Data Min Knowl Disc (2015) 29:299–323
DOI 10.1007/s10618-014-0343-4

Addressing the cold-start problem in location recommendation using geo-social correlations

Huiji Gao · Jiliang Tang · Huan Liu

Received: 5 April 2013 / Accepted: 2 January 2014 / Published online: 12 January 2014

Abstract Location-based social networks (LBSNs) have attracted an increasing number of users in recent years, resulting in large amounts of geographical and social data. Such LBSN data provide an unprecedented opportunity to study the human movement from their socio-spatial behavior, in order to improve location-based applications like location recommendation. As users can check-in at new places, traditional work on location prediction that relies on mining a user's historical moving trajectories fails as it is not designed for the cold-start problem of recommending new check-ins. While previous work on LBSNs attempting to utilize a user's social connections for location recommendation observed limited help from social network information. In this work, we propose to address the cold-start location recommendation problem by capturing the correlations between social networks and geographical distance on LBSNs with a geo-social correlation model. The experimental results on a real-world LBSN dataset demonstrate that our approach properly models the geo-social correlations of a user's cold-start check-ins and significantly improves the location recommendation performance.

Keywords Location-based social networks · Location recommendation · Location prediction · Cold-start · Geo-social correlation

Responsible editor: M.J. Zaki.

H. Gao (✉) · J. Tang · H. Liu
Arizona State University, Tempe, AZ 85287, USA
e-mail: Huiji.Gao@asu.edu

J. Tang
e-mail: Jiliang.Tang@asu.edu

H. Liu
e-mail: Huan.Liu@asu.edu

1 Introduction

Location-based social media attracts millions of users, and generates large location-based social networks (Kessler 2012). A recent survey from the Pew Internet and American Life Project reports that over 28 % of Americans use mobile or social location-based services (Zickuhr and Smith 2011). Typical online location-based social networking sites such as Foursquare[1] and Facebook Places[2] provide location-based services for users to "check-in" at a physical place, and automatically include the location into their posts. The online "check-in" posts a user's current geographical location, making known to his friends the information on when and where he is. Compared with many other online activities, "check-in" reflects a user's geographical action in the real world, residing where the online world and real world intersect. Thus, the study of check-ins provides an ideal environment to understand human behavior, and could also benefit a variety of location-based services such as mobile marketing (Barnes and Scornavacca 2004; Li 2011) and disaster relief (Goodchild and Glennon 2010; Gao et al. 2011). Among various applications on LBSNs, location recommendation has become a significant task in recent years since it is proposed to help users filter out uninteresting items and reduce time in decision making, which could also benefit virtual marketing.

One of the most significant properties of check-in behavior is the user-driven property (Noulas et al. 2011). When using location-based social networking services, a user is able to choose where and when to make a check-in. It is reported in previous research that a user's check-ins displays a power-law distribution on LBSNs, i.e., a user goes to a few places many times and to many places a few times (Gao et al. 2012b), indicating that users do visit new places, resulting in the cold-start check-in problem. Recommending a none cold-start location to a user (also referred to as "location prediction") has been widely studied by taking advantage of spatial trajectories (Monreale et al. 2009; Spaccapietra et al. 2008), periodical patterns (Thanh et al. 2007), spatial-temporal patterns (Scellato et al. 2011a; Gao et al. 2012a), etc. The success of these methods relies on sufficient numbers of observations on the target location in an individual's check-in history; hence, it is difficult to apply them to the cold-start check-ins as there is no historical information on the user for the new place he will go to.

Facing the difficulty of recommending cold-start check-in locations, researchers resort to social network information on LBSNs and investigate if it could help solve the recalcitrant cold-start problem. As suggested by social theories (e.g., social correlation (Anagnostopoulos et al. 2008)), human movement is usually affected by their social networks, such as watching movies with families, visiting friends, traveling by following friends' recommendations, and so on, providing the potential opportunity to solve the cold-start recommendation problem from a user's social friends. However, recent work on utilizing social information for location recommendation has reported limited improvement (Cho et al. 2011; Gao et al. 2012b; Ye et al. 2010, 2011). One

[1] https://foursquare.com.

[2] https://www.facebook.com/about/location.

explanation of this phenomenon could be the check-in characteristics of LBSNs. It has been reported that in general users with social connections only share less than 10 % common check-in locations (Gao et al. 2012b; Cho et al. 2011), which provide very limited observation for social recommendation.

Since the check-in action connects a user's geographical movement and his social networks, it actually provides a new perspective to study a user's cold-start check-in behavior not only through social aspect but also from the closely correlated geographical aspect, i.e., geo-social perspective. Researchers investigated how geographical distance influences social networks, and how social networks influence human movement on LBSNs (Scellato et al. 2011b, c; Cheng et al. 2011) indicating the necessity to consider these two factors together when studying human mobile behavior, and suggesting the potential opportunity to improve current location recommendation approaches. In this paper, we propose the concept of geo-social correlations to combine both social networks and geographical distance for recommending cold-start check-in locations. In particular, we study the following issues:

- Are user's cold-start check-ins correlated to their social ties on LBSNs?
- How to capture the social correlations on LBSNs? and
- How to utilize the social correlations for solving the cold-start location recommendation problem?

To the best of our knowledge, this work presents the first comprehensive study of geo-social correlations for the cold-start problem on location-based social networks. The contributions of our work are summarized below:

- We study the usability of social network information on LBSNs, and propose a feasible solution for the cold-start location recommendation problem by taking advantage of geo-social correlations.
- We investigate the social correlations in geo-social perspective, and observe that users in different geo-social circles have various correlation strength.
- We suggest various correlation measures to capture the geo-social correlations of a user's check-in behavior on the cold-start problem, and determine the most effective correlation measures for each geo-social circle.
- We propose a geo-social correlation model (**gSCorr**) to solve the cold-start location recommendation problem by considering four types of geo-social circles with corresponding correlation strength.

The remainder of this paper is organized as follows. We first introduce the concept of geo-social correlations of check-in behavior on LBSNs in Sect. 2, present the proposed model for geo-social correlations in Sect. 3, discuss the experimental design and results on the real-world dataset in Sect. 4, followed by related work in Sect. 5, and provide some conclusions with future work in Sect. 6.

2 Geo-social correlations on LBSNs

When we observe a check-in from a user, there are two scenarios: checking in at a previous visited location, or a new location that the user has never checked in before. In this paper, we define the former one as "existing check-in(s)", and the latter one as "'cold-start' check-in(s)", with respect to a user's check-in history. In Gao et al. (2012b), the authors explored the social-historical ties of check-in behavior on LBSNs, and found that both ties have effects on explaining a user's check-in behavior. To investigate the social correlations on a user's check-in behavior, we need a controlled social environment that excludes the effects of users' historical ties. However, to distinguish whether a user's existing check-in is correlated to his historical ties or social ties is actually a big challenge (Gao et al. 2012b), while on the other hand, when a user performs a cold-start check-in, the effect of this behavior is more likely from his social ties than his historical ties, which indicates the chance to study the correlation between such check-ins and his social networks, while in turn also provides a feasible perspective of solving the traditional cold-start location recommendation problem. Therefore, we focus on investigating the social correlations with a user's cold-start check-ins by eliminating the historical tie effect to the largest extent.

Figure 2a shows the percentage of cold-start check-ins over the total number of observed check-ins in a period of a half year (January 1, 2011 to June 30, 2011) with 11,326 users and 1,171,521 check-ins on Foursquare (more details about this dataset in Sect. 4.1). The x-axis represents the number of observed check-ins in a chronological order, and the y-axis represents the percentage of cold-start check-ins. There are around 50 % cold-start check-ins within 2×10^5 observed check-ins, and around 35 % 'cold-start" check-ins among 1.2×10^6 observed check-ins, indicating that a user would like to go to a new location when he does not have much check-in history at early time; and then, as time goes by, the user would gradually shift his check-ins from new locations to existing locations. Furthermore, the high cold-start check-in ratio suggests cold-start check-ins take a big proportion of a user's check-in behavior. With half-year check-in history, a user would still have approximated one third probability to perform a cold-start check-in. Therefore, capturing a user's cold-start check-in location is necessary for designing improving location recommendation services.

Social scientists found that geographical distance plays an important role in social connections (Mok et al. 2010; Goldenberg and Levy 2009; Cairncross 2001). Previous work on LBSNs studied the spatial property of social networks, and reported that the probability of having a social connection between two individuals is a function of their distance (Scellato et al. 2011b). Therefore, to study the social correlation of a user's cold-start check-in behavior, we divide the social correlations into four sub-correlations, namely geo-social correlations, corresponding to four social circles with respect to the factors of social friendship and geographical distance. The confusing matrix of the four social circles is listed in Table 1, where F indicates observed social friendship, $\bar{F}$ indicates non-friendship, D indicates long geographical distance, and $\bar{D}$ indicates short geographical distance.

