Exploring Social-Historical Ties on Location-Based Social Networks

Huiji Gao, Jiliang Tang, and Huan Liu
Computer Science and Engineering
Arizona State University
{Huiji.Gao, Jiliang.Tang, Huan.Liu}@asu.edu

Abstract

Location-based social networks (LBSNs) have become a popular form of social media in recent years. They provide location related services that allow users to "check-in" at geographical locations and share such experiences with their friends. Millions of "check-in" records in LBSNs contain rich information of social and geographical context and provide a unique opportunity for researchers to study user's social behavior from a spatial-temporal aspect, which in turn enables a variety of services including place advertisement, traffic forecasting, and disaster relief. In this paper, we propose a social-historical model to explore user's check-in behavior on LBSNs. Our model integrates the social and historical effects and assesses the role of social correlation in user's check-in behavior. In particular, our model captures the property of user's check-in history in forms of power-law distribution and short-term effect, and helps in explaining user's check-in behavior. The experimental results on a real world LBSN demonstrate that our approach properly models user's check-ins and shows how social and historical ties can help location prediction.

Introduction

Social media extends the physical boundary of user activities. As a new type of online social media, location-based social networks (LBSNs) provide location services and allow users to share their locations with friends and find others who are nearby. A recent survey reports that 4% of people in the United States use location services like Foursquare[1], Gowalla[2] and Facebook Places[3]; About 1% of Internet users are using these services daily (Zickuhr and Smith 2010). Such location-based social networks form a new generation of online social media with both user's social friendships and his historical geographical trajectory, which provides challenges for researchers in investigating a user's social behavior in a spatial-temporal aspect.

People share their experiences and interact with friends through LBSNs via various online activities such as making online friends, sharing events, checking in, etc. Among these activities, "checking in" (an online activity that tells your friends when and where you are through social media) is a typical online action that reflects an actual interaction between a user and the real world, which is different from many other online activities (following, grouping, voting, tagging, etc.) in which users interact in the virtual world. Hence, it provides opportunities to study a user's real world behavior through virtual media, and devise location-based services such as mobile recommendation (Barwise and Strong 2002) and disaster relief (Gao, Barbier, and Goolsby 2011).

To understand a user's check-in behavior, the historical analysis of the user is inevitable, because the historical check-ins provide rich information about a user's interests and hints about when and where a particular user would like to go. In addition, social correlation (Anagnostopoulos, Kumar, and Mahdian 2008) suggests to consider users' social ties since human movement is usually affected by their social context, such as visiting friends, going out with colleagues, traveling while following friends' recommendations, and so on. These two relationship ties can shape the user's check-in experience on LBSNs, while each tie gives rise to a different probability of check-in activity, which indicates that people in different spatial-temporal-social circles have different interactions. Thus, exploring a user's social-historical ties is crucial to analyze his check-in behavior and therefore understand the corresponding movement. Sociologists studied the effect of social-historical context in realms of sports (Hargreaves 1986), sociologies (Gordon 1973), economy (Hodgson 2001), disaster (Quarantelli 1987), and so on. In this paper, we propose to investigate the effect of social-historical ties on users' check-in behavior in the real world, and understand how social and historical ties affect users' behavior through LBSNs. Figure 1 illustrates an example of the social-historical tie effect on user's check-ins from time T_1 to T_5. A user's next check-in could be affected by his historical check-ins and social ties, while historical check-ins and social ties have varying tie strengths represented by line thickness.

The historical ties of a user's check-in behavior have two properties on LBSNs. First, a user's check-in history approximately follows a power-law distribution, i.e., a user goes to a few places many times and to many places a few times. Second, the historical ties have short-term effect. As illus-

[1]http://foursquare.com/

[2]http://gowalla.com/

[3]http://www.facebook.com/facebookplaces

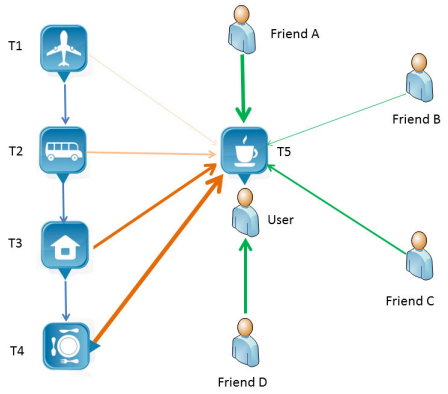


Figure 1: An example: How social and historical ties may affect a user's check-ins

trated in Figure 1, a user arrives at the airport and then takes a shuttle to the hotel. After his dinner, he sips a cup of coffee. The historical ties of the previous check-ins at airport, shuttle stop, hotel and restaurant have different strengths to the latest check-in of drinking coffee. Furthermore, the historical tie strength decreases over time.

