity of a location), we need to capture the difference between two people co-occur at a popular place and a non-popular one. Take train station as an example which normally has a high location entropy, two users may co-occur many times. However, this should not be considered as a strong indicator for their friendship since they may just take the train everyday. On the other hand, location entropy can also strengthen the influence of co-occurrences that happen at a non-popular place. As introduced before, two people visiting the same location within a large time interval indicates a lower friendship probability. However, if the large time interval happens at a low-entropy location, then the situation may be changed, i.e., they may both visit a common friend's home at a different time.

Therefore, we believe it is necessary to consider location popularity and time intervals on friendship prediction.

C. Model Evaluation

Experiment setup. We exploit the downsampling techniques to construct a balanced dataset. The dataset is partitioned 70/30 for training and testing. Our experiments are performed on a machine with duo Intel Xeon 2.26 GHz and 24 GB memory. All the experiments are implemented in MATLAB R2013a. Regarding the machine learning techniques, we adopt logistic regression as our classifier. In all sets, we perform 10-fold cross validation

Metrics. We exploit precision-recall curve and receiver operating characteristic (ROC) curve to measure the performance of our prediction. In addition, we also use the AUC value (area under the ROC curve) as a metric.

Experimental results. Figure 2 gives the precision-recall and ROC curves of experimental results of model $\mathcal{I}$.

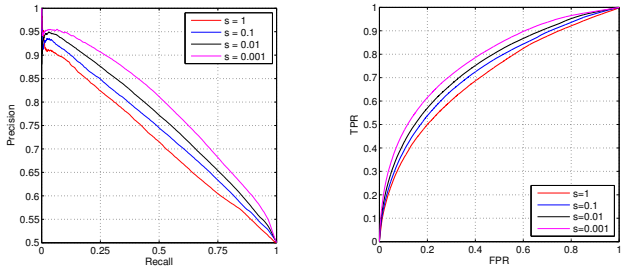
As we can see from Figure 2(a), the precision-recall results get better when the value of cell size s decreases. When $s = 0.001°$, we obtain the best performance of the prediction model. Practically, a smaller s indicates more precisely defined locations. In this case, we can capture user pairs who co-occurred at such locations with more accuracy. On the contrary, if s is large, the location information becomes sparse. A cell can cover the whole NYC when $s = 1°$ and two users both visiting the same city is not a strong indicator of their friendship. In general, as shown from the curve of $s = 0.001°$, when precision is 80%, recall is larger than 50%. On the other hand, when recall is as high as 80%, precision reaches at nearly 65%. This is a promising result as the information we used in our model is limited to only one of two users' co-locations.

The ROC curves of the prediction model under different cell sizes are presented in Figure 2(b). Similarly, the performance increases as s becomes smaller. Moreover, AUC values under $s = 1°, 0.1°, 0.01°$ and $0.001°$ are 0.7043, 0.7294, 0.7519 and 0.7807, respectively. This also indicates $s = 0.001°$ has the best performance.

Comparison with the state-of-the-art model. The model proposed in [6] (we call it CS model for short) solves the same problem with model $\mathcal{I}$. In order to perform comparison between results of CS model and model I, we conduct experiments on CS model with exactly the same Gowalla dataset and classifier as in our experiments for model $\mathcal{I}$.

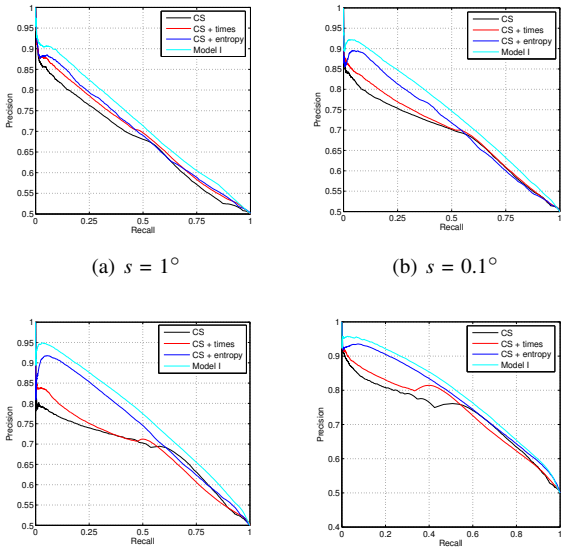
The CS model extracted the following three features from the check-in dataset for machine learning:

1) the number of check-ins conducted by all users at location ℓ, denoted by $|C_\ell|$
2) the number of check-ins conducted by user u at location ℓ, denoted by $|C_u^\ell|$



(a) Precision-recall curve. (b) ROC curve.