Table 1 Geo-social correlations

	F	$\bar{F}$
$\bar{D}$	$S_{F\bar{D}}$: Local friends	$S_{\bar{F}\bar{D}}$: Local non-friends
D	S_{FD}: Distant friends	$S_{\bar{F}D}$: Distant non-friends

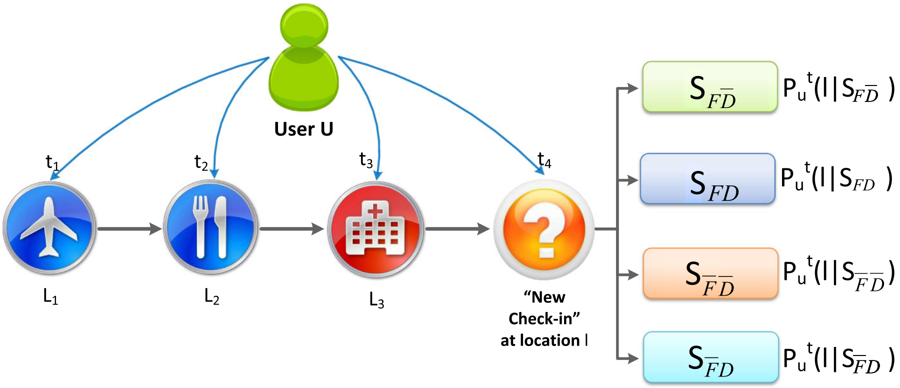

Fig. 1 The geo-social correlations of cold-start check-in behavior

- $S_{F\bar{D}}$: user's social circle consisting of his friends who live close;
- S_{FD}: user's social circle consisting of his friends who live distant;
- $S_{\bar{F}\bar{D}}$: user's social circle consisting of non-friend users who live close; and
- $S_{\bar{F}D}$: user's social circle consisting of non-friend users who live distant.

We define the four social circles as "**geo-social circles**". In Cho et al. (2011), it is reported that the relative influence of a friend who lives 1,000 km away is ten times greater than the influence of a friend who lives 40 km away on a user making check-ins. Therefore in this paper, we consider a pair of users within the same state/province as living close with short geographical distance, and a pair of users in different states/provinces as living distant with long geographical distance.

Figure 1 illustrates a user's cold-start check-in behavior in different social correlation aspects. User u goes to the airport at t_1, and then the restaurant at t_2 followed by the hospital at t_3. When u performs a cold-start check-in at t_4, i.e., the check-in location does not belong to $\{l_1, l_2, l_3\}$, then it may be correlated to those users that are from u's different geo-social circles $S_{F\bar{D}}$, S_{FD}, $S_{\bar{F}\bar{D}}$ and $S_{\bar{F}D}$.

The investigation of geo-social correlations between a user's cold-start check-in behavior and the four geo-social circles, i.e., $S_{F\bar{D}}$, S_{FD}, $S_{\bar{F}\bar{D}}$ and $S_{\bar{F}D}$, enables us to study a user's check-in behavior in four aspects. The geo-social circle $S_{F\bar{D}}$ captures a user's local social correlations, sometimes also including local influence, such as going out with friends, or following friends' recommendations. The geo-social circle S_{FD} captures a user's distant social correlations, such as visiting friends in another state. The third geo-social circle, $S_{\bar{F}\bar{D}}$, indicates that a user goes to a place where his local non-friends usually go to, usually referred to as "confounding" effect (Easley and Kleinberg 2010). The last geo-social circle, i.e., $S_{\bar{F}D}$, suggests that a user would randomly visit some new locations due to an unknown effect regardless of what his friends or local users do. This could be, for example, visiting famous points of interest. Note that there could be some cold-start check-ins that cannot be correlated to any of the four geo-social circles. In our foursquare data, we found that such kind of cold-start check-ins only correspond to a small proportion (to discuss later in Table 3), therefore we consider it as an unknown effect and combine it to $S_{\bar{F}D}$ as well.

3 Modeling geo-social correlations

3.1 Problem formulation

To model the geo-social correlations of a user's cold-start check-in behavior, we consider the probability of a user u checking-in at a new location l at time t as $P_u^t(l)$. With the four geo-social circles defined above, we further define this probability as a combination of the four geo-social correlations,

$$\begin{aligned} P_u^t(l) = {} & \Phi_1 P_u^t(l|S_{\bar{F}\bar{D}}) + \Phi_2 P_u^t(l|S_{F\bar{D}}) \\ & + \Phi_3 P_u^t(l|S_{FD}) + \Phi_4 P_u^t(l|S_{\bar{F}D}). \end{aligned} \tag{1}$$

where Φ_1, Φ_2 and Φ_3 and Φ_4 are correlation strength of different geo-social correlations, $P_u^t(l|S_x)$ indicates the geo-social correlation probability, which is the probability of user u checking-in at location l that is correlated to u's geo-social circle S_x. For example, $P_u^t(l|S_{FD})$ indicates the probability of user u checking-in at l that is correlated to u's distant friends. In the following sections, we will further discuss how to model the geo-social correlation strength and correlation probabilities.

3.2 Modeling geo-social correlation strength

To explicitly model the correlation strength Φ_1, Φ_2, Φ_3 and Φ_4, we investigate the intrinsic patterns of correlations between a user's check-ins and his geo-social circles. We plot the percentage of cold-start check-ins that can be found from the different geo-social circles versus the total number of observed cold-start check-ins in Fig. 2b, with the same data set used in Fig. 2a. The x-axis represents the number of observed cold-start check-ins in a chronological order, and the y-axis represents the percentage of cold-start check-in locations that have been checked-in before by users from that specific geo-social circle. For example, the blue line represents the percentage of cold-

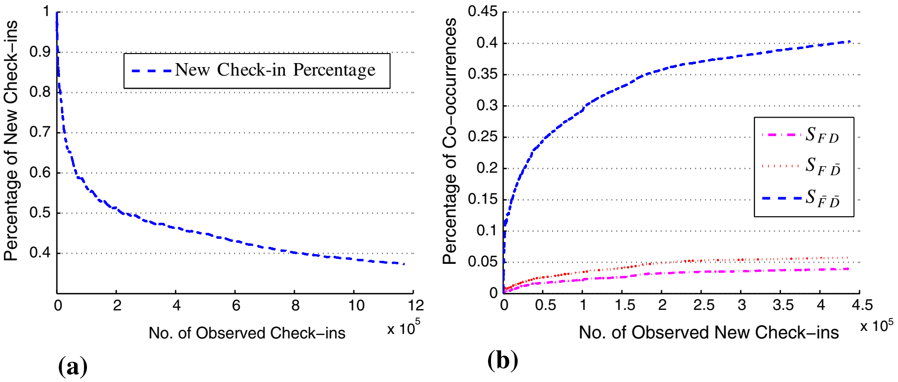

Fig. 2 The cold-start check-in rate and social correlation on Foursquare Data, **a** The ratio of cold-start check-ins, **b** observed social correlations on cold-start check-ins

start check-in locations that have been visited by the user's local non-friends before. The percentage of cold-start check-ins from $S_{\bar{F}D}$ is not presented, since it can be deduced from the other three. Note that the geo-social correlations of the four geo-social circles may overlap. For example, a user may visit a new location l where both of his local friends and distant friends have visited before.

Equation (1) indicates that with probability Φ_1, the current cold-start check-in is correlated to $S_{\bar{F}\bar{D}}$. According to the observation in Fig. 2b, the correlation between cold-start check-ins and the geo-social circle $S_{\bar{F}\bar{D}}$ (blue line) increases with the increment of the number of observed cold-start check-ins. It keeps increasing rapidly early on, and then gradually becomes stable. The reason for this trend may come from two parts: (1) user u would like to go to new locations when he does not have many historical check-ins, therefore in the early time, a lot of cold-start check-ins correlated to $S_{\bar{F}\bar{D}}$ are observed; and (2) as time goes by, the number of check-ins from u's geo-social circle is also increasing, which provides opportunities of co-occurrent check-ins between u and his geo-social circle, hence the social correlation keeps increasing. Therefore, we set Φ_1 as an active function to control the social correlation strength from local non-friend users, which considers a set of features capturing u's historical check-in behavior and his different geo-social circles.

$$\Phi_1 = f\left(\mathbf{w}^T \mathbf{f}_u^t + b\right), \quad 0 \leq \Phi_1 \leq 1, \tag{2}$$

where $\mathbf{f}_u^t$ is a check-in feature vector of a single user u at time t, $\mathbf{w}$ is a vector of the weights of $\mathbf{f}_u^t$, and b controls the bias. In this work, we define a user's check-in and social features $\mathbf{f}_u^t$ in Table 2. Note that $\mathbf{f}_u^t$ is time sensitive, where all the features in