For the purposes of understanding the social ties of users' check-in behavior, we propose a social-historical model (SHM) to integrate both social ties and historical ties to investigate the relationship between the social ties and the user's check-in behavior. Location prediction is used as an application to evaluate the social-historical ties effect on LBSNs. The contributions of our work are summarized below:

- We introduce the HPY language model for modeling the user's historical check-in sequences of LBSNs since HPY naturally captures the power-law distribution and short-term effect of check-in behavior.
- We propose a social-historical model (SHM) that enables us to study the importance of social-historical ties in affecting user's check-in behavior.
- We design experiments to evaluate how social and historical ties affect a user's check-in behavior. For example, which tie will play a determining role and how both ties work under what circumstances.

The remainder of this paper is organized as follows. We first introduce a language model based on analyzing the user's historical ties on check-in behavior, next present the proposed models for historical ties and social-historical ties, then discuss experimental design and results on the real-world dataset, followed by a brief review of some related work, and last, conclude this work with future work.

Analyzing User's Historical Ties

On an LBSN, a user's individual check-in behavior exhibits power-law distribution and short-term effect as described above. To capture these two properties, we introduce a language model to help in analyzing check-in behavior. There are many common features shared between language processing and LBSNs mining. First, the text data and check-in data have similar structures, as shown in Table 1. For example, a document in language processing can correspond

Table 1: Correspondences between language and LBSN modeling

Language Modeling		LBSN Modeling	
Corpus		Check-in collection	
Document		Individual check-ins	
Document Structure	Paragraph	Check-in Structure	Monthly check-in sequence
	Sentence		Weekly check-in sequence
	Phrase		Daily check-in sequence
	Word		Check-in location

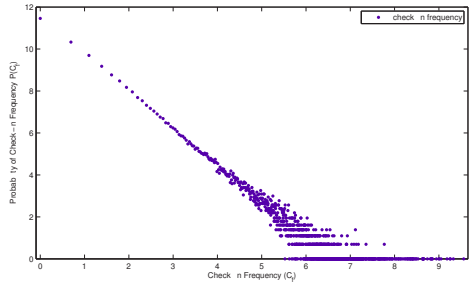
to a individual check-in sequence in LBSNs, while a word in the sentence corresponds to a check-in location. Second, the power-law distribution and short term effect observed in LBSNs have also been found in natural language processing, where the word distribution is closely approximated by power-law (Zipf 1949); and the current word is more relevant to its adjacent words than distant ones. Thus, the language model that works for language processing is potential applicable to LBSNs given these common features. For example, the unigram language model that ignores the relationship between a word to its nearest neighbors can be applied to LBSNs while considering the current check-in and ignoring its latest check-ins, and so does the n-gram language model. Therefore, to model the historical ties of a user, specifically, the power-law distribution and short-term effect of historical ties, we introduce Pitman-Yor process to the location based social networks.

Pitman-Yor process (Pitman and Yor 1997; Pitman 2006; Ishwaran and James 2001) is a state-of-the-art language model that generates a power-law distribution of word tokens (Goldwater, Griffiths, and Johnson 2006). Furthermore, its hierarchical extension, i.e., Hierarchical Pitman-Yor (HPY) process (Teh 2006a; 2006b), assumes that the earliest word has least importance to the latest word, which has potential to be leveraged to capture the short-term effect in LBSNs. Therefore, we propose to utilize the power of language model in LBSNs for modeling check-in behavior. We first demonstrate the power-law distribution in check-in behavior. Figure 2(a) shows the distribution of check-in frequency on our collected data (more details in the experiment section). Note that both x-axis and y-axis are in the log scale. The figure suggests that the check-in history follows a power-law distribution and the corresponding exponent is approximately 1.42. The check-in distribution of an individual also shows power-law property, as shown in Figure 2(b). We now introduce how the PY process captures the power-law property. The PY process generates a distribution over distributions over a probability space. Given a user with his/her check-in history, the next check-in location distribution is formulated as:

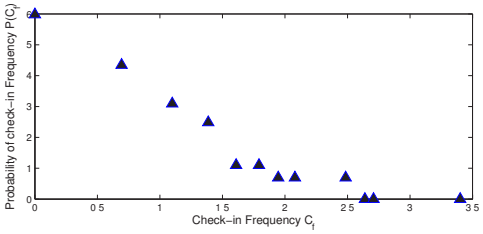
$$G \sim PY(d, \gamma, G_0), \tag{1}$$

where G is the next check-in location distribution based on the observed check-in history, $d \in [0, 1)$ is a discount parameter to control the power-law property, γ is a strength parameter, and G_0 is a base distribution over the location space. Let $\mathcal{L}$ be the location space which is a fixed and finite vocabulary of m locations, i.e., $m = |\mathcal{L}|$. The base distribution G_0 is a uniform distribution providing a prior probability of the location before observing any data. It satisfies

$G_0(l) = 1/m$, where $G_0(l)$ is the probability of location $l \in \mathcal{L}$ being checked-in. Furthermore, when the discount parameter d is regarded as zero, this process reduces to the Dirichlet process (Ferguson 1973).



