Modeling User Mobility for Location Promotion in Location-based Social Networks

Wen-Yuan Zhu[1], Wen-Chih Peng[1], Ling-Jyh Chen[2], Kai Zheng[3] and Xiaofang Zhou[3]
[1]National Chiao Tung University, Taiwan
[2]Academia Sinica, Taiwan
[3]The University of Queensland, Australia
{wyzhu, wcpeng}@cs.nctu.edu.tw, ccljj@iis.sinica.edu.tw, {kevinz, zxf}@itee.uq.edu.au

ABSTRACT

With the explosion of smartphones and social network services, location-based social networks (LBSNs) are increasingly seen as tools for businesses (e.g., restaurants, hotels) to promote their products and services. In this paper, we investigate the key techniques that can help businesses promote their locations by advertising wisely through the underlying LBSNs. In order to maximize the benefit of location promotion, we formalize it as an influence maximization problem in an LBSN, i.e., given a target location and an LBSN, which a set of k users (called seeds) should be advertised initially such that they can successfully propagate and attract most other users to visit the target location. Existing studies have proposed different ways to calculate the information propagation probability, that is how likely a user may influence another, in the settings of static social network. However, it is more challenging to derive the propagation probability in an LBSN since it is heavily affected by the target location and the user mobility, both of which are dynamic and query dependent. This paper proposes two user mobility models, namely Gaussian-based and distance-based mobility models, to capture the check-in behavior of individual LBSN user, based on which location-aware propagation probabilities can be derived respectively. Extensive experiments based on two real LBSN datasets have demonstrated the superior effectiveness of our proposals than existing static models of propagation probabilities to truly reflect the information propagation in LBSNs.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining, spatial databases and GIS*

General Terms

Algorithms, Experimentation.

KDD'15, August 11-14, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2783331

Keywords

Influence maximization; propagation probability; check-in behavior; location-based social network

1. INTRODUCTION

Due to the success of viral marketing, more and more advertisements appear in social networks. The key to the success of viral marketing is the influence among social connections. Users are more likely to accept the advertisements from their friends in social networks than from media directly. By observing this phenomenon, prior works have elaborated on the influence maximization problem in social networks. In general, a social network is modeled as a graph $G = (U, E)$, where U is the set of users, E refers to the social connections among users, and the weight of edges infers the influence degree between users. Given a graph, the influence maximization problem is to select a set of users as seeds with the purpose of maximizing the number of influenced users (i.e., influence spreads) in a social network.

With the popularity of smart phones and location-based social networks (LBSNs), users are able to check-in at some locations and share their check-in records with their friends. In view of the social influences of friends, recently, many POIs (e.g., restaurants, stores) have explored check-in sharing to attract users to stay or visit. We mention in passing that the prior work in [16] formulated a location-aware influence maximization problem in which, given a query region and the location of users, a set of seed users is determined with the purpose of maximizing influence spreads in the query range. In reality, locations in social media are referred to as POIs which could be restaurants, hotels or theme parks. From the perspective of POIs, each would like to attract users to visit, and via the check-in records of users, more users (e.g., friends of check-in users) will be influenced and then visit this POI. Thus, we formulate this problem as a location promotion problem in which given one target location and the number of seeds, the purpose is to maximize the number of influenced users. Note that the location promotion problem is different from the prior work [16] in that a target location is specified.

Intuitively, the location promotion problem can be modeled as an influence maximization problem in LBSNs. Explicitly, given a set of nodes and edges with the propagation probability, a target location and the number of seeds, a set of seed users is derived to maximize the number of influenced users. Notice that one challenging issue behind the location promotion problem is to determine the propagation probabilities of edges. By investigating the information of LBSNs

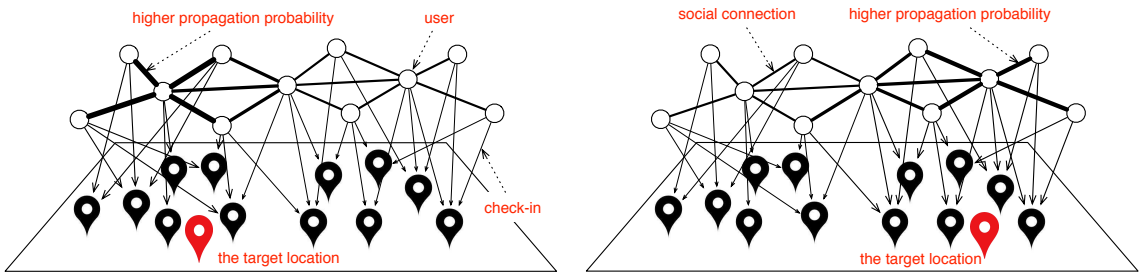


Figure 1: An example of dynamic propagation probability when the target location is changed in an LBSN