Table 2 Check-in and social features

Features	Description
N^c	Number of check-ins in u's history
N^{nc}	Number of cold-start check-ins in u's history
$N_{F\bar{D}}$	Number of friends in $S_{F\bar{D}}$
$N^c_{F\bar{D}}$	Number of check-ins from $S_{F\bar{D}}$
$N^{uc}_{F\bar{D}}$	Number of unique check-ins from $S_{F\bar{D}}$
$N^{vc}_{F\bar{D}}$	Number of visited check-ins from $S_{F\bar{D}}$
$N^{uvc}_{F\bar{D}}$	Number of visited unique check-ins from $S_{F\bar{D}}$
N_{FD}	Number of friends in S_{FD}
N^c_{FD}	Number of check-ins from S_{FD}
N^{uc}_{FD}	Number of unique check-ins from S_{FD}
N^{vc}_{FD}	Number of visited check-ins from S_{FD}
N^{uvc}_{FD}	Number of visited unique check-ins from S_{FD}
$N_{\bar{F}\bar{D}}$	Number of users in $S_{\bar{F}\bar{D}}$
$N^c_{\bar{F}\bar{D}}$	Number of check-ins from $S_{\bar{F}\bar{D}}$
$N^{uc}_{\bar{F}\bar{D}}$	Number of unique check-ins from $S_{\bar{F}\bar{D}}$
$N^{vc}_{\bar{F}\bar{D}}$	Number of visited check-ins from $S_{\bar{F}\bar{D}}$
$N^{uvc}_{\bar{F}\bar{D}}$	Number of visited unique check-ins from $S_{\bar{F}\bar{D}}$

$\mathbf{f}_u^t$ are computed at time t, and $S_{F\bar{D}}$, S_{FD}, and $S_{\bar{F}\bar{D}}$ are related to user u's geo-social circles.

$f(\bullet)$ is a real-valued and differentiable function that guarantees the range of Φ_1 limited in [0, 1]. In this case, a sigmoid function is often used (Anderson et al. 1986), which can approximately capture the observations about $S_{\bar{F}\bar{D}}$ in Fig. 2b.

$$f\left(\mathbf{w}^T\mathbf{f}_u^t + b\right) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{f}_u^t+b)}}, \tag{3}$$

Similarly, we observe that the social correlations of S_{FD} and $S_{F\bar{D}}$ are fairly constant in Fig. 2b, therefore we define,

$$\begin{aligned}\Phi_2 &= (1-\Phi_1)\phi_1\\ \Phi_3 &= (1-\Phi_1)(1-\phi_1)\phi_2\\ \Phi_4 &= (1-\Phi_1)(1-\phi_1)(1-\phi_2),\end{aligned} \tag{4}$$

where $\phi_1 \in [0, 1]$, $\phi_2 \in [0, 1]$ are two constants to govern the social correlation strength of local friends and distant friends respectively.

Based on above definitions, we can rewrite the probability $P_u^t(l)$ in Eq. (1) as below,

$$\begin{aligned}P_u^t(l) = f\left(\mathbf{w}^T\mathbf{f}_u^t + b\right) & P_u^t(l|S_{\bar{F}\bar{D}})\\ &+ \left(1 - f(\mathbf{w}^T\mathbf{f}_u^t + b)\right)\phi_1 P_u^t(l|S_{F\bar{D}})\\ &+ \left(1 - f(\mathbf{w}^T\mathbf{f}_u^t + b)\right)(1-\phi_1)\phi_2 P_u^t(l|S_{FD})\\ &+ \left(1 - f(\mathbf{w}^T\mathbf{f}_u^t + b)\right)(1-\phi_1)(1-\phi_2) P_u^t(l|S_{\bar{F}D}).\end{aligned} \tag{5}$$

3.3 Modeling geo-social correlation probabilities

In this section, we discuss the modeling of geo-social correlation probabilities, i.e., $P_u^t(l|S_x)$, representing the probability of user u checking in at location l at time t that is correlated to u's social circle S_x, $S_x = \{S_{FD}, S_{F\bar{D}}, S_{\bar{F}D}, S_{\bar{F}\bar{D}}\}$.

Gao et al. (2012b) reported that check-in sequence and text sentence share a large number of common properties, where a check-in location can be analog to a word. Thus, inspired by the "TF-IDF" strategy which is commonly used in text mining and information retrieval to determine the importance of a word, we propose *location frequency* (LF), *user frequency* (UF), and LF.UF correspondingly. The underlying assumption of "LF" is that a user tends to go to a place where his friend usually goes to. Previous work has also reported such property that the number of check-ins previously made by friends of a user is a good predictor for the user's next check-in location (Chang and Sun 2011). Furthermore, to consider the uncertainty of normalization, we also propose a normalized version of "LF", i.e., "NLF". On the other hand, the underlying assumption of "UF" is that a user tends to go to a place where many of his friends have been to, which can be considered as another way to determine the importance of check-in location.

In Cho et al. (2011), *user similarity* is also considered as important to explain the user's check-in behavior as different friendship may present different behavior similarity. In this work, to examine the probability $P_u^t(l|S_x)$, we first propose five geo-social correlation measures according to LF and UF without considering user similarity, and then propose another five geo-social correlations measures with user similarity accordingly, as described below,

- **Location frequency (LF)**

  A user may go to a new location that has been frequently visited by his geo-social circle before, therefore we define the probability of a user u checking at location l at time t that is correlated with his geo-social circle $\mathcal{S}_x$ as:

  $$P_u^t(l|\mathcal{S}_x) = \frac{\sum_{v \in \mathcal{S}_x} N_v^t(l)}{\sum_{v \in \mathcal{S}_x} N_v^t}, \tag{6}$$

  where $N_v^t(l)$ represents the number of check-ins at location l by user v before time t, and N_v^t the total number of locations visited by user v that user u has not visited before time t.

- **Normalized location frequency (NLF)**

  Normalized location frequency (**NLF**) calculates the ratio of check-ins at location l for each user in S_x, and then normalized by the total number of users. The purpose of introducing this normalized measure of **LF** is due to the uncertainty of normalization in improving performance of user behavior modeling (Wang et al. 2010).

  $$P_u^t(l|\mathcal{S}_x) = \frac{\sum_{v \in \mathcal{S}_x} \frac{N_v^t(l)}{N_v^t}}{N_{S_x}}, \tag{7}$$

  where N_{S_x} represents the number of users in S_x.

- **User frequency (UF)**

  User frequency (**UF**) computes the probability $P_u^t(l|\mathcal{S}_x)$ as the ratio of users in $\mathcal{S}_x$ who have checked-in at l,

  $$P_u^t(l|\mathcal{S}_x) = \frac{\sum_{v \in \mathcal{S}_x} \delta_v^t(l)}{N_{\mathcal{S}_x}}, \tag{8}$$

  where $\delta_v^t(l)$ equals to 1 if user v has checked-in at l before t, and 0 otherwise.

- **Location frequency & user frequency (LF.UF)**

  As reported in Gao et al. (2012b), location sequences and document segments share a lot of common features. Traditional language model on language processing also achieves good performance when applied to the location prediction task. Therefore, inspired by the *Tf-idf* (a state-of-the-art weighting strategy widely used in language processing and information retrieval), we propose a **LF.UF** strategy to explore the geo-social correlations on LBSNs.

$$P_u^t(l|\mathcal{S}_x) = \frac{\sum_{v\in\mathcal{S}_x} N_v^t(l)}{\sum_{v\in\mathcal{S}_x} N_v^t} \cdot \frac{\sum_{v\in\mathcal{S}_x} \delta_v(l)}{N_{\mathcal{S}_x}}, \tag{9}$$

– **Normalized location frequency & user frequency (NLF.UF)**

Similar to the **LF.UF** measure, **NLF.UF** is defined as,

$$P_u^t(l|\mathcal{S}_x) = \frac{\sum_{v\in\mathcal{S}_x} \frac{N_v^t(l)}{N_v^t}}{N_{\mathcal{S}_x}} \cdot \frac{\sum_{v\in\mathcal{S}_x} \delta_v(l)}{N_{\mathcal{S}_x}}, \tag{10}$$

To integrate the effect of user similarities, we further propose another five measures that consider user similarities, corresponding to the five measures above.

– **Sim-location frequency (S.LF)**

$$P_u^t(l|\mathcal{S}_x) = \frac{\sum_{v\in\mathcal{S}_x} s(u,v)N_v^t(l)}{\sum_{v\in\mathcal{S}_x} s(u,v)N_v^t}, \tag{11}$$

where $s(u, v)$ represents the user similarity between user u and user v.