Fig. 2. Results of model I under logistic regression.



(a) s = 1° (b) s = 0.1°

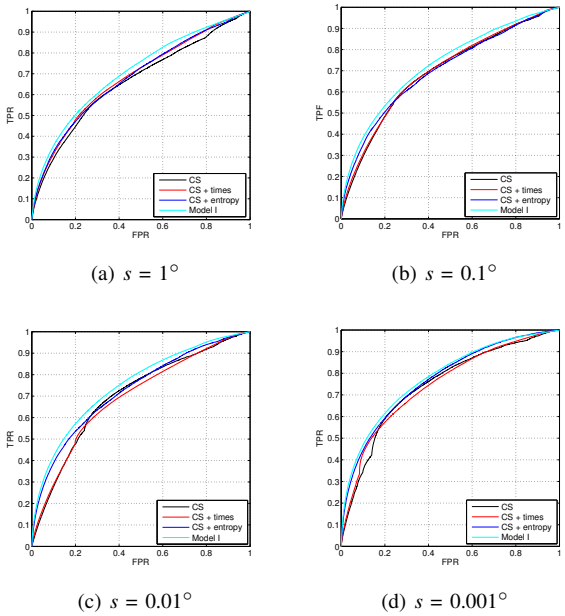
(c) s = 0.01° (d) s = 0.001°

Fig. 3. Result comparisons of model I (precision-recall curve).

3) the number of check-ins conducted by user u' at location ℓ, denoted by $|C^{\ell}_{u'}|$.

To demonstrate time and location entropy's contributions on friendship prediction, we construct another two models: CS + time model and CS + entropy model. In CS + time model, for each tuple $\langle u, u', \ell \rangle$, we have following six features: $|C_{\ell}|$, $|C^{\ell}_{u}|$, $|C^{\ell}_{u'}|$, $mean(TIS^{\ell}_{u,u'})$, $min(TIS^{\ell}_{u,u'})$ and $max(TIS^{\ell}_{u,u'})$. In CS + entropy model, we extract four features: $|C_{\ell}|$, $|C^{\ell}_{u}|$, $|C^{\ell}_{u'}|$ and $H(\ell)$.

Figure 3 shows the precision-recall curves of those four models under four different cell sizes. As we can see, regardless of the cell size, both CS+time and CS+entropy model outperform the CS model. This indicates that time and location entropy both contribute to the prediction results. Besides, blue lines are generally higher than red lines, which implies that location entropy has better influence on friendship prediction than time interval. By combining time interval and entropy together into one model, i.e., our model $\mathcal{I}$, we obtain the best results among all sets. The ROC curves (Figure 4) have demonstrated the same results.



(a) s = 1° (b) s = 0.1°

(c) s = 0.01° (d) s = 0.001°

Fig. 4. Result comparisons of model I (ROC curve).

Cell size	CS model	CS	CS	Model I
1°	0.6638	0.6834	0.6814	0.7063
0.1°	0.6946	0.6989	0.7005	0.7285
0.01°	0.7038	0.6956	0.7271	0.7522
0.001°	0.7457	0.7464	0.7705	0.7808

TABLE I
AUC OF DIFFERENT MODELS UNDER DIFFERENT CELL SIZES.

V. FRIENDSHIP PREDICTION MODEL $\mathcal{II}$

In this section, we first give the formal definition of the friendship prediction problem considering all the co-locations between two users. We then give the solution of the problem (i.e., model $\mathcal{II}$). The evaluation results are shown in the end.

A. Problem Definition

With the information of check-ins at all co-locations of two users, a better analysis and solution can be proposed to solve the friendship prediction problem. The formal definition of this problem is given as follows:

Problem definition: Given $u, u' \in \mathcal{U}$, and a check-in dataset $\mathcal{C}$ in the form of $\langle u, t, \ell \rangle$, the problem is to predict whether u and u' are friends or not.

B. Model Description

Similar to model $\mathcal{I}$, the problem defined above can be transformed into a binary classification problem. For each pair of users u and u', they have a co-location set $L_{u,u'}$. For each $\ell \in L_{u,u'}$, we can calculate the average time interval - $mean(TIS^{\ell}_{u,u'})$, the minimum time interval - $min(TIS^{\ell}_{u,u'})$, the maximum time interval - $max(TIS^{\ell}_{u,u'})$, its number of co-occurrences - $|C^{\ell}_{u,u'}|$, and its location entropy $H(\ell)$. To gather all the information of each co-location together, we combine these data in following ways. For three time intervals,

we take the average values of them from all co-locations separately. For the co-occurrence number, we propose a notion called the *weighted number of co-occurrences*, which will be explained in detail later. In addition, we can also obtain another useful information - the number of co-locations. Similarly, we propose a notion called the *weighted number of co-locations* to measure it. In total, we extract the following five features for prediction.