(a) Power-law distribution of check-ins in whole dataset



(b) Power-law distribution of individual check-ins

Figure 2: The power-law distribution of check-ins

Next, we illustrate how to generate a check-in sequence with this process. Let $c_1, c_2, ..., c_n$ be a sequence of check-ins coming one by one. The first arrived check-in chooses a location drawn from the distribution G_0, then uses this location to form a location node and adheres to it. The subsequent check-in could either choose to adhere to a previous location node as its check-in location, or choose a new location node with its check-in location drawn from G_0. The choosing rule is: the k-th location node with probability $\frac{N_k - d}{\gamma + n}$ while a new location node with probability $\frac{\gamma + td}{\gamma + n}$, where N_k denotes the number of check-ins adhered to location node k, $n = \sum_k N_k$ the length of check-in sequence, and t the current number of location nodes. Notice that each location node represents a check-in location. Since a new draw from G_0 may generate a previously appeared location, there may be multiple location nodes corresponding to one check-in location. Therefore, by marginalizing on the location node, the predictive probability of a new check-in c_{n+1} at location l given the previous check-in sequence is,

$$P(c_{n+1} = l | c_1, c_2, ..., c_n) = \sum_k \frac{N_k - d}{\gamma + n} \delta_k^l + \frac{\gamma + td}{\gamma + n} G_0 = \frac{N_l - t_l d}{\gamma + n} + \frac{\gamma + td}{\gamma + n} G_0,$$

where δ_k^l is a function that satisfies:

$$\delta_k^l = \begin{cases} 1 & \text{location node k represents location } l \\ 0 & \text{location node k does not represent location } l, \end{cases}$$

$N_l = \sum_k N_k \delta_k^l$ denotes the current number of check-ins adhered to the location node at location l, which is the current number of check-ins at location l, and t_l is the current number of location nodes that represent location l. This generating process indicates that a new check-in would either choose a previously appeared location l with probability proportional to $N_l - t_l d$, or choose a location drawn from G_0 with probability proportional to $\gamma + td$.

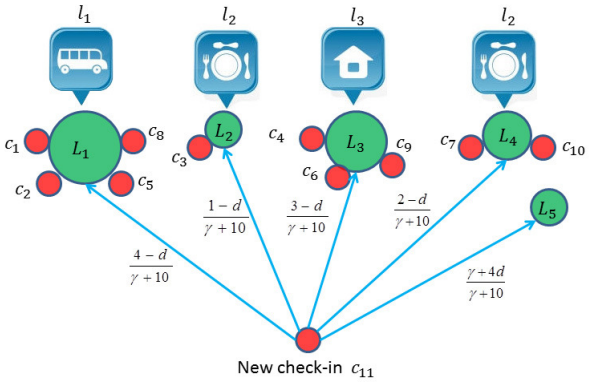


Figure 3: The generating process of check-in sequence

Figure 3 illustrates a generating process of the next check-in c_{11} with 10 previous check-ins $\{c_1 = l_1, c_2 = l_1, c_3 = l_2, c_4 = l_3, c_5 = l_1, c_6 = l_3, c_7 = l_4, c_8 = l_1, c_9 = l_3, c_{10} = l_4\}$. The green nodes are location nodes and each one represents a location corresponding to a location icon. Red nodes are check-ins that adhered to the location nodes, which indicates the check-ins happened at that location. The probability of next check-in c_{11} at location l_2 consists of three parts: (1) c_{11} adheres to the location node L_2 with probability $\frac{1-d}{\gamma+10}$; (2) c_{11} adheres to the location node L_4 with probability $\frac{2-d}{\gamma+10}$; and (3) c_{11} forms a new location node representing the check-in location l_2 with probability $\frac{\gamma+4d}{\gamma+10} G_0(l_2)$. Therefore, the probability of the next check-in $c_{11} = l_2$ is:

$$P(c_{11} = l_2 | c_1, ..., c_{10}) = \frac{3 - 2d}{\gamma + 10} + \frac{\gamma + 4d}{\gamma + 10} G_0(l_2), \quad (2)$$

This generating process shows two properties. First, the rich-get-richer property indicates that a user has a tendency to visit some places more frequently than others. Second, the more check-ins occurred, the more new locations would appear as drawn from the base distribution G_0.

The PY process models the power-law property and generates the unigram check-in distribution for a check-in sequence. However, a unigram distribution is not sufficient to capture the short-term effect, therefore we adopt the hierarchical extension of PY process, i.e., Hierarchical Pitman Yor process (Teh 2006a; 2006b) to consider the historical context of a particular check-in. It is an n-gram model that naturally captures the short-term effect while keeping the power-law property in distribution. It models the probability of the next check-in, denoted as G_u, given a history context u as:

$$G_u \sim PY(d_{|u|}, \gamma_{|u|}, G_{\pi(u)}), \quad (3)$$

where $G_u(l), l \in \mathcal{L}$, is the probability of the next check-in occurring at location l given the history context u. The discount parameter $d_{|u|}$ and strength parameter $\gamma_{|u|}$ are func-

tions of the historical context u. $\pi_{(u)}$ is the suffix of u consisting of all but the earliest check-in, therefore $G_{\pi(u)}$ is the probability of next check-in given all but the earliest check-in in the history context u. $G_{\pi(u)}$ is then computed with the parameter $d_{|\pi(u)|}$,$\gamma_{|\pi(u)|}$ and $G_{\pi(\pi(u))}$. This process is repeated until we get the empty historical context $\emptyset$,

$$G_{\emptyset} \sim PY(d_0, \gamma_0, G_0). \tag{4}$$

Note that this iterative process drops the earliest check-in first in each iteration. It assumes that the earliest check-in has the least importance in determining the distribution over the next check-ins, which in turn captures the short-term effect progressively.