– **Sim-normalized location frequency (S.NLF)**

$$P_u^t(l|\mathcal{S}_x) = \frac{\sum_{v\in\mathcal{S}_x} s(u,v)\frac{N_v^t(l)}{N_v^t}}{\sum_{v\in\mathcal{S}_x} s(u,v)}, \tag{12}$$

– **Sim-user frequency (S.UF)**

$$P_u^t(l|\mathcal{S}_x) = \frac{\sum_{v\in\mathcal{S}_x} \delta_v(l)s(u,v)}{\sum_{v\in\mathcal{S}_x} s(u,v)}, \tag{13}$$

– **Sim-location frequency & user frequency (S.LF.UF)**

$$P_u^t(l|\mathcal{S}_x) = \frac{\sum_{v\in\mathcal{S}_x} s(u,v)N_v^t(l)}{\sum_{v\in\mathcal{S}_x} s(u,v)N_v^t} \frac{\sum_{v\in\mathcal{S}_x} \delta_v(l)}{N_{\mathcal{S}_x}}, \tag{14}$$

– **Sim-normalized location frequency & user frequency (S.NLF.UF)**

$$P_u^t(l|\mathcal{S}_x) = \frac{\sum_{v\in\mathcal{S}_x} s(u,v)\frac{N_v^t(l)}{N_v^t}}{\sum_{v\in\mathcal{S}_x} s(u,v)} \frac{\sum_{v\in\mathcal{S}_x} \delta_v(l)}{N_{\mathcal{S}_x}}. \tag{15}$$

In our model, the correlation probability from each circle, i.e., $P_u^t(l|\mathcal{S}_{\bar{F}\bar{D}})$, $P_u^t(l|\mathcal{S}_{F\bar{D}})$, $P_u^t(l|\mathcal{S}_{FD})$, and $P_u^t(l|\mathcal{S}_{\bar{F}D})$, can be calculated with various measures. Due to the different user and check-in distributions of each geo-social circle, the measures may perform variously. Therefore measure selection is necessary for each sub-correlation to achieve a better model performance. We'll discuss the measure performance and selection in the experiment section.

3.4 Parameter inference

With the definitions described in the last section, we discuss the process of inferring the parameters defined in Eq. (5). We define (u, l, t) as a check-in action at location l performed by user u at time t, the likelihood of the observation over the whole data set is the product of the probability of each (u, l, t) action, defined as:

$$P(\mathcal{C}|\Theta) = \prod_{(u,l,t)\in\mathcal{C}} P_u^t(l), \tag{16}$$

where $\mathcal{C}$ is the set of all the observed (u, l, t) actions, and Θ is the parameter set consisting of $\mathbf{w}$, b, ϕ_1, ϕ_2. We learn these parameters through maximum likelihood, which is equivalent to the following minimization problem:

$$\min \sum_{(u,l,t)\in\mathcal{C}} - \ln P(\mathcal{C}|\Theta) + \lambda \left(||\mathbf{w}||_2^2 + ||b||_2^2 + ||\phi_1||_2^2 + ||\phi_2||_2^2 \right) \tag{17}$$

where parameter λ controls the quadratic regularized term to avoid overfitting. In this paper, we set the value of λ as 0.05, and get the objective function below,

$$\begin{aligned}
\min \sum_{(u,l,t)\in\mathcal{C}} & - \ln \Big(f(\mathbf{w}^T \mathbf{f}_u^t + b) P_u^t(l|S_{\bar{F}\bar{D}}) \\
& + \big(1 - f(\mathbf{w}^T \mathbf{f}_u^t + b)\big)\phi_1 P_u^t(l|S_{F\bar{D}}) \\
& + \big(1 - f(\mathbf{w}^T \mathbf{f}_u^t + b)\big)(1 - \phi_1)\phi_2 P_u^t(l|S_{FD}) \\
& + \Big(1 - f(\mathbf{w}^T \mathbf{f}_u^t + b)\big)(1 - \phi_1)(1 - \phi_2) P_u^t(l|S_{\bar{F}D})\Big) \\
& + \lambda \Big(||\mathbf{w}||_2^2 + ||b||_2^2 + ||\phi_1||_2^2 + ||\phi_2||_2^2 \Big) \\
& \text{s.t.} \qquad 0 \leq \phi_1 \leq 1, \quad 0 \leq \phi_2 \leq 1
\end{aligned} \tag{18}$$

We take the projected gradient method (Boyd and Vandenberghe 2004) to solve Eq. (18). The basic idea is to update each current parameter towards an optimal direction (determined by the first derivative of the objective function) with an appropriate step size in each learning step. In each step, if the parameter value runs out of the constraints (e.g., $0 \leq \phi_1 \leq 1$, $0 \leq \phi_2 \leq 1$), we project it back to the corresponding range. The process will go iteratively to update the parameters until convergence. As shown below, the parameters are updated as,

$$\begin{aligned}
\mathbf{w} &\leftarrow \mathbf{w} - \gamma_{\mathbf{w}} \nabla_{\mathbf{w}} \\
b &\leftarrow b - \gamma_b \nabla_b
\end{aligned}$$

$$\phi_1 \leftarrow \begin{cases} 0 & \phi_1 - \gamma_{\phi_1}\nabla_{\phi_1} < 0 \\ 1 & \phi_1 - \gamma_{\phi_1}\nabla_{\phi_1} > 1 \\ \phi_1 - \gamma_{\phi_1}\nabla_{\phi_1} & else \end{cases}$$
$$\phi_2 \leftarrow \begin{cases} 0 & \phi_2 - \gamma_{\phi_2}\nabla_{\phi_2} < 0 \\ 1 & \phi_2 - \gamma_{\phi_2}\nabla_{\phi_2} > 1 \\ \phi_2 - \gamma_{\phi_2}\nabla_{\phi_2} & else \end{cases} \tag{19}$$

where $\gamma_{\mathbf{w}},\ \gamma_b,\ \gamma_{\phi_1}$ and γ_{ϕ_2} are learning step sizes, which are chosen to satisfy Goldstein conditions (Nocedal and Wright 1999). $\nabla_{\mathbf{w}},\ \nabla_b,\ \nabla_{\phi_1}$ and ∇_{ϕ_2} are the partial derivatives of the objective function in Eq. (18) with respect to $\mathbf{w},\ b,\ \phi_1$ and ϕ_2 respectively,

$$\begin{aligned} \nabla \mathbf{w} &= 2\lambda \mathbf{w} - \sum_{(u,l,t)\in\mathcal{C}} \frac{B}{A}\frac{e_1}{(1+e_1)^2}\mathbf{f}_u^t \\ \nabla b &= 2\lambda b - \sum_{(u,l,t)\in\mathcal{C}} \frac{B}{A}\frac{e_1}{(1+e_1)^2} \\ \nabla \phi_1 &= 2\lambda \phi_1 - \sum_{(u,l,t)\in\mathcal{C}} \frac{(1-\Phi_1)}{A}C \\ \nabla \phi_2 &= 2\lambda \phi_2 - \sum_{(u,l,t)\in\mathcal{C}} \frac{(1-\Phi_1)(1-\phi_1)}{A}D \end{aligned} \tag{20}$$

where

$$\begin{aligned} e_1 &= e^{-(\mathbf{w}^T\mathbf{f}_u^t+b)} \\ A &= \Phi_1 P_u^t(l|S_{\bar{F}\bar{D}}) + (1-\Phi_1)\phi_1 P_u^t(l|S_{F\bar{D}}) \\ &\quad + (1-\Phi_1)(1-\phi_1)\phi_2 P_u^t(l|S_{FD}) \\ &\quad + (1-\Phi_1)(1-\phi_1)(1-\phi_2) P_u^t(l|S_{\bar{F}D}), \\ B &= P_u^t(l|S_{\bar{F}\bar{D}}) - \phi_1 P_u^t(l|S_{F\bar{D}}) - (1-\phi_1)\phi_2 P_u^t(l|S_{FD}) \\ &\quad - (1-\phi_1)(1-\phi_2) P_u^t(l|S_{\bar{F}D}) \\ C &= P_u^t(l|S_{F\bar{D}}) - \phi_2 P_u^t(l|S_{FD}) - (1-\phi_2) P_u^t(l|S_{\bar{F}D}) \\ D &= P_u^t(l|S_{FD}) - P_u^t(l|S_{\bar{F}D}) \end{aligned} \tag{21}$$

4 Experiments

In this work, we use location prediction to evaluate our proposed geo-social correlation model (**gSCorr**).[3] In particular, we evaluate the following: (1) how well the proposed geo-social correlation measures capture the geo-social correlation probabilities; (2) how the geo-social correlation strengths and measures affect the cold-start check-in behavior; and (3) whether social correlations help cold-start location recommendation. Before we delve into experiment details, we first discuss an LBSN dataset and experiment settings.

[3] The code can be downloaded at http://www.public.asu.edu/~hgao16/code/gSCorr.zip.