1) the average value of the maximum check-in time intervals of u and u' at all their co-locations, denoted by $mean(\sum_{\ell \in L_{u,u'}} max(TIS^{\ell}_{u,u'}))$
2) the average value of the minimum check-in time intervals of u and u' at all their co-locations, denoted by $mean(\sum_{\ell \in L_{u,u'}} min(TIS^{\ell}_{u,u'}))$
3) the average value of the average check-in time intervals of u and u' at all their co-locations, denoted by $mean(\sum_{\ell \in L_{u,u'}} mean(TIS^{\ell}_{u,u'}))$
4) the weighted number of co-locations users u and u', denoted by $WL(u, u')$
5) the weighted number of co-occurrences of users u and u', denoted by $WO(u, u')$.

Weighted number of co-locations. Given two users u and u' and their co-location set $L_{u,u'}$, the *weighted number of co-locations*, which is denoted by $WL(u, u')$, is formally defined as follows:

$$WL(u, u') = \sum_{\ell \in L_{u,u'}} exp(-H(\ell)).$$

This notion is used to measure the number of important co-locations of two users. The importance of a location is decided by its location entropy. As discussed in Section III-C, if two users co-occur at a location with low location entropy, their possibility of being friends is larger than two people who co-occur at a high-entropy location. Therefore, the co-locations with low entropy can be considered to be more important than high-entropy ones. As shown in the definition, we use $exp(-H(\ell))$ to be the weight of location ℓ and to measure its importance. Table II gives four examples to demonstrate this notion. Most locations in row 2 have higher entropy, and its corresponding value of weighted number of co-locations is small. It indicates that high-entropy co-locations have small contribution to $WL(u, u')$. On the contrary, as we can see from row 1, low-entropy co-locations make larger contribution to this value.

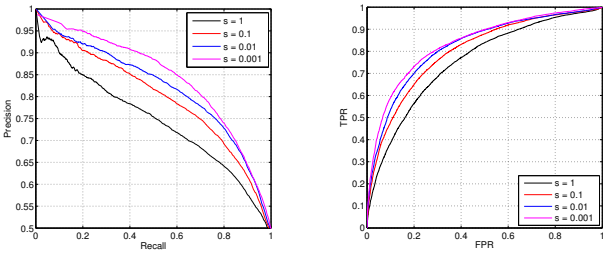
$H(\ell_1)$	$H(\ell_2)$	$H(\ell_3)$	$H(\ell_4)$	$WL(u, u')$
0.02	0.18	0.23	0.09	3.2481
1.25	2.18	3.34	2.29	0.5362
0.65	0.16	2.67	3.14	1.4867
3.24	0.54	0.38	0.66	1.8226

TABLE II
EXAMPLES OF THE WEIGHTED NUMBER OF CO-LOCATIONS.

Weighted number of co-occurrences. Given two users u and u' and their co-location set $L_{u,u'}$, the *weighted number of co-occurrences*, which is denoted by $WO(u, u')$, is formally defined as follows:

$$WO(u, u') = \sum_{\ell \in L_{u,u'}} |C^{\ell}_{u,u'}| \times exp(-H(\ell)).$$

Similar with the weighted number of co-locations, this notion strengthens the co-occurrences happened at non-popular places and, in the meantime, weakens the co-occurrences happened at popular places based on location entropy. In this way, it can provide more useful information for friendship prediction than the number of co-occurrences. It works in the same way with the weighted number of co-locations.



(a) Precision-recall curve (b) ROC curve

Fig. 5. Results of model $\mathcal{II}$ under logistic regression.

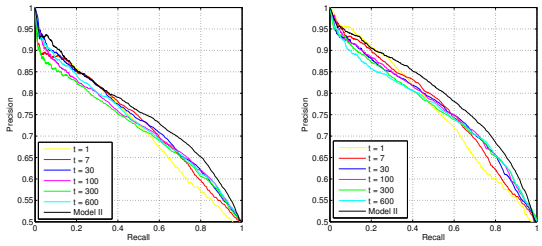
C. Model Evaluation

The experiment setup and the metrics are the same as model $\mathcal{I}$.

Experimental results. The experiment results are presented in Figure 5. For the precision-recall curves, the result gets better as the cell size decreases. Among these, the magenta line is the highest one, indicating that, when $s = 0.001°$, model $\mathcal{II}$ achieves its best performance. Specifically, when precision is as high as 80%, recall is larger than 70%, i.e., our prediction achieves a strong performance.

