# Spatio-Temporal Prediction of Social Connections

Guolei Yang
Iowa State University
Ames, Iowa, USA 50011
yanggl@iastate.edu

Andreas Züfle
George Mason University
Fairfax, Virginia, USA 22030
azufle@gmu.edu

## ABSTRACT

It is long known that a user's mobility pattern can be affected by his social connections. Users tend to visit same locations visited by their friends. In this paper we investigate the inverse problem: How does a set of users' trajectory reflects their social connections. To this end, we define the social connection prediction problem. Given two users, predict the probability that they are friends by mining their historical trajectories. A naive method to do so is to exam how often the two users visits the same location at the same time, which suffers from the problem that different locations/times may have different predictive power. We propose a comprehensive prediction model that is able to capture this difference between locations and time slots. To demonstrate its effectiveness, we trained the proposed model using the publicly available Foursquare dataset. The result shows the proposed model is able to predict existence of social connections between randomly selected users significantly more accurate comparing with the naive method.

## CCS CONCEPTS

**•Information systems → Location based services; Data mining; •Computing methodologies →** *Machine learning;*

## KEYWORDS

Location-Based Social Network, social connection prediction, feature selection, spatio-temporal data

**ACM Reference format:**
Guolei Yang and Andreas Züfle. 2016. Spatio-Temporal Prediction of Social Connections. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 3 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

In the past decade, with the rise of Location-Based Social Networks (LBSN), huge amount of geo-spatial data is collected on a daily basis. For example, the Foursquare[13] dataset contains more than 30 millions of self-reported check-ins from thousands of user around the world. As a result, it becomes possible to mine spatio-temporal data and study human mobility pattern at unprecedented large scale.

Several studies on human mobility pattern reveal that a user's movement can be affected by his social connections [2, 14]. For example, a group of close friends tend to check-in to the same locations at the same time period. As such, it is possible to predict a user's future movement by mining the historical trajectory of his friends on LBSN. These studies has since inspired a seriers of research efforts towards the prediction of future individual movements (e.g., [3, 5, 8, 10]). Another research direction (e.g., [12]) focuses on exploring historical trajectories to identify similar users, which can be used for friend recommendation and community detection on LBSN.

Towards the goal of a more thorough understanding of human mobility patterns, we propose to investigate the inverse problem: How does a set of users' trajectory reflects their social connections. In particular, we define the social connection prediction problem: Given the trajectories of two LBSN users $u_i$ and $u_j$, we aim to model the probability that $u_i$ and $u_j$ are friends on the LBSN using their trajectories. Social connection prediction is a long standing research topic. Most existing methods relies on link prediction techniques, which exploit a user's profile and existing social connections to make predictions of hiding links between users, but not users' trajectory. The focus of this paper is not to compete with, but to supplement existing methods.

A straightforward way to predict the social connection, or the lack thereof, between two users is to exam the *spatio-temporal overlap* of their trajectories, i.e., find events where the two users visit the same location at the same time on their trajectories. We define such an event as a *co-visitation* of the two users. The assumption is, if two users frequently visits the same location during the same time peroid, they might be friend with each other. Thus the occurance of co-visitations could reflect when and where they were meeting. The same assumption is used to identify similar users in [12]. Algorithms such as co-location mining [11] can be used to discover co-visitations among users.

Although the above assumption is reasonable, this naive solution suffers from two problems. First, it treats all locations equally in predicting social connections, which is not realistic. For example, if two users frequently meet at private locations like someone's house, or a small coffee shop, it is very likely that they know each other. However, if they both check-in to the same Walmart supermarket after work, it might be just an coincidence simply because there it is the only supermarket near their home. Second, this method ignores the time difference of check-ins behaviours. If two users both check-in to a restaurant at 6:00pm, it is not as significant as two users visit the same location at 10:00pm. This is because most customer of the restaurant may choose to dine there around 6:00pm, but if two users both decide to check-in there at 10:00pm, the chance that it is purely an coincidence is relatively lower. Although the technique proposed in [12] considered the impact of different granularity


*Conference'17, Washington, DC, USA*

DOI: 10.1145/nnnnnnn.nnnnnnn

of locations (e.g., the same state v.s. the same city), it does not explicitly distinguish the predictive power of different locations.

We propose to employ a more comprehensive methodology to study the social connection prediction problem. Unlike the naive solution, we assume different locations and different time slots have different predictive power. We propose a social connection prediction model that is able to capture the inherent difference among locations and times. Specifically, we first convert the trajectory of two users into a high-dimensional co-visitation feature vector, and then use feature selection to filter out less significant co-visitations. A predictive model is then learnt based on the selected features using the Foursquare dataset. Using the users' social connections on Foursquare as ground truth, we show that the proposed model outperforms the naive algorithm that counts only the number of co-visitations. We summarize our contributions as follows:

- We study how the trajectories of a set of users reflect their social connections. To this end, we define the social connection prediction problem: Given the trajectories of two LBSN users $u$ and $v$, we aim to model the probability that $u$ and $v$ are friends on the LBSN.
- Our key observation is: different locations and time may have different predictive power, which is in accordance with common sense. As such, we propose a social connection model that is able to capture this difference among locations and times.
- We demonstrate effectiveness of the proposed model using the Foursquare dataset. The result shows the proposed method outperforms the naive trajectory overlap based solution in prediction accuracy.