Modeling User's Check-in Behavior

In this section, we propose a historical model to capture the user's check-in behavior in terms of historical ties and then a social-historical model to integrate both social and historical ties modeling user's check-in behavior.

Historical Model (HM)

Based on Eq. (3), the predictive probability of the next check-in c_{n+1} at location l with context u is defined as:

$$\begin{aligned} &P_u^{HPY}(c_{n+1} = l | c_1, c_2, ..., c_n) \\ &= \frac{N_{ul} - t_{ul} d_{|u|}}{\gamma_{|u|} + n_{u\cdot}} + \frac{\gamma_{|u|} + t_u d_{|u|}}{\gamma_{|u|} + n_{u\cdot}} G_{\pi(u)}(c_{n+1} = l | c_1, c_2, ..., c_n), \end{aligned} \tag{5}$$

where N_{ul} is the number of check-ins at l following the history context u and $n_{u\cdot} = \sum_{l'} N_{ul'}$. $t_u = \sum_l t_{ul}$ is the sum of all t_{ul}, which is a latent variable satisfying:

$$\begin{cases} t_{ul} = 0 & \text{if } N_{ul} = 0; \\ 0 \leq t_{ul} \leq N_{ul} & \text{if } N_{ul} > 0; \end{cases}$$

Since we always consider a user's complete check-in history as historical context u, we remove the notion u in the following sections. To model the historical tie effect, we define our *historical model* (HM) as:

$$P_H^i(c_{n+1} = l) = P_{HPY}^{i,i}(c_{n+1} = l). \tag{6}$$

where $P_{HPY}^{i,i}(c_{n+1} = l)$ is the probability of u_i's check-in c_{n+1} at location l generated by HPY process with user u_i's observed check-in history.

Social-Historical Model (SHM)

As a user's movement may also be influenced by his social ties, we explore the social tie effect by proposing a social-historical model to understand the user's check-in behavior on LBSNs. First, we investigate the social correlation of check-in behavior, more specifically, we ask whether the friendships a user has affect his check-in behavior. We first compare the number of common check-ins between two friends and two strangers. As shown in Table 2, on average, a pair of strangers share approximately *4.32* check-ins, while a pair of friends share approximately *11.83* check-ins, which is as almost *3* times large as the former.

Table 2: Average number of check-ins between two users

	Common check-ins
between friends	11.8306
between strangers	4.3226

Next, we define the check-in similarity between two users and compare the similarity between users with friendship and those without. For each user, let $\mathbf{f} \in \mathbb{R}^m$ be his check-in vector with each element $\mathbf{f}(k)$ equal to the number of check-ins at location $l_k \in \mathcal{L}$, where $m = |\mathcal{L}|$ is the vocabulary size. The cosine similarity of two users u_i and u_j is defined as:

$$sim(u_i, u_j) = \frac{\mathbf{f}_i \cdot \mathbf{f}_j}{|\mathbf{f}_i|_2 \times |\mathbf{f}_j|_2}, \tag{7}$$

where $|\bullet|_2$ is the 2-norm of a vector.

We define the similarity between u_i and a group G of other users as the average similarity between user u_j and the users in group G,

$$S_G(u_i) = \frac{\sum_{u_j \in G} sim(u_i, u_j)}{|G|}. \tag{8}$$

For each u_i, we calculate two similarities, i.e., $\mathbf{S}_F(u_i)$ is the average similarity of u_i and his friendship network; $\mathbf{S}_R(u_i)$ is the average similarity of u_i and randomly chosen users, who are not in the friendship network of u_i. The number of the randomly chosen users is the same as that of u_i's friends.

We conduct a two-sample t-test on the vectors $\mathbf{S}_F$ and $\mathbf{S}_R$. The null hypothesis is H_0: $\mathbf{S}_F \leq \mathbf{S}_R$, i.e., users with friendship share less common check-ins than those without, and the alternative hypothesis is H_1: $\mathbf{S}_F > \mathbf{S}_R$. In our experiment, the null hypothesis is rejected at significant level $\alpha = 0.001$ with p-value of $2.6e$-6, i.e., users with friendship have higher check-in similarity than those without.