(i.e., check-in locations of users and social connections of users), we thus propose some static propagation probabilities. A success influence in LBSNs is that users should follow their friends to check-in at the target location specified. To achieve this, we claim that user mobility should be considered. If a user never appears in the nearby area of the target location, this user is not likely to visit the target location at all. Figure 1 shows an example of dynamic propagation probabilities in an LBSN. If the target location (the red location) is on the left side (e.g., US west in Figure 1(a)), users on the left side have more chance to visit it since the target location is near their check-in records. The propagation probabilities of edges among users in the left side are higher (bold social connections) than the propagation probabilities of edges among users on the right side. On the other hand, if the target location on the right side (e.g., US east in Figure 1(b)), the edges of users on the right side have larger propagation probabilities. Therefore, we consider a user mobility model for the propagation probability to truly reflect the propagation of information on social connections. If the target location is changed, the propagation probability should be updated.

From the above example, one should consider the mobility model for the location promotion problem. To the best of our knowledge, there are some mobility models to describe users' check-in behavior in LBSNs [5][6][17]. The mobility models proposed [5][6][17] are all based on the bivariate Gaussian distribution. However, it is hard to decide the area around the target location to evaluate the probability of a user visiting the target location from Gaussian-based mobility models, which are based on two-dimensional density functions. Moreover, the Gaussian-based mobility models only describe the coordinates of check-in records but do not consider their order. We argue that different orders of check-in records represent different check-in behavior. Therefore, we propose distance-based mobility models to represent individual check-in behavior in LBSNs. We exploit the distance between consequent check-in records to model individual check-in behavior, where the distances represent individual movement preferences. Moreover, it is easy to evaluate the probability of user check-in at the target location. Thus, our proposed distance-based mobility models are suitable not only to describe individual check-in behavior but also to determine the propagation probability of edges.

In summary, our major contributions are outlined as follows:

- We formulate the location promotion problem in an LBSN as an influence maximization problem in a graph.
- By investigating the check-in records of users and the social connections of users, we propose some approaches to derive static propagation probability in LBSNs.
- We take the target location and users' check-in records into consideration, and propose two types of mobility model, Gaussian-based mobility models and distance-based mobility models, to evaluate location-aware propagation probability in LBSNs.
- Gaussian-based mobility models and distance-based mobility models consider the spatial, temporal and social features hidden in LBSNs.
- We have conducted comprehensive experiments on two real datasets, and the experimental results show that the proposed models are suitable for effective influence maximization in LBSNs.

The remainder of this paper is organized as follows: Section 2 discusses related works. Section 3 gives a formal definition of the location promotion problem and the relation between location promotion and influence maximization in LBSNs. Section 4 presents how to evaluate the propagation probabilities in LBSNs. The experimental results are shown in Section 5. Section 6 concludes this paper.

2. RELATED WORKS

In this paper, our work is related to the influence maximization in social networks. Thus, we will present some existing works of influence maximization in social networks. Since the main theme of this paper is to derive the propagation probabilities, we will describe how to set the propagation probabilities in existing works. Note that as our work is to explore mobility models from check-in records to set the propagation probabilities, we present some research works of modeling user mobility from social media.

2.1 Influence Maximization in Social Networks

The influence maximization problem is to find k users as seeds which can maximize the influence spreads in a social network, and this problem is NP-hard [13]. In general, most works utilize greedy-based algorithms to select the user as a seed who could maximize the number of influenced users until k seeds are selected [13][14]. In [15], the authors propose the CELF algorithm which exploits the submodular property to significantly boost the traditional greedy approach. Moreover, another problem is to evaluate the influence spreads from a seed set. A traditional way is to utilize

the Monte Carlo approach [13][14]. However, the computation cost of this approach is expensive since it has to run about 10,000 times for evaluation [13]. In [4], the authors proposed a heuristic algorithm, MIA, to estimate the influence spreads by using the paths to represent the routes of influence among users. For LBSNs, the authors of [21] designed a greedy algorithm to select users with a higher degree and more closely related to the location as seeds in their one-wave diffusion model for LBSNs. However, the one-wave diffusion model is not close to real LBSNs. In [16], the authors target the influence maximization problem in LBSNs. Explicitly, given a query region and users' representative locations, the output is to derive k seeds with the purpose of maximizing influence spreads. This work is the first work for the influence maximization problem in LBSNs. However, this problem is different from the location promotion problem addressed in this paper. In location promotion, a specified target location is given. Though the location promotion problem could be solved via the influence maximization problem, we explore user mobility to infer propagation probabilities. These features differentiate our work from existing influence maximization works.