The rest of the paper is organized as follows: Related works are summarized in Section 2. We formally define our problem and give an overview of our methodology in Section 3. Section 4 discuses the feature extraction process. Section 5 presents details of model training. Experiment results are showed and analysed in Section 6. And finally, Section 7 concludes the paper.

## 2 RELATED WORK

The spatio-temporal social connection prediction problem we study in this paper is directly related to link prediction problem on social networks. Given the snapshot of a social network at time $t$, the goal of link prediction is to predict links, i.e., social connections, that will emerge at a later time, or to identify missing links at $t$. Such missing links could be the result of privacy settings, e.g., a user may want to hide his friend list from the general public.

Existing works in the field mainly explore two types of information in predicting links: 1) Network structure, i.e., existing social connections, and 2) node attributes such as user profiles. We briefly summarize some representative works. The relational learning [7, 9, 16] and matrix factorization-based [6] techniques both leverage attribute information for link prediction. The Supervised Random Walk (SRW) technique proposed in [1] combines networks structure and edge attributes to improve prediction accuracy, but does not fully explore node attributes. In [15], network structure and node attributes are integrated with a Social Attribute Network (SAN) model, which is later generalized in [4] to both predict links and infer missing attributes.

Our problem is also closely related to [12], which proposes to explore trajectory data to identify similar users. Their goal is to find users who share similar interests in locations, which serves as a friend recommendation tool on LBSN. The proposed technique employs clustering over users location history to identify similar users. In contrast, our work focus on studying the predictive power of trajectories in terms of reflecting existing or missing social connections among users. And our methodology is to model the probability of the existence of such social connections between two users. Form this perspective, our work intends to complement existing studies on human mobility patterns.

## 3 OVERVIEW

### 3.1 Problem Statement

We define a user's trajectory as a series of timestamped check-ins, where each check-in indicates the exact place (i.e., a restaurant, a coffee shop, etc.) the user visits, instead of a geo-graphical coordinate. The Foursquare dataset is an example of such trajectory that consists of self-reported check-ins. Note that coordinate-based trajectory can be converted into such check-ins by joining the coordinates with a database of Point-of-Interests (PoI), such as provided by Open-Stree Map. For simplicity, we consider only check-in-based trajectory in this paper. We formally define the notion of *Check-in* and *User Trajectory* as follows.

*Definition 3.1 (Check-in).* Let $U$ denote a set of unique user identifiers, $L$ denote a set of locations, and $T$ denote the time domain. A check-in $c$ is a triple $(u, l, t) \in U \times L \times T$, which indicates the user $u$ has visited $l$ at time $t$.

*Definition 3.2 (User-trajectory).* Let $C$ be a collection of check-ins and $u \in U$ a user, then the set $C_u := \{(u', l, t) \in C | u = u'\}$ is the user-trajectory (or simply trajectory) of $u$.

The proposed social connection prediction model is based on the concept of *Co-visitation*, which is defined as follows:

*Definition 3.3 (Co-visitation).* A co-visitation of two users $u_i$ and $u_j$ to a location $l$ is defined as the event that $u_i$ and $u_j$ report two check-ins $(u_i, l, t_i)$ and $(u_j, l, t_j)$ respectively, where $|t_i - t_j| \leq \tau$.

Here $\tau$ is an experience-based parameter called the co-visitation time window. We formulate the social connection prediction problem as a classification problem. Given the trajectory of two users $u_i$ and $u_j$, the goal is to assign the pair of users $(u_i, u_j)$ into one of the two classes: *Friends* or *Non-friends*.

### 3.2 Methodology

We model the probability of the existence of social connection between two users based on the hypothesis that socially connected users tend to visit same locations at same time periods, which is defined as co-visitations. We propose a three step model learning process.

- **Co-visitation feature extraction** Given the trajectories of two users $u_i$ and $u_j$, we first convert their trajectories into a high-dimensional co-visitation feature vector that reflects the time and location of their co-visitations.
- **Feature selection** A feature selection process is applied on the co-visitation feature vector in order to filter out

less significant co-visitations. It helps us to identify the locations and time slots that are most related to user's social connections.
- **Model learning** A model is then learnt from the selected features. To find the most suitable model for our problem, we explored several modelling techniques including the logistic regression, Support Vector Machine (SVM), and artificial neural network.

# 4 CO-VISITATION FEATURE EXTRACTION

# 5 FEATURE SELECTION AND MODEL LEARNING

# 6 EVALUATION

In this section, we report the preliminary experiment results of the proposed methodology on the trajectory of a selected subset of Foursquare users.