The ROC curves in Figure 5(b) presents the similar results. On the other hand, we get the same result referring to AUC value. The AUC values are 0.7541, 0.7999, 0.8213 and 0.8359 for $s = 1°, 0.1°, 0.01°$ and $0.001°$, respectively. When $s = 0.001°$, the curve has the largest AUC. It indicates the best performance of the prediction model.

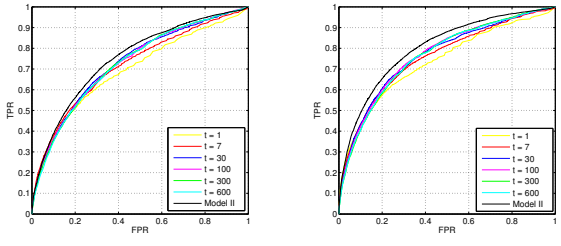
Comparison with the state-of-the-art model. We also compare our model $\mathcal{II}$ with a state-of-the-art model, namely EBM [7]. In EBM, for each pair of users, two features are extracted from check-in dataset, namely *Renyi entropy-based diversity* and *weighted frequency*. Diversity is used to measure the diversity of the co-occurrences of two users u and u' at different locations. It weakens the influence of frequent coincidences. Frequency is used to tell how important the co-occurrences at non-crowded places are to two users' friendship. It strengthens the influence of co-occurrences that happened at non-popular places. We set all the parameters following the experiments in [7]. In addition, there is a parameter for time interval τ on defining the co-occurrence



(a) $s = 1°$ (b) $s = 0.1°$

(c) $s = 0.01°$ (d) $s = 0.001°$

Fig. 6. Result comparisons of model $\mathcal{II}$ and the EBM model (precision-recall curve).



(a) $s = 1°$ (b) $s = 0.1°$

(c) $s = 0.01°$ (d) $s = 0.001°$

Fig. 7. Result comparisons of model $\mathcal{II}$ and the EBM model (ROC curve).

in EBM. We treat it as an application dependent parameter and exploit several including one day, one week, one month, 100 days, 300 days and 600 days in our experiments.

Figure 6 shows the precision-recall curves of model $\mathcal{II}$ and EBM model under different cell sizes. Note that, no matter what the cell size is, the largest check-in time interval of two users in a cell is smaller than 600 days. Besides, over the constrains of cell size s and time interval τ, the number of friends who co-occurred in the same cell within τ varies a lot. Table III shows the numbers of data belonging to friends we are able to obtain under different time intervals and different cell sizes. As shown in Figure 6, the variation of time interval τ does not lead to obvious changes on the performances of EBM model. On the other hand, when $s = 1°$, $0.1°$ and $0.01°$, black curves that represent the result of model $\mathcal{II}$ are higher than all the other curves belongs to EBM model. This implies that, in such cases, our model has better performance than EBM model.

	$s = 1°$	$s = 0.1°$	$s = 0.01°$	$s = 0.001°$
model $\mathcal{II}$	155,710	142,642	122,130	97,872
$\tau = 1$	83,769	69,293	49,396	37,645
$\tau = 7$	104,492	90,918	69,465	51,917
$\tau = 30$	121,624	108,554	87,777	67,533
$\tau = 100$	140,893	127,639	106,551	84,066
$\tau = 300$	155,107	142,094	121,535	97,273
$\tau = 600$	155,710	142,642	122,130	97,872

TABLE III
OF DATA BELONGS TO FRIENDS FOR MODEL $\mathcal{II}$ AND EBM MODEL UNDER DIFFERENT TIME INTERVALS AND DIFFERENT CELL SIZES.

Figure 7 shows the ROC curves of model $\mathcal{II}$ and the EBM model under different cell sizes. The time interval τ is set to be different values as before in EBM model. Same conclusion can obtained. Again, we cannot get a clear pattern of the relationship between time interval τ and EBM model's performance. Besides, model $\mathcal{II}$ performs better than EBM model when cell size is set to be $1°$, $0.1°$ and $0.01°$, as black curves are always higher than other curves in first three sub-figures. However, when $s = 0.001°$, model $\mathcal{II}$ has a similar result as EBM. Table IV further gives AUC values of each curve in the figure, model $\mathcal{II}$ has larger AUC values than EBM model except when the cell size is $0.001°$.