The evidence from both shared check-ins and t-test suggests that with high probability, users with friend relationships have larger check-in correlation than those without, which demonstrates that a user's social ties contain important evidence for the user's movement. In this paper, we propose an effective model to integrate both effects, in order to explore the social-historical ties. To do so, we add a user's social ties as a regularization part to his historical ties. A parameter $\eta \in [0, 1]$ is introduced to control the weight between historical and social ties. For a user u_i, the probability of the next check-in location is defined as:

$$P_{SH}^i(c_{n+1} = l) = \eta P_H^i(c_{n+1} = l) + (1 - \eta) P_S^i(c_{n+1} = l). \tag{9}$$

We denote this model as *social-historical model* (SHM), where $P_H^i(c_{n+1} = l)$ is the probability of u_i's check-in at location l observed from his historical ties, defined in Eq. (6); $P_S^i(c_{n+1} = l)$ is the check-in probability calculated based on u_i's social ties, defined as:

$$P_S^i(c_{n+1} = l) = \sum_{u_j \in \mathcal{N}(u_i)} sim(u_i, u_j) P_{HPY}^{i,j}(c_{n+1} = l). \tag{10}$$

where $\mathcal{N}(u_i)$ is the set of u_i's friends. $P_{HPY}^{i,j}(c_{n+1} = l)$ is the probability of u_i's next check-in at location l computed

Table 3: Statistical information of the dataset

number of users	18,107
number of check-ins	2,073,740
number of links	123,325
average check-ins per user	101
clustering coefficient	0.1841
average degree	10.58

by HPY process with u_j's check-in history as training data. Note that only the check-ins before the prediction time are included in the training data.

Experiments

In this work, we use location prediction as an application to evaluate our proposed models: historical model and social-historical model. In particular, we evaluate the following: (1) how the proposed historical model fares in comparison with baseline models; (2) how the proposed historical model behaves over time; (3) whether social ties help location prediction as we discussed earlier; and (4) under what circumstances, the two types of ties complement each other. Before we delve into experiment details, we first discuss an LBSN dataset, evaluation metrics, and baseline models.

Data

We choose Foursquare, one of the most popular LBSNs, to study the social-historical ties on LBSNs. Foursquare has more than 15 million members as of June, 2011[4] and keeps growing every month. For a particular user on Foursquare, we get his check-in history with timestamps and his friendship information. Since Foursquare does not provide APIs to collect personal check-ins, we are not able to get the check-in history directly from Foursquare. However, members on Foursquare can choose to list on their Twitter account and automatically publish their check-in messages as tweets on Twitter. We can access these tweets through Twitter's public REST API. A check-in tweet contains a unique URL that points to a Foursquare web page including the geographical information of this check-in location. We get check-ins with timestamps ranging from August, 2010 to November, 2011. Instead of crawling the friendships on Twitter as done in (Scellato et al. 2011), we collect the user's social ties directly from Foursquare to keep the friendships identical to the Foursquare social circle.

In our experiment, we consider the users who have at least 10 check-ins. We obtain 43,108 unique geographical locations as the location vocabulary. Some key statistics of the dataset are shown in Table 3.

Evaluation Metrics

We separate the check-in sequence of each user into 9 time bins, and each time bin has approximately equal time interval. Let the timestamp at the end of each time bin be $\mathcal{T} = \{T_1, T_2, ..., T_9\}$. We predict the check-in location at each timestamp for the user, with his historical check-ins before that time as observed context. Denote the prediction for user u at time t as $P_t(u)$, the prediction accuracy is defined as:

$$accuracy(T_i) = \frac{|\{u|u \in \mathcal{U}, P_{T_i}(u) = l_{T_i}(u)\}|}{|\mathcal{U}|}. \quad (11)$$

where $\mathcal{U}$ is the set of users, $l_{T_i}(u)$ denotes the actual check-in location l of user u at time T_i.

Baseline Models

To evaluate the historical model (HM) and social-historical model (SHM), we choose three baseline models, i.e., Most Frequent Check-in model (MFC), Most Frequent Time model (MFT), and Order-k Markov Model based on our review of related work (to discuss later). The MFC baseline model considers the power-law property simply in aspect of rich-get-richer effect. The MFT model considers the temporal pattern only, which was used in (Cho, Myers, and Leskovec 2011) for comparison with their periodic model. Since our proposed models do not attempt to model periodic behavior, we focus on the social and historical sequence of check-ins. Integrating periodic patterns in HM and SHM will be an extension of this work. The Order-k Markov Model considers the short-term effect of historical check-ins, which is reported as a state-of-the-art prediction algorithm for location prediction (Song et al. 2004). We give detailed information of these three baselines below:

- **Most Frequent Check-in Model**: In (Chang and Sun 2011), a logistic regression model was proposed and found that the strongest predictor is the check-in frequency of the historical check-ins made by the user. In this paper, we use this rule as one baseline, denoted as the most frequent check-in model (MFC). It assigns the probability of next check-in c_{n+1} at location l as the probability of l appearing in the check-in history,

$$P_{MFC}(c_{n+1} = l|\mathcal{C}) = \frac{|\{c_r|c_r \epsilon \mathcal{C}, c_r = l\}|}{|\{c_r|c_r \epsilon \mathcal{C}\}|}, \quad (12)$$

  where $\mathcal{C} = \{c_1, c_2, ..., c_n\}$ is the set of check-in history.