## 6.1 Dataset Description

We evaluate the proposed model on the widely-used Foursquare check-in dataset [13]. In our experiments, we mine the check-in data from two of the most popular cities, including New York City (NYC) and Tokyo. The dataset contains about 227,428 check-ins reported in NYC and 573,703 in Tokyo. The check-ins were collected for about 10 month. From each check-in, we extract the user ID, location ID, and a timestamp. Using the user ID or location ID, we retrieve the profile of that user or location on Foursquare. The user profile include the social connection between users ("follower - followee") and the location profile includes its category (*Food, Coffee, Nightlife, Fun, and Shopping*), coordinates, and user rating. The check-ins are grouped by user ID/location ID and sorted by their timestamps.

Select most active users...Ground truth...etc.

## 6.2 Experiment Design

For comparison purpose, we have implemented the following schemes:

- **Random** This scheme randomly assigns a user pair as friends or non-friends, each with a probability of 50%.
- **Naive** This scheme simple counts the number of co-visitations of two users. If the number is higher than a threshold, the two users are predicted to be friends, and otherwise non-friends. The threshold is set to be the average number of co-visitations of each pair of friends in the dataset we used.
- **Proposed** The proposed model that assumes co-visitations occurred at different locations and time slots have different predictive power.

## 6.3 Results

# 7 CONCLUSION

In this paper, we study the predictive power of user trajectories in reflecting their social connections. Based on the hypothesis that friends tend to visit same locations at same time, we propose to model the probability of that social connection exists between two users using their co-visitations. We propose a three step model learning process. First, a co-visitation feature vector is generated based on the trajectory of two users. We then employ feature selection to filter out less significant co-visitations from the feature vector. Finally, we explore several statistic models in order to find the one that is able to yield the best performance for our problem. In our preliminary experiments using a subset of users selected from the Foursquare dataset, we find the proposed methodology shows promising performance, in terms of prediction accuracy, comparing with a naive solution based on simple counting the number of co-visitations between users. As for future work, we aim to develop a more comprehensive social connection prediction framework that combines spatio-temporal data with network structure as well as node attribute.

# REFERENCES


[1] Lars Backstrom and Jure Leskovec. 2011. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 635–644.

[2] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD conference on knowledge discovery and data mining*. ACM, 1082–1090.

[3] Huiji Gao, Jiliang Tang, and Huan Liu. 2012. Mobile location prediction in spatio-temporal context. In *Nokia mobile data challenge workshop*, Vol. 41. 44.

[4] Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine Runting Shi, and Dawn Song. 2014. Joint link prediction and attribute inference using a social-attribute network. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 2 (2014), 27.

[5] Defu Lian, Vincent W Zheng, and Xing Xie. 2013. Collaborative filtering meets next check-in location prediction. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 231–232.

[6] Aditya Krishna Menon and Charles Elkan. 2011. Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases*. Springer, 437–452.

[7] Kurt Miller, Michael I Jordan, and Thomas L Griffiths. 2009. Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*. 1276–1284.

[8] Anastasios Noulas, Salvatore Scellato, Neal Lathia, and Cecilia Mascolo. 2012. Mining user mobility features for next place prediction in location-based services. In *Data mining (ICDM), IEEE 12th international conference on*. IEEE, 1038–1043.

[9] Ben Taskar Ming-Fai Wong Pieter and Abbeel Daphne Koller. 2003. Link prediction in relational data. (2003).

[10] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. 2011. NextPlace: a spatio-temporal prediction framework for pervasive systems. In *International Conference on Pervasive Computing*. Springer, 152–169.

[11] Michael Weiler, Klaus Arthur Schmid, Nikos Mamoulis, and Matthias Renz. 2015. Geo-Social Co-location Mining. In *Second International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data*. ACM, 19–24.

[12] Xiangye Xiao, Yu Zheng, Qiong Luo, and Xing Xie. 2010. Finding similar users using category-based location history. In *ACM SIGSPATIAL*. 442–445.

[13] Dingqi Yang, Daqing Zhang, Longbiao Chen, and Bingqing Qu. 2015. NationTelescope: Monitoring and visualizing large-scale collective behavior in LBSNs. *Journal of Network and Computer Applications* 55 (2015), 170–180.

[14] Jihang Ye, Zhe Zhu, and Hong Cheng. 2013. What's your next move: User activity prediction in location-based social networks. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 171–179.

[15] Zhijun Yin, Manish Gupta, Tim Weninger, and Jiawei Han. 2010. Linkrec: a unified framework for link recommendation with user attributes and graph structure. In *Proceedings of the 19th international conference on World wide web*. ACM, 1211–1212.

[16] Kai Yu, Wei Chu, Shipeng Yu, Volker Tresp, and Zhao Xu. 2006. Stochastic relational models for discriminative link prediction. In *NIPS*. 1553–1560.