2.2 Propagation Probability in Social Networks

To detect the propagation probability in a social network, some existing approaches use the number of in-degrees [13], a fixed value [10] and a uniformly random value [10]. However, these approaches are some baseline methods. In social networks, users' actions indeed reflect real influences. As such, some works focus on learning the propagation probability from action history. In [20], the authors exploited the EM algorithm to learn the probability from the Independent Cascading Model (ICM). In [9], the authors utilized the Bernoulli distribution to model the influence on each social connection. The probability parameter of the Bernoulli distribution is learned from data for every social connection, and each probability parameter is the propagation probability of the corresponding edge. Moreover, prior works [1][3][18] noticed that different topics have different diffusion results since users have different topic preferences. In LBSNs, this phenomenon is more observable. Users have their locality such that they would not like to move to a far location, even if this location is recommended by their friends. Thus, different target locations have different propagation probabilities.

2.3 Capturing Movement Behavior

A trajectory is a sequence of locations ordered by visiting time so that many works aimed at discovering spatio-temporal patterns from trajectories [7][8][12]. To capture an individual user's check-in behavior, we have to deal with the spatio and temporal sparsity. In [19], the authors discovered spatio-temporal patterns from all users' check-in records in Foursquare. In [22], the authors showed that the major reason of check-in is based on their mobility, and the chance of social influence is very low. In [5], the authors had similar perspectives on weak social influence in LBSNs. The authors proposed PSMM which divides each user's check-ins into three types: work, home and social. The check-in records of each type are modeled by a bivariate Gaussian distribution. Moreover, to distinguish the home and work states of check-in records, they also select the Gaussian distribution to classify the time of check-in records. In [6], there are two factors considered for users' check-in behaviors, personal preference and social influence. Personal preference is modeled by a bivariate Gaussian distribution, and the social influence is from the similarity between the user's check-in records and their friends' check-in records. In [17], the authors exploit kernel density estimation (KDE) to model the individual user's check-in records. They provide how to set the parameters of KDE for each user from their check-in records. The KDE approach is Gaussian-based since it is a mixture model consisting of bivariate Gaussian distributions. In our work, it is hard for the models to evaluate the probability of a location since we have to decide on an area to compute the probability (the probability of a point in a two-dimensional distribution is 0). Therefore, we use the distance-based approach to model each user's check-in behavior.

3. PRELIMINARIES

3.1 Influence Maximization in LBSNs

In this section, we will give a formal definition of the location promotion problem in an LBSN. First, the formal definition of an LBSN is as follows:

Definition 1. (LBSN) An LBSN $\langle G, C\rangle$ consists of a social network $G = \langle U, E\rangle$, where U is the set of users, $E = \{(u_i, u_j)|$ a social connection from u_i to u_j, $u_i, u_j \in U, u_i \neq u_j\}$ and the set of check-in records $C = \{(u, \ell, t)\}$, where (u, ℓ, t) represents a check-in record where a user u checks in at location ℓ at time t, and $\ell \in L$. A location ℓ is a coordinate which consists of latitude and longitude.

Then, the formal definition of the location promotion problem on an LBSN is as follows:

Definition 2. (Location Promotion Problem) Given an LBSN $\langle G, C\rangle$, a target location ℓ and a constant k, the location promotion problem is to select a set of seeds S, $S \in U$, which has k seeds (to distinguish from other users) to maximize the number of expected influenced users $\sigma(S)$ who will visit the target location ℓ.