	$s = 1°$	$s = 0.1°$	$s = 0.01°$	$s = 0.001°$
model $\mathcal{II}$	0.7510	0.8011	0.8237	0.8369
$\tau = 1$	0.6960	0.7352	0.7597	0.7596
$\tau = 7$	0.7183	0.7550	0.7698	0.8268
$\tau = 30$	0.7298	0.7652	0.7836	0.8329
$\tau = 100$	0.7282	0.7673	0.7673	0.8439
$\tau = 300$	0.7274	0.7632	0.7863	0.8419
$\tau = 600$	0.7260	0.7629	0.7848	0.8419

TABLE IV
AUC VALUES OF ROC CURVES UNDER DIFFERENT TIME INTERVALS AND DIFFERENT CELL SIZES.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have focused on friendship prediction from LBSN dataset. We proposed two friendship prediction models. In model $\mathcal{I}$, we studied the problem of predicting whether two people are friends under the situation that only the check-ins happened at one certain location can be obtained. Compared with the state-of-the-art CS model [6], we take check-in time interval and location entropy into consideration, which leads to a more effective friendship prediction. In model $\mathcal{II}$, differently, we focus on utilizing all the check-in information that belong to any co-location of two users to predict their

relationship. We consider five elements that would make a difference on friendship prediction - the weighted number of co-occurrences, the weighted number of co-locations, the average time interval, the minimum time interval and maximum time intervals. The experimental results shows that model $\mathcal{II}$ outperforms the EBM model [7] in most cases.

In the future, we would like to further investigate the relation between check-in time and friendship. Specifically, if the co-occurrence happens in the evening or weekend, its influence on friendship should be stronger since evening and weekend can be considered as social time. Moreover, location prediction from social relationships is another interesting problem worth investigation (e.g., see [15], [16]).

REFERENCES

[1] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma, "Mining User Similarity Based on Location History," in *Proc. 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2008, p. 34.

[2] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring Friendship Network Structure by Using Mobile Phone Data," *Proceedings of the National Academy of Sciences*, vol. 106, no. 36, pp. 15 274–15 278, 2009.

[3] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the Gap Between Physical Location and Online Social Networks," in *Proc. 12th ACM International Conference on Ubiquitous computing*. ACM, 2010, pp. 119–128.

[4] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and Mobility: User Movement in Location-based Social Networks," in *Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2011, pp. 1082–1090.

[5] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, "Inferring Social Ties from Geographic Coincidences," *Proceedings of the National Academy of Sciences*, vol. 107, no. 52, pp. 22 436–22 441, 2010.

[6] J. Chang and E. Sun, "Location^3: How Users Share and Respond to Location-based Data on Social Networking Sites," in *Proc. 5th International Conference on Weblogs and Social Media*. The AAAI Press, 2011.

[7] H. Pham, C. Shahabi, and Y. Liu, "EBM- An Entropy-based Model to Infer Social Strength from Spatiotemporal Data," in *Proc. 2013 International Conference on Management of Data*. ACM, 2013, pp. 265–276.

[8] X. Chen, J. Pang, and R. Xue, "Constructing and comparing user mobility profiles for location-based services," in *Proc. 28th ACM Symposium on Applied Computing*. ACM Press, 2013, pp. 261–266.

[9] X. Chen, J. Pang, and R. Xue, "Constructing and comparing user mobility profiles," *ACM Transactions on the Web*, vol. 8, no. 4, p. article 21, 2014.

[10] J. Tang, Y. Chang, and H. Liu, "Mining social media with social theories: a survey," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 2, pp. 20–29, 2014.

[11] X. Chen, R. Lv, X. Ma, and J. Pang, "Measuring user similarity with trajectory patterns: Principles and new metrics," in *Proc. 16th Asia-Pacific Web Conference*, ser. Lecture Notes in Computer Science, vol. 8709. Springer, 2014, pp. 437–448.

[12] X. Chen, P. Kordy, R. Lv, and J. Pang, "MinUS: Mining user similarity with trajectory patterns," in *Proc. 7th European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, vol. 8726. Springer, 2014, pp. 436–439.

[13] H. Wang, Z. Li, and W.-C. Lee, "PGT: Measuring mobility relationship using personal, global and temporal factors," in *Proc. 14th IEEE International Conference on Data Mining (ICDM)*. IEEE, 2014, pp. 570–579.

[14] Y. Zhang and J. Pang, "Distance and friendship: A distance-based model for link prediction in social networks," in *Proc. 17th Asia-Pacific Web Conference*, ser. LNCS. Springer, 2015, accepted.

[15] J. Pang and Y. Zhang, "Exploring communities for effective location prediction," in *Proc. 24th World Wide Web Conference (Companion Volume)*. ACM, 2015, pp. 87–88.

[16] J. Pang and Y. Zhang, "Event prediction with community leaders," in *Proc. 10th Conference on Availability, Reliability and Security*. IEEE CS, 2015, accepted.