Table 3 Statistical information of the dataset

Duration	January 1, 2011–December 31, 2011
No. of users	11,326
No. of check-ins	2,290,997
No. of unique locations	187,218
No. of links	47,164
Average check-ins per user	202
Clustering coefficient	0.1560
Average degree	8.33

4.1 Data collection

We use a Foursquare dataset[4] to study the geo-social correlations of check-in behavior on location-based social networks. Foursquare is one of the most popular online LBSNs. It has more than 20 million users and 2 billion check-ins as of April, 2012.[5] The web site itself does not provide a public API to access users' check-in data, however, it provides an alternative way for users to link their twitter accounts with Foursquare, and then pop out the check-in messages as tweets to Twitter. Previous work (Scellato et al. 2011b; Gao et al. 2012b) uses this way to collect the data from Twitter for studying check-in behavior. Similarly, by getting access to the check-in tweets through the Twitter REST API, we collected public Foursquare check-in data from January 2011 to December 2011. We also collected the user friendships and hometown information through Foursquare. Note that the friendships on Foursquare are undirected. The statistics of the final dataset are shown in Table 3. The user distributions w.r.t. the world and the USA are given in Fig. 3a, b, respectively.

4.2 Experiment setup

We test our proposed model **gSCorr** on the data of each month from July to December respectively, with the corresponding training data from the previous 6 months to learn our model parameters as in Eq. (19). For example, when testing **gSCorr** on September data, we use the data from March to August to train our model.

For each month from July to December, we construct its test set and ground truth based on the observation of their corresponding cold-start check-in distributions in four geo-social circles. Table 4 lists detailed statistical information of the observed cold-start check-in distribution in four geo-social circles on the check-in data in July. Due to the space limit, we do not present the statistical information from the other months since they all have the similar distributions. We define ***"Social Co-occurrence Check-ins" (SCCs)*** as the cold-start check-ins whose check-in locations can be found from the user's different social circles before its checking in time. The check-in data

[4] The dataset is publically available at http://www.public.asu.edu/~hgao16/dataset.html.

[5] https://foursquare.com/about/.

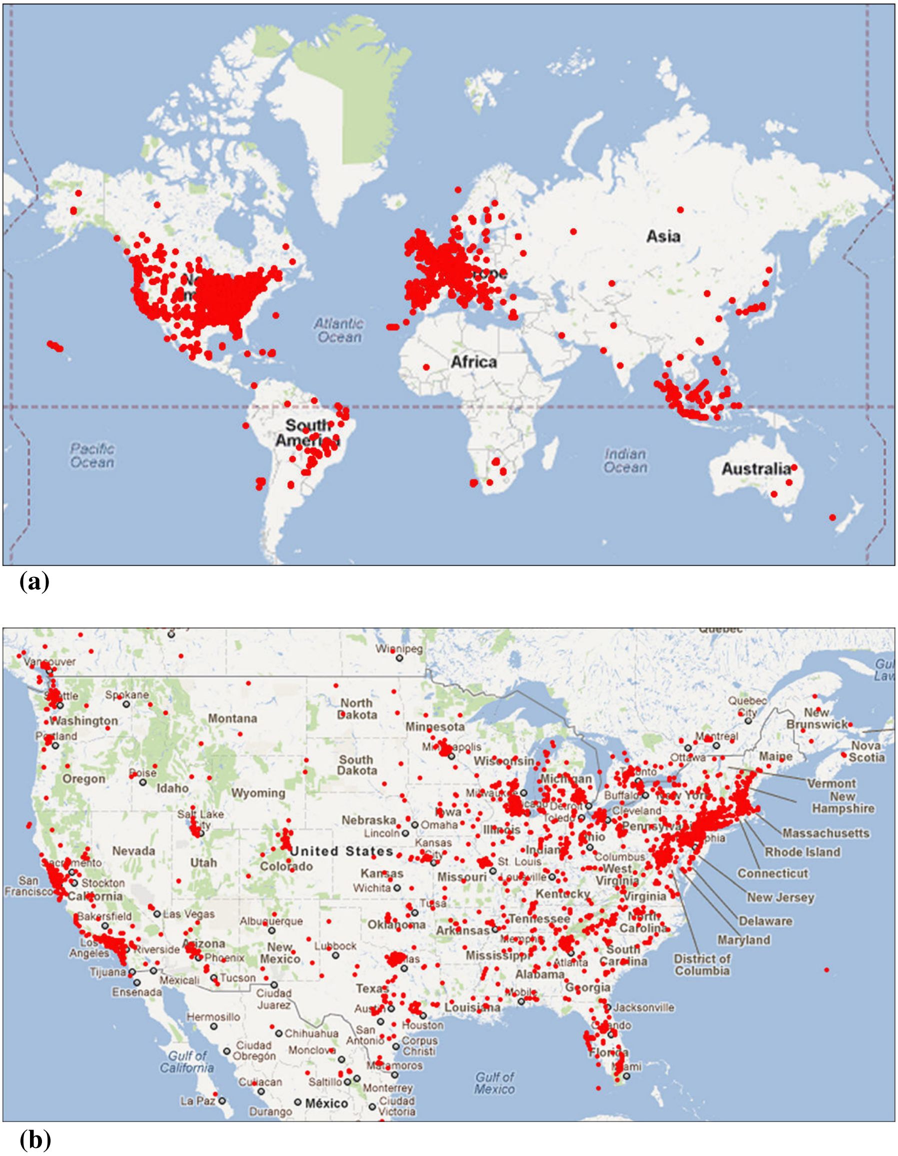

Fig. 3 The user distribution on Foursquare, **a** the user distribution over the world, **b** the user distribution over the USA

in July contains 213,702 check-ins, with 77,581 cold-start check-ins performed at the locations that have never been visited before (the July test data is a closed set in the sense that it does not consider the historical check-ins before July, as the same as the test data from other months). Among the 77,581 cold-start check-ins, around 44.5 % SCCs can be found from the $S_{\bar{F}\bar{D}}$, 7.26 % from $S_{F\bar{D}}$, 4.62 % from S_{FD} and 50.82 % from $S_{\bar{F}D}$. Only 10.61 % SCCs are from a user's direct friendship circle. In other words, only 8,235 among 77,581 cold-start check-ins co-occurred with check-ins of the user's

Table 4 Statistical information of the July data

Social circle	No. of SCCs	Ratio (%)
$S_{\bar{F}\bar{D}}$	34,523	44.50
$S_{F\bar{D}}$	5,636	7.26
S_{FD}	3,588	4.62
$S_{\bar{F}D}$	39,423	50.82
Others	1,672	2.2
$S_{\bar{F}\bar{D}} \cup S_{F\bar{D}}$	35,277	45.47
$S_{\bar{F}\bar{D}} \cup S_{FD}$	35,784	46.12
$S_{F\bar{D}} \cup S_{FD}$	8,235	10.61
$S_{\bar{F}\bar{D}} \cup S_{F\bar{D}} \cup S_{FD}$	36,486	47.03

friendships. $S_{\bar{F}\bar{D}}$ has a large proportion of co-occurrences, indicating that user would like to go to a new location where his local non-friends in the state usually go. The number of SCCs of $S_{\bar{F}\bar{D}} \cup S_{F\bar{D}} \cup S_{FD}$ doesn't increase much compared to $S_{\bar{F}\bar{D}}$, indicating that local non-friends have already covered most of the co-occurrences. Finally, we found that more than 50 % of SCCs are correlated to $S_{\bar{F}D}$, which is difficult to capture for location prediction as the unknown effect. Note that there are 2.2 % "Others", indicating that at the time of check-in, 1,672 cold-start check-ins cannot be found from any of the four social circles. We consider this as an unknown effect and merge it into $S_{\bar{F}D}$.

We use location recommendation to evaluate our correlation measures and model performance. The **user similarities** are computed based on the check-in data in the first half year by cosine similarity, while each user is represented by a check-in vector, and the entry in the vector indicates the visiting frequency of the user at the location. For each test month, the **test set** is selected as the SCCs of $S_{\bar{F}\bar{D}} \cup S_{F\bar{D}} \cup S_{FD}$, and the **ground truth** is the corresponding check-in locations. We do not consider $S_{\bar{F}D}$ because from a user's perspective, friends and local non-friends are the ones that are reachable, while the distant non-friend users are too weak in relation for the user to correlate.