- **Most Frequent Time Model**: People tend to go the the same place at the similar time of the day as a routine activity. For example, an individual might like to have coffee after lunch; therefore, it would be common for him to check-in at Starbucks around 1pm. We choose the most frequent time model (MFT) as another baseline considering the temporal patterns of the check-ins. Let $t_{n+1} = h$ denote that the time at the $(n+1)$-th check-in is h, where $h \in \mathcal{H} = \{1, 2, ..., 24\}$ is a discrete set of 24 hours. MFT model assigns the probability of next check-in c_{n+1} at location l at time h as the probability of the location l occurring at time h in the previous check-in history,

$$P_{MFT}(c_{n+1} = l|\mathcal{C}, t_{n+1} = h) = \frac{|c|c_r \epsilon \mathcal{C}, c_r = l, t_r = h|}{|c_r|c_r \epsilon \mathcal{C}, t_r = h|}. \quad (13)$$

- **Order-k Markov Model**: The third baseline is the order-k Markov Model. It considers the latest k check-in context, and searches for frequent patterns to predict the next

[4] http://mashable.com/2011/12/05/fourquare-15-million-users

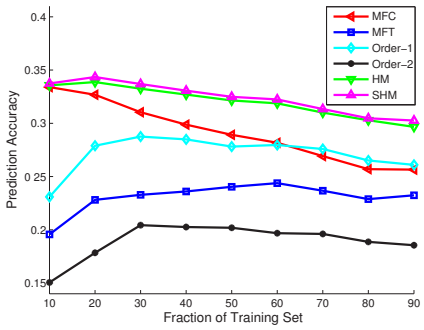


Figure 4: The performance comparison of prediction models

location. The probability of the next check-in c_{n+1} at location l with order-k Markov model is defined as:

$$P_{order-k}(c_{n+1} = l|\mathcal{C}) = P(c_{n+1} = l|c_{n-k+1}, ..., c_n)$$
$$= \frac{|c_r|c_r \epsilon \mathcal{C}, c_r = l, c_{r-j} = c_{n-j+1}|}{|c_r|c_r \epsilon \mathcal{C}, c_{r-j} = c_{n-j+1}|}, 0 < j < k, j \in Z. \quad (14)$$

We consider the Order-1 and Order-2 Markov models as baseline methods, note that the MFC is actually Order-0 Markov model.

Results and Discussions

Figure 4 shows the comparison results at 9 time stamps. The Order-2 Markov model performs the worst while SHM obtains the best performance for all the 9 time stamps. MFC model performs well but its accuracy decreases greatly as time goes by. The Order-1 Markov model has better performance than MFC after T6, while the MFT model performs stable without impressive accuracy. To further investigate their performance, we summarize several interesting observations below:

- The MFC performs better than MFT, Order-1, and Order-2 Markov models before T6. Since it predicts the next location as the most frequent one in history, it considers the rich-get-richer property of power-law effect. However, it ignores the short-term effect; therefore, as time goes by, it can not distinguish which check-ins are more important in the long history and its accuracy decreases quickly. In contrast, the Order-1 Markov model outperforms MFC after T6. That is because Order-1 Markov model considers the short-term effect more than power-law property, it is not affected by the length of the history as much as MFC.
- All the models have a decreasing trend in accuracy after certain time point, especially the SHM, HM, Order-1, and Order-2 Markov models have similar decreasing rate with time. This phenomenon can be explained by the increasing number of appeared unique check-ins. We prove this by launching a random guess of next check-in location at each time point. We denote the average number of unique check-ins per user that appeared before time t as W_t. The probability of accurately predict the next location at this time by random guess is the inverse of W_t, which reflects the difficulty of prediction. We denote this as the random guess accuracy A_{W_t}. The statistics information of W_t and its corresponding A_{W_t} from our data is shown in Table 4. From T1 to T9, the accuracy of random guess keeps decreasing from 20.49% to 4.33% (approximately 78.87% relative decrement), while our historical model only decreases 3.9% from 33.56% to 29.66% (approximately 11.62% relative decrement). In sum, the performance of our historical model is considerable, and even slight improvement in experiment is significant considering the difficulty of this prediction task.
- The MFC has the highest decreasing rate among all the models. This phenomenon is caused by two factors. Firstly, MFC is affected by the number of appeared unique check-ins as well as SHM, HM, Order-1, and Order-2 Markov models as described above. Secondly, it suffers from the short-term effect. Since even the number of appeared unique check-ins does not increase, it cannot distinguish the most important check-ins to current time through the long history. Therefore, suffering from both unique check-ins and short-term effect, it has the greatest decreasing rate among all the five models.
- In our data, there are 14.47% of users with check-in sequence length between 10 to 20. For these users at time T1, only 1 to 2 check-ins are observed, which significantly intensifies the prediction difficulty. Specially, SHM, HM and MFC are very close to each other at T1, because all are suffering from the lack of observed data. The MFT, Order-1, and Order-2 Markov models perform even worse than SHM, HM, and MFC due to their strict pattern rules. With insufficient data, few patterns can be found and used to determine the next location by these three models; while as time goes by, more and more data are observed which improves their performance. This suggests that SHM, HM and MFC are more robust to the situation when the observation sequence is insufficient. The Order-2 Markov model is too strict on its pattern rule therefore it performs the worst due to over-fitting.
- The HM obtains better performance than all baselines, which considers both power-law property and short-term effect. Furthermore, the smoothing strategy on the n-gram context gives it better performance than the Order-2 Markov model, which suffers severely from over-fitting. The MFT performs stable, which suggests the importance of temporal information. We will further investigate the temporal effect on check-in behavior in our future work.