4.3 Geo-social correlation measure selection

Before we discuss the performance of our proposed model **gSCorr**, we first evaluate the ten geo-social correlation measures described in Sect. 3.3. Each measure can be directly applied to the test set and generates a ranking list of location probabilities $P_u^t(l|S_x)$ with respect to the geo-social circles. We select the location with the highest $P_u^t(l|S_x)$ as recommended location for the cold-start check-in, and evaluate the performance with accuracy. The purpose of this comparison is to select the best correlation measure for each geo-social circle, and utilize the most suitable ones for $P_u^t(l|S_x)$ in Eq. (1). The results are shown in Tables 5, 6 and 7 with some observations summarized below:

- The user similarity consistently improves the recommendation performance. Comparing with measures without considering user similarities, measures with user similarities on average have 5.36 % relative improvement on $S_{F\bar{D}}$, 30.53 % relative

Table 5 Location recommendation for measure selection on $S_{F\bar{D}}$

Ranking strategy	$S_{F\bar{D}}$					
	July (%)	August (%)	September (%)	October (%)	November (%)	December (%)
LF	5.85	6.24	6.49	6.76	6.47	7.29
NLF	5.60	6.07	6.11	6.42	6.32	6.88
UF	5.18	5.54	5.68	5.70	5.69	6.33
LF.UF	6.16	6.50	6.72	6.99	6.70	7.33
NLF.UF	5.92	6.39	6.49	6.78	6.58	7.22
S.LF	6.30	6.73	6.99	7.32	7.04	**7.90**
S.NLF	5.89	6.31	6.34	6.64	6.62	7.21
S.UF	5.38	5.83	5.77	5.97	5.96	6.58
S.LF.UF	**6.51**	**6.85**	**7.02**	**7.37**	**7.11**	7.76
S.NLF.UF	6.23	6.68	6.75	7.07	6.92	7.55

The best performance of each month in bold

Table 6 Location recommendation for measure selection on S_{FD}

Ranking strategy	S_{FD}					
	July (%)	August (%)	September (%)	October (%)	November (%)	December (%)
LF	2.39	2.39	2.91	3.35	3.26	3.65
NLF	2.65	2.50	3.07	3.39	3.33	3.38
UF	2.23	2.22	2.66	3.15	3.03	3.15
LF.UF	2.65	2.70	3.20	3.66	3.52	3.59
NLF.UF	2.79	2.70	3.26	3.64	3.52	3.56
S.LF	**3.65**	3.52	4.15	**4.63**	**4.37**	**4.91**
S.NLF	3.45	3.46	3.92	4.31	4.14	4.40
S.UF	3.14	3.00	3.43	3.86	3.76	4.01
S.LF.UF	3.64	**3.57**	**4.19**	4.56	4.31	4.64
S.NLF.UF	3.58	3.52	4.11	4.47	4.25	4.47

The best performance of each month in bold

improvement on S_{FD}, and 15.89 % relative improvement on $S_{\bar{F}\bar{D}}$, suggesting that user similarity is a significant factor to capture human mobile behavior.

- The comparison of **LF** and its normalized version **NLF** (including those measures containing **LF** and **NLF**) indicates that normalization does not always improve the performance, which is consistent to the findings in Wang et al. (2010). Depending on which social circle we apply the measure to, normalization may result in various performances in capturing the geo-social correlations.
- **S.Lf.Uf** is the best measure for capturing the social correlations of local friends $S_{F\bar{D}}$. It also performs well on the other two geo-social circles especially on $S_{F\bar{D}}$. It considers the user frequency, location frequency and user similarities together, and obtains 1 % relative improvement compared to the second best rated (**S.LF**), and 24.88 % relative improvement compared to the worst rated (**Uf**).

Table 7 Location recommendation for measure selection on $S_{\bar{F}\bar{D}}$

Ranking strategy	$S_{\bar{F}\bar{D}}$					
	July (%)	August (%)	September (%)	October (%)	November (%)	December (%)
LF	14.42	15.47	15.34	16.17	16.32	18.70
NLF	15.23	16.30	16.35	17.32	17.45	19.67
UF	15.35	16.54	16.50	17.64	17.77	19.81
LF.UF	15.22	16.44	16.40	17.41	17.53	19.87
NLF.UF	15.44	16.59	16.58	17.66	17.77	19.84
S.LF	17.84	18.75	18.60	19.56	19.72	22.38
S.NLF	18.00	19.01	19.04	19.81	19.84	22.39
S.UF	**18.37**	**19.40**	**19.45**	**20.21**	**20.34**	**22.82**
S.LF.UF	17.75	18.86	18.80	19.74	20.10	22.34
S.NLF.UF	17.68	18.66	18.66	19.69	19.80	22.36

The best performance of each month in bold

- **S.Lf** shows good performance in capturing the social correlations of distant friends S_{FD}. It considers the location frequency and user similarity without the user frequency. One possible reason of this may be due to the smaller number of distant friends (2.68 per user on average) compared with the number of local friends (5.64 per user on average), which makes it a weak measure by counting the user frequency of distant friends.
- The performance on $S_{\bar{F}\bar{D}}$ indicates that its best correlation measure is **S.Uf**, suggesting that a user would like to go to a location that has been visited by a large proportion of local non-friend users, no matter how frequently the location is visited by each individual user. This is consistent to the confounding effect that people who live in similar environment tend to share similar behavior, which is exactly the geo-social circle $S_{\bar{F}\bar{D}}$ supposed to capture.

Due to the varied performances of each correlation measure on each geo-social circle, we conclude that measure selection is necessary for computing geo-social correlation probabilities. Hence, we apply **S.Lf.Uf**, **S.Lf** and **S.Uf** to compute $P_u^t(l|S_{F\bar{D}})$, $P_u^t(l|S_{FD})$ and $P_u^t(l|S_{\bar{F}\bar{D}})$ respectively in the following experiments, considering their good performance on the corresponding geo-social circles. We do not report the results on $S_{\bar{F}D}$, since for the unknown effect $P_u^t(l|S_{\bar{F}D})$, all the measures applied to $S_{\bar{F}D}$ perform as a random guess in our experiment, one possible reason may be the large number of users and candidate locations within this geo-social circle. Therefore, to reduce the time complexity, we consider $P_u^t(l|S_{\bar{F}D})$ as a probability of a random jump to a location in current location vocabulary that user u has not checked-in before.

4.4 Performance of **gSCorr**

In this section, we discuss the performance of **gSCorr** on cold-start check-in location recommendation problem with the correlation measures selected in the above section. Note that **gSCorr** is different from traditional recommendation approaches. The traditional recommendation methods, such as collaborative filtering, usually perform recommendation based on a single user-location matrix generated from training data, which ignore the temporal information among user's check-ins. While in **gSCorr**, for each test case (a cold-start check-in location in test set), only check-ins before the check-in time of test case are used for prediction. Of course, methods based on collaborative filtering can still be applied with the consideration of temporal information, by generating a user-location matrix considering only the previous user check-ins for each test case in the test set. We will use this approach as one baseline to compare with **gSCorr** in our experiment, and show its relationship to our proposed correlation measures.

We compare **gSCorr** with four baselines, one is from the observation of the measure selection in Tables 5, 6 and 7, the other three are selected as the existing most popular location recommendation model on LBSNs.

- **S.LF.UF**: We select **S.LF.UF** to capture the geo-social correlations and predict cold-start check-ins. It performs well on all the geo-social circles, and achieve the best performance on $S_{F\bar{D}}$ and many times on $S_{F\bar{D}}$. We apply it to the whole test set to predict the cold-start check-ins.
- **Periodic & social mobility model (PSMM)**: **PSMM** ranks the locations based on a user's periodic and social patterns (Cho et al. 2011). Since the periodic patterns can only recommend existing locations, we adopt the social patterns to recommmend the cold-start check-ins.
- **Social-historical model (SHM)**: **SHM** integrates a user's historical ties and social ties to recommend/predict the next check-in location (Gao et al. 2012b). Similar to PSMM, we leverage the social model which utilizes the social ties to recommend cold-start check-in locations.
- **Collaborative filtering (CF)**: **CF** is a state-of-the-art approach for recommender systems. It computes a user's interest in a location based on other users' interests in that location. Since it can recommend new locations to a user, we apply it to each test case of our test set and consider that a correct recommendation happens when the recommended location is the same as the ground truth of the test case. We choose user-based collaborative filtering for such recommendation (Su and Khoshgoftaar 2009) as shown below:

$$P_u^t(l) = \frac{\sum_{v \in \mathcal{U}} s(u, v) r_{v,l}}{\sum_{v \in \mathcal{U}} s(u, v)}. \tag{22}$$

  where $\mathcal{U}$ is the set of users who have visited l, $r_{v,l}$ is the preference of user v on location l, which in our experiment is chosen as proportional to number of v's check-ins on l normalized by v's total number of check-ins, i.e., $\frac{N_v^t(l)}{N_v^t}$.

The results are shown in Table 8, we summarize several interesting observations below:

- Both **PSMM** and **SHM** do not perform well in recommending the cold-start check-in locations. **SHM** performs better than **PSMM**, but still only achieve a low accuracy.

Table 8 Performance comparison for location recommendation

Dataset	July (%)	August (%)	September (%)	October (%)	November (%)	December (%)
S.LF.UF	18.31	19.58	19.71	20.79	21.10	23.53
PSMM	1.04	1.19	1.24	1.22	1.26	1.23
SHM	5.30	5.08	5.39	5.65	5.03	5.58
CF	18.24	19.57	19.45	20.74	20.84	23.59
gSCorr	**19.21**	**20.25**	**20.36**	**21.26**	**21.42**	**24.13**

The best performance of each month among all the approaches is in bold

They recommend a user's next location based on the observation of his friends' check-in history. The performance indicates that a user does not follow his friends' check-in sequence a lot on LBSNs, especially when performing a cold-start check-in.

- **CF** has comparable performance with **S.LF.UF**. Comparing to Eq. (12), applying **S.NLF.UF** to the whole test set is actually equivalent to user-based collaborative filtering, resulting in a close performance to **S.LF.UF** according to Tables 5, 6 and 7. This also demonstrates the practicability of our proposed correlation measures.
- **gSCorr** performs the best among all the approaches. To demonstrate the significance of its improvement over other baseline methods, we launch a random guess approach to recommend the cold-start check-ins. The recommendation accuracy of the random guess is always below 0.005 %, indicating that **gSCorr** significantly improves the baseline methods.

4.5 Effect of geo-social correlation strength and measures

To further investigate **gSCorr**, we consider the effect of both geo-social correlation strength and measures in capturing the user's cold-start check-in behavior. Therefore, we set up five alternative approaches with various correlation strength and measures, to compare the location recommendation performance with **gSCorr**, as shown in Table 9 with the details below:

- **EsSm**. We select the measure **S.LF.UF**, which works well in all the geo-social circles, to compute the geo-social correlation probabilities for each social circle. We set all the geo-social correlation strength equaling to 1.
- **EsVm**. We select the same measures as in **gSCorr**, but set all the geo-social correlation strength as 1.

Table 9 Evaluation measures

	Single measure	Various measures
Equal strength	EsSm	EsVm
Random strength	RsSm	RsVm
Various strength	VsSm	gSCorr

Table 10 Location recommendation with various geo-social correlation strength and measures

Dataset	July (%)	August (%)	September (%)	October (%)	November (%)	December (%)
EsVm	17.88	18.60	18.86	19.48	19.64	21.94
EsSm	16.20	16.86	17.11	17.94	18.16	20.57
VsSm	16.49	17.94	18.08	18.17	18.45	20.90
RsSm	14.93	15.49	15.88	16.70	16.97	19.29
RsVm	15.23	15.78	16.17	16.81	17.02	19.10
gSCorr (VsVm)	**19.21**	**20.25**	**20.36**	**21.26**	**21.42**	**24.13**

The best performance of each month in bold

- **RsSm**. We select the same measure as in EsSm, and randomly assign the geo-social correlation strength.
- **RsVm**. We select the same measures as in **gSCorr**, and randomly assign the geo-social correlation strength.
- **VsSm**. We select the same measure as in EsSm, and perform the same training procedure to obtain the geo-social correlation strength.

Note that **gSCorr** is a various strength and various measures approach. The results are shown in Table 10; For each random strength approach (RsSm and RsVm), we run 30 times and report the average accuracy. We summarize the essential observations below:

- The geo-social correlations from different geo-social circles contribute variously to a user's check-in behavior. Both ***VsSm*** and ***gSCorr*** perform better than their equal strength versions (i.e., ***EsSm*** and ***EsVm***), respectively, indicating that the geo-social correlations are not equally weighted.
- The randomly assigned strength approaches (***RsSm*** and ***RsVm***) perform the worst comparing to the other approaches, where the average performance of ***VsSm*** has a 9.40 % relative improvement over ***RsSm***, and ***gSCorr*** has a 27.51 % relative improvement over ***RsVm***, indicating that social correlation strength do affect the check-in behavior.
- The single measure approaches (***EsSm***, ***RsSm***, ***VsSm***) always perform worse than the various measures approaches (***EsVm***, ***RsVm***, ***gSCorr***), which suggests that for different social circles, there are different suitable correlation measures.
- **gSCorr** performs the best on all the test data, suggesting the advantage of **gSCorr** as considering different geo-social correlation strength and measures for each geo-social circle, which results in a flexible model for capturing the geo-social correlations on a user's check-in behavior.

4.6 Effect of different geo-social circles

To further investigate the contribution of different geo-social circles, we compare the recommendation results by utilizing various combinations of geo-social circles, as shown in Table 11. The geo-social correlation measures are all selected as the best one for the corresponding social circles, and the geo-social correlation strength is learned in the previous section through **gSCorr**.

Table 11 Location recommendation with various social circle combinations

Methods	July (%)	August (%)	September (%)	October (%)	November (%)	December (%)
$S_{F\bar{D}}$	6.51	6.85	7.02	7.37	7.11	7.76
S_{FD}	3.65	3.52	4.15	4.63	4.37	4.91
$S_{\bar{F}\bar{D}}$	18.37	19.40	19.45	20.21	20.34	22.82
$S_{\bar{F}\bar{D}} \cup S_{F\bar{D}}$	18.62	19.60	19.62	20.49	20.73	23.22
$S_{\bar{F}\bar{D}} \cup S_{FD}$	19.01	20.09	20.23	21.08	21.17	23.97
$S_{F\bar{D}} \cup S_{FD}$	8.33	8.46	9.04	9.43	9.23	10.09
$S_{\bar{F}\bar{D}} \cup S_{F\bar{D}} \cup S_{FD}$	**19.21**	**20.25**	**20.36**	**21.26**	**21.42**	**24.13**

The results show that the social correlations of user's direct friendships S_{FD} and $S_{F\bar{D}}$ are significantly lower than the local non-friend users $S_{\bar{F}\bar{D}}$. The latter contributes more than 95% of accurate recommendation, which indicates that there is a big overlap of check-in locations between local non-friend users and direct friends. On the other hand, the correlations of S_{FD} and $S_{F\bar{D}}$ do not overlap much, where the combination of them has significant improvement over S_{FD} and $S_{F\bar{D}}$ individually. This is due to the diversity of friend distribution since local friends and distant friends do not share much common geographical environment. Furthermore, the combination of $S_{\bar{F}D} \cup S_{F\bar{D}}$ performs much better than $S_{\bar{F}D} \cup S_{FD}$, indicating that local non-friend users share more common check-in locations with local friends than distant friends. Finally, **gSCorr** always performs the best among all the combinations of social circles (in bold font), demonstrating that by taking advantage of both social networks and geographical distance, our approach properly captures the user's cold-start check-in behavior on LBSNs, and could be utilized to benefit cold-start location recommendation.

4.7 Discussion

We summarize the experiment results in this section, and explain a set of observations of user check-in behavior on LBSNs as below:

- Social correlations do exist on LBSNs. The correlation is more relevant to a user's local non-friends than direct social friends, where the latter only contribute a small proportion in a user's check-in behavior. This in turn explains the previous findings (Cho et al. 2011; Gao et al. 2012b; Chang and Sun 2011;Ye et al. 2010, 2011) that utilizing social friends' check-ins can only slightly improve the location prediction/recommendation on LBSNs.
- To capture the social correlations on LBSNs, a set of factors need to be considered, which consists of user similarity, location frequency and user frequency. Furthermore, the factors affect variously on user's different geo-social circles.
- Social correlations can be utilized to solve the cold-start problem to a certain extent. From the results in Tables 10 and 11, **gSCorr** could accurately recommend 19.21 % cold-start check-in locations from $S_{\bar{F}\bar{D}} \cup S_{F\bar{D}} \cup S_{FD}$. Considering the total number of cold-start check-ins (77,581) in the data set, it is equivalent to around 10 % accuracy among the whole dataset, while a random guess of a user's next location

from our testing set is below 0.005 %, indicating that the improvement of **gSCorr** in recommending the cold-start location is actually very significant.

4.8 Limitations of **gSCorr**

gSCorr considers both geographical distance and social friendships, providing a better perspective to compute user similarity for recommendation purpose. According to the comparison between **gSCorr** and state-of-the-art baseline methods, the geographical property does present significant effect in improving the location recommendation. However, there are several limitations of **gSCorr** that could be considered for future improvements.

- **Discrete geographical distance**

  The geographical separation of social relationships in this work is discrete due to the limitation of data availability. According to the observed user profiles, hometown information is usually provided in city or state level. Considering geographical distance in state level may loss valuable local information, while adopting city level distance measure may result in the incorrect use of state level information. One possible way is to consider all the check-ins of a user and take the average; however this strategy is highly affected by the check-in outliers, i.e., check-in locations far away from the hometown. Thus, under which granularity to compute the geographical distance is still an open question. We will continue to study this problem and investigate an appropriate social correlation function of continuous geographical distance.