We note that SHM consistently outperforms HM, and SHM considers both historical and social ties. To investigate the contribution of social ties and historical ties in affecting user's behavior, we increase the parameter η from 0 to 1 with an increment step of 0.01 and observe the prediction performance at each η. We only show the prediction accuracy at times T3, T6 and T9 in Figure 5, since similar performance can be observed at other time points. Some interesting insights can be observed:

- When $\eta = 0$, the social-historical model only considers social ties. Its performance is always worst, suggesting that considering social information only is not enough to

capture the check-in behavior.

- By increasing η, the performance shows the following pattern: first increasing, reaching its peak value and then decreasing. Most of the time, the best performance is achieved at around $\eta = 0.7$. A big weight is given to historical ties, indicating that historical ties are more important than social ties.
- When $\eta = 1$, the social-historical model boils down to the historical model. Its performance is not the best, suggesting that social ties are also important.
- Comparing with the previous time, the social ties make the greatest improvement on performance of historical ties at T9, indicating that social ties are complementary to the historical ties, especially when the historical model does not perform well due to the long and noisy history.

Related Work

Previous work mainly focuses on studying the effect of historical ties and social ties to user's movement independently. In (Zheng et al. 2009), the authors modeled multiple individuals' location histories to mine the interesting locations and travel sequences with GPS logs. Petzold et al. (Petzold et al. 2005) investigated user's historical ties for in-door next location prediction within an office building. In (Zheng et al. 2008), the authors proposed an supervised learning approach to infer people's motion modes from their GPS histories. Yavas et al. (Yavas et al. 2005) proposed an algorithm for mobile prediction with communication histories by mobile rules. In (Goldenberg and Levy 2009; Mok, Wellman, and Carrasco 2010; Cairncross 2001), the authors studied the relationship between the social ties of two users and their real geographical distance. Gong et al. (Gong et al. 2011) introduced social networks into location prediction and predicts a user's next location as his closest friend's recent location without considering the user's own location history. Their proposed model has similar changing tendency to the order-2 Markov model.

Researchers have attempted to understand the check-in property on location based social networks. Cheng et al. (Cheng et al. 2011) reported a quantitative assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects associated with check-ins. They found that the check-in historical ties follow the "Lèvy Flight" and the social status is affected by geographic constraints. In (Scellato et al. 2011; Scellato and Mascolo 2011; Scellato et al. 2010), the authors studied the spatial properties of the social networks on three main popular LBSNs. They observed strong heterogeneity across users with different characteristic geographic scales of interaction across social ties.

Efforts have also been made to integrate user's social and historical information of check-ins for location prediction. Chang et al. (Chang and Sun 2011) proposed a logistic regression model and found that the strongest predictor is the check-ins frequency of the historical check-ins made by user, while the check-in frequency in the user's friends' check-in history is also a good predictor. Cho et al. (Cho, Myers, and Leskovec 2011) proposed a Periodic & Social Mobility Model which considers the user's movement as a 2-dimensional time-independent Gaussian distribution. They consider the temporal pattern other than the historical check-in sequence.

Conclusion

In this paper, we explore the pattern of user check-ins on LBSNs with respect to social-historical ties. We find that users with friendship tend to go to similar locations than those without. We observe the power-law property and short-term effect in historical ties and introduce a historical model to capture these properties. We devise an approach for integrating social-historical ties to model user's check-in behavior.

The experimental results on location prediction demonstrate that our proposed approach suitably captures user's check-in property and outperforms current mobile models.

In our current work, we do not consider the temporal information for check-in modeling, and the social tie strengths are computed using cosine similarity in terms of "bag of check-in". Furthermore, Tang et al. (Tang, Gao, and Liu 2012) investigate the power of user preference on user behavior prediction, which can also be considered as a potential improvement. An interesting direction for future work is to consider all the social-spatial-temporal information with multi-faceted user preference and the social tie strengths computed in sequence constrains, which may lead to a better understanding of the social geographical check-in behavior on LBSNs.

Acknowledgments

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

References

Anagnostopoulos, A.; Kumar, R.; and Mahdian, M. 2008. Influence and correlation in social networks. In *KDD*, 7–15. ACM.

Barwise, P., and Strong, C. 2002. Permission-based mobile advertising. *Journal of interactive Marketing* 16(1):14–24.

Cairncross, F. 2001. *The death of distance: How the communications revolution is changing our lives*. Harvard Business Press.