- **Temporal dynamics of check-in behavior**

  As human movement is a stochastic process over the time, the corresponding geo-social correlations may also change over the time. Adopting geo-social correlation measures analog to “TF-IDF” is under the assumption of “bag of check-ins”, where check-in locations are independent to each other. This may result in the temporal information loss as the older check-in could have a decreasing correlation to the current check-in. We have performed preliminary experiments to evaluate such effect. The experimental results show that geo-social correlations do decrease over the time. In our observations, we have found that using the recent 30 % check-ins in half-year duration from a user is sufficient to compute the user similarity and perform recommendation. This observation could be potentially utilized to improve our algorithm efficiency significantly, as human check-in sequence is usually too long to be efficiently leveraged in similarity computation.

5 Related work

Researchers have investigated the social network and check-in properties on location-based social networks (Gao and Huan 2013). Noulas et al. (2011) studied the spatio-temporal patterns of user activity on Foursquare, and found that the check-in activity varies within the course of a day and a week. Long et al. (2012) investigated the local geographical topics of check-ins with LDA, and studied the spatial and temporal properties of discovered topics. Cheng et al. (2011) reported that user's check-in behavior on LBSNs follows the "Lèvy Flight" mobility pattern, and is influenced by the social status, sentiment and geographic constraints. Gao et al. (2013a, b) investigated the temporal properties of geographical check-ins, and leverage them for location prediction and location recommendation. In Scellato et al. (2010, 2011b), the authors studied the correlation between friends and their average distance on three LBSNs, and observed that the probability of having a social connection between two users is a function of their geographical distance. In Backstrom et al. (2010), the authors found that the probability of friendship is roughly inversely proportional to distance, and this information has been further studied to predict a user's home address with Facebook data.

Efforts have also been made to utilize user's social network information on LBSNs for improving location based services. In Chang and Sun (2011), the authors investigated various features for location prediction on LBSNs, and reported that the number of check-ins made by friends is a significant predictor. In Long and Joshi (2013), the authors proposed a HITS-based POI recommendation algorithm to recommend POIs to LBSN users with the consideration of social relationships. In Ye et al. (2010), the check-in information from nearby friends was utilized for location recommendation while other users were ignored. The results indicate that social network only brings minor improvement. Gao et al. (2012c) investigated geo-social correlations on LBSNs to solve the "cold start" location prediction problem, which can be analog to the location recommendation problem specifically on a user's next check-in. In Ye et al. (2011), the authors utilized both user-based and friend-based collaborative filtering for location recommendation. This approach did not consider the geographical property of social correlations, and could be related to the equal strength and single measure version of our **gSCorr**. Cho et al. (2011) studied the periodic patterns of check-in behavior on LBSNs, and proposed a Gaussian mixture model together with the social network information considered for location prediction, while their results also show limited improvements from social network. Gao et al. (2012b) studied the social-historical ties on Foursquare, and found that both ties have contributions to the user's check-in behavior, while social ties are complementary to the historical ties, especially when the historical model does not perform well due to the long and noisy history.

6 Conclusions and future work

In this paper, we propose a geo-social correlation model to capture the social correlations of check-in behavior on LBSNs. We investigate the correlations in context of social networks and geographical distance. The work presented in this paper suggests many future directions. Firstly, the geographical separation of social relationships in this work is binary. It would be interesting to consider a continuous function of social correlations with the changing of geographical distance. Secondly, in this work we focus on utilizing social network information to solve the cold-start problem, while ignoring a user's own check-in history. In the future we will continue to study how to take advantage of both social correlations and historical check-ins, and explore novel usage of such information.

Acknowledgments This work is supported, in part, by ONR (N000141010091).

References

Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. In: KDD. ACM, New York, pp. 7–15

Anderson J, Michalski R, Michalski R, Carbonell J, Mitchell T (1986) Machine learning: an artificial intelligence approach. Morgan Kaufmann, Burlington

Backstrom L, Sun E, Marlow C (2010) Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th international conference on World wide web. ACM, New York, pp. 61–70

Barnes S, Scornavacca E (2004) Mobile marketing: the role of permission and acceptance. Int J Mobile Commun 2(2):128–139

Boyd SP, Vandenberghe L (2004) Convex optimization. Cambridge university press, Cambridge

Cairncross F (2001) The death of distance: how the communications revolution is changing our lives. Harvard Business Press, Cambridge

Chang J, Sun E (2011) Location 3: how users share and respond to location-based data on social networking sites. In: ICWSM

Cheng Z, Caverlee J, Lee K, Sui D (2011) Exploring millions of footprints in location sharing services. In: ICWSM

Cho E, Myers S, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: KDD. ACM, New York, pp. 1082–1090

Easley D, Kleinberg J (2010) Networks, crowds, and markets. Cambridge University Press, Cambridge

Gao H, Barbier G, Goolsby R (2011) Harnessing the crowdsourcing power of social media for disaster relief. Intell Syst IEEE 26(3):10–14

Gao H, Tang J, Liu H (2012a) Mobile location prediction in spatio-temporal context. In: Nokia mobile data challenge workshop

Gao H, Tang J, Liu H (2012b) Exploring social-historical ties on location-based social networks. In: ICWSM

Gao H, Tang J, Liu H (2012c) gSCorr: modeling geo-social correlations for new check-ins on location-based social networks. In: The 21st ACM International Conference on Information and Knowledge Management

Gao H, Huan (2013) Data analysis on location-based social netwoks. In: Chin A, Zhang D (eds) Mobile social networking: an innovative approach. Springer, Berlin, pp 165–194

Gao H, Tang J, Hu X, Liu H (2013a) Exploring temporal effects for location recommendation on location-based social networks. In: Proceedings of the 7th ACM conference on Recommender systems. ACM, New York, pp. 93–100

Gao H, Tang J, Hu X, Liu H (2013b) Modeling temporal effects of human mobile behavior on location-based social networks. In: Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, New York, pp. 1673–1678

Goldenberg J, Levy M (2009) Distance is not dead: social interaction and geographical distance in the internet era (Arxiv, preprint arXiv:0906.3202)

Goodchild M, Glennon J (2010) Crowdsourcing geographic information for disaster response: a research frontier. Int J Digit Earth 3(3):231–241

Kessler S (2012) Foursquare tops 20 million users. http://mashable.com/2012/04/16/foursquare-20-million/. Accessed 16 Apr 2012

Li S (2011) Location based services marketing. Master thesis, Royal Institute of Technology

Long X, Jin L, Joshi J (2012) Exploring trajectory-driven local geographic topics in foursquare. In: UbiComp, pp. 927–934

Long X, Joshi J (2013) A hits-based poi recommendation algorithm for location-based social networks. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Mok D, Wellman B, Carrasco J (2010) Does distance matter in the age of the internet? Urban Stud 47(13):2747

Monreale A, Pinelli F, Trasarti R, Giannotti F (2009) Wherenext: a location predictor on trajectory pattern mining. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, pp. 637–646

Nocedal J, Wright S (1999) Numerical optimization. Springer, Berlin

Noulas A, Scellato S, Mascolo C, Pontil M (2011) An empirical study of geographic user activity patterns in foursquare. In: ICWSM

Scellato S, Mascolo C, Musolesi M, Latora V (2010) Distance matters: geo-social metrics for online social networks. In: Proceedings of the 3rd conference on online social networks. USENIX Association, Boston, pp. 8–8

Scellato S, Musolesi M, Mascolo C, Latora V, Campbell A (2011a) Nextplace: a spatio-temporal prediction framework for pervasive systems. Pervasive Comput 6696:152–169

Scellato S, Noulas A, Lambiotte R, Mascolo C (2011b) Socio-spatial properties of online location-based social networks. In: ICWSM, p 11

Scellato S, Noulas A, Mascolo C (2011c) Exploiting place features in link prediction on location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp. 1046–1054

Spaccapietra S, Parent C, Damiani M, De Macedo J, Porto F, Vangenot C (2008) A conceptual view on trajectories. Data Knowl Eng 65(1):126–146

Su X, Khoshgoftaar TM (2009) A survey of collaborative filtering techniques. Adv Artif Intell 2009:4

Thanh N, Phuong T (2007) A gaussian mixture model for mobile location prediction. In: 2007 IEEE International Conference on Research, Innovation and Vision for the Future, pp. 152–157

Wang X, Tang L, Gao H, Liu H (2010) Discovering overlapping groups in social media. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on IEEE, pp. 569–578

Ye M, Yin P, Lee W (2010) Location recommendation for location-based social networks. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, New York, pp. 458–461

Ye M, Yin P, Lee W, Lee D (2011) Exploiting geographical influence for collaborative point-of-interest recommendation. In: Annual International ACM SIGIR Conference on Research and Development in, Information Retrieval, pp. 325–334

Zickuhr K, Smith A (2011) 28 % of american adults use mobile and social location-based services Pew Internet and American Life Center Report. Accessed 8 Sept 2011