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

Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D. 2011. Exploring millions of footprints in location sharing services. In *ICWSM*.

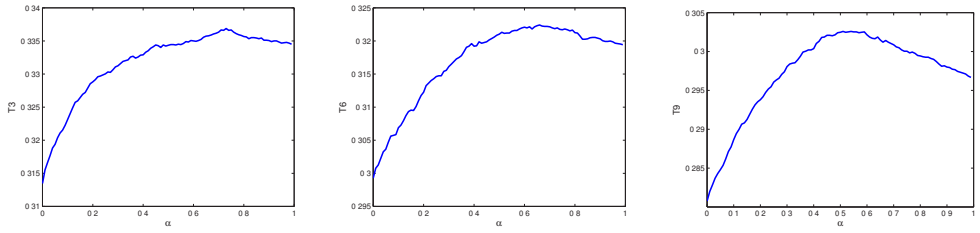
Cho, E.; Myers, S.; and Leskovec, J. 2011. Friendship and mobility: user movement in location-based social networks. In *KDD*, 1082–1090. ACM.

Ferguson, T. 1973. A bayesian analysis of some nonparametric problems. *The annals of statistics* 209–230.

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

Table 4: Number of unique check-ins at each time point

	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9
No. of Unique Check-ins W_t	4.88	7.578	9.95	12.20	14.40	16.59	18.75	20.92	23.11
Random Guess Accuracy A_{W_t}	20.49%	13.20%	10.05%	8.20%	6.94%	6.03%	5.33%	4.78%	4.33%



(a) T3 (b) T6 (c) T9

Figure 5: The performance of social-historical model w.r.t. η

Goldenberg, J., and Levy, M. 2009. Distance is not dead: Social interaction and geographical distance in the internet era. *Arxiv preprint arXiv:0906.3202*.

Goldwater, S.; Griffiths, T.; and Johnson, M. 2006. Interpolating between types and tokens by estimating power-law generators. *Advances in neural information processing systems* 18:459.

Gong, Y.; Li, Y.; Jin, D.; Su, L.; and Zeng, L. 2011. A location prediction scheme based on social correlation. In *VTC Spring*, 1–5. IEEE.

Gordon, M. 1973. *The American family in social-historical perspective*. St. Martin's Press.

Hargreaves, J. 1986. *Sport, power and culture: A social and historical analysis of popular sports in Britain*. St. Martin's Press New York.

Hodgson, G. 2001. *How economics forgot history: The problem of historical specificity in social science*. Psychology Press.

Ishwaran, H., and James, L. 2001. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 96(453):161–173.

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

Petzold, J.; Bagci, F.; Trumler, W.; and Ungerer, T. 2005. Next location prediction within a smart office building. *Cognitive Science Research Paper-University of Sussex CSRP* 577:69.

Pitman, J., and Yor, M. 1997. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability* 25(2):855–900.

Pitman, J. 2006. *Combinatorial stochastic processes*, volume 1875. Springer-Verlag.

Quarantelli, E. 1987. Disaster studies: An analysis of the social historical factors affecting the development of research in the area.

Scellato, S., and Mascolo, C. 2011. Measuring user activity on an online location-based social network. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, 918–923. IEEE.

Scellato, S.; Mascolo, C.; Musolesi, M.; and Latora, V. 2010. Distance matters: Geo-social metrics for online social networks. In *Proceedings of the 3rd conference on Online social networks*, 8–8. USENIX Association.

Scellato, S.; Noulas, A.; Lambiotte, R.; and Mascolo, C. 2011. Socio-spatial properties of online location-based social networks. *ICWSM* 11.

Song, L.; Kotz, D.; Jain, R.; and He, X. 2004. Evaluating location predictors with extensive wi-fi mobility data. In *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, volume 2, 1414–1424. IEEE.

Tang, J.; Gao, H.; and Liu, H. 2012. mtrust: discerning multi-faceted trust in a connected world. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 93–102. ACM.

Teh, Y. 2006a. A bayesian interpretation of interpolated kneser-ney.

Teh, Y. 2006b. A hierarchical bayesian language model based on pitman-yor processes. In *ACL*, 985–992. Association for Computational Linguistics.

Yavas, G.; Katsaros, D.; Ulusoy, O.; and Manolopoulos, Y. 2005. A data mining approach for location prediction in mobile environments. *Data & Knowledge Engineering* 54(2):121–146.

Zheng, Y.; Li, Q.; Chen, Y.; Xie, X.; and Ma, W. 2008. Understanding mobility based on gps data. In *UbiComp*, 312–321. ACM.

Zheng, Y.; Zhang, L.; Xie, X.; and Ma, W. 2009. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, 791–800. ACM.

Zickuhr, K., and Smith, A. 2010. 4% of online americans use location-based services. *Pew Internet & American Life Project*.

Zipf, G. 1949. Selective studies and the principle of relative frequency in language (cambridge, mass, 1932). *Human Behavior and the Principle of Least-Effort (Cambridge, Mass, 1949*.