3.2 Framework

Figure 2 shows the framework of location promotion in an LBSN. In the first step, given a target location, we determine the propagation probability of edges in the social

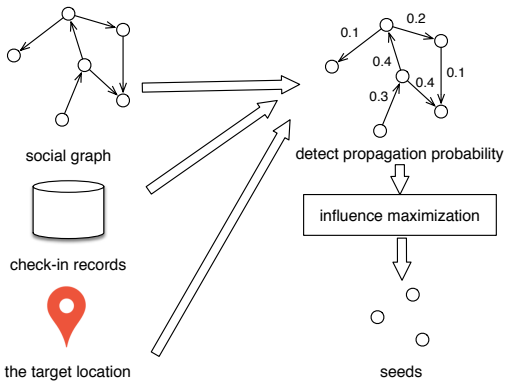


Figure 2: The framework for location promotion

graph based on check-in records. In this step, the input of location promotion in an LBSN is transformed to the input of influence maximization on a graph. Based on the social graph and propagation probability of edges from the first step, one could derive the seed set via existing solutions of influence maximization [4] and [15]. One challenging issue is how to set the propagation probability on social connections to truly reflect the propagation of information in an LBSN. The following sections will present how to derive propagation probability of influences among users in the social graph.

4. PROPAGATION PROBABILITIES IN LBSNS

In this section, we will propose some methods to derive propagation probability in an LBSN. First, we borrow the concept of prior works in [2][4][9][10][13] and then derive some baselines to determine the propagation probability. Note that in the location promotion problem, the given target location has an impact on the propagation probabilities. Thus, we explore the mobility models to derive the propagation probability in Section 4.2.

4.1 Static Propagation Probabilities

In this part, we will show the traditional approaches for determining the propagation probability on edges in social networks. By referring to prior works [2][4][9][10][13], we have seven approaches to derive the propagation probability $p_{u\to v}$ of edge (u, v) that is formulated as follows:

Uniform probability: All edges are assigned to the same probability [10].

$$p_{u\to v} = 0.01$$

Trivalency: All edges are assigned to the probability selected from $\{0.1, 0.01, 0.001\}$ uniformly [4].

In-degree of nodes: The propagation probability from u to v is the in-degree of v [13].

$$p_{u\to v} = \frac{1}{deg(v)}$$

where $deg(v)$ denotes the in-degree of v. The propagation probability is higher if the in-degree of v is lower.

Jaccard index of friends: The Jaccard index is used to measure the similarity between two sets. Here, we select the Jaccard index to measure the similarity between two users' friendships as the propagation probability if the edge (u, v) exists [9].

$$p_{u\to v} = \frac{|adj(u) \cap adj(v)|}{|adj(u) \cup adj(v)|}$$

where $adj(u)$ denotes the set of u's friends. The propagation probability is higher if most friends of u and v are the same.

Jaccard index of locations: For LBSNs, we selected the Jaccard index of locations from two users' check-ins as the propagation probability if the edge (u, v) exists [2].

$$p_{u\to v} = \frac{|loc(u) \cap loc(v)|}{|loc(u) \cup loc(v)|}$$

where $loc(u)$ denotes the set of u's visited locations. The propagation probability is higher if most visited locations of u and v are the same.

Cosine of locations: To consider the check-in times of locations, we convert the user's check-in records as a vector, in which the dimensions are locations and the values of each dimension are the check-in times of this location. Thus, we can measure the similarity via cosine from two users' check-in records.

$$p_{u\to v} = \frac{\sum_{\ell\in loc(u)\cap loc(v)} \#_u^L(\ell) \cdot \#_v^L(\ell)}{\sqrt{\sum_{\ell\in loc(u)} \#_u^L(\ell)^2}\sqrt{\sum_{\ell\in loc(v)} \#_v^L(\ell)^2}}$$

where $\#_u^L(\ell)$ denotes the number of times u visited ℓ. The propagation probability is higher if not only most visited locations of u and v are the same, but also the if check-in times of the visited locations are similar.

Bernoulli distribution: Given two check-in records of two users connected by a social connection, we can calculate the times of influence. For each check-in, we can know whether another user has checked-in at the same location after the check-in. If yes, then the influence is success; otherwise, it is failure. The probability of successful influence is from the Bernoulli distribution. To find the success probability in Bernoulli distribution from the records, the result from the maximum likelihood estimator is as follows [9].

$$p_{u\to v} = \frac{\#_u^I(v)}{\#\text{ of attempts from } u}$$

where $\#_u^I(v)$ denotes the times of successful influence from u to v. The propagation probability is higher if the success chance of influence attempt from u to v is higher.

4.2 Location-aware Propagation Probabilities

In LBSNs, the information propagation is different from in traditional social networks since users have to perform check-in at the target location to spread the information. Figure 3 shows an example of information propagation in an LBSN. u_1 is active such that u_1's friend u_2 will receive the information of the target location from u_1's check-in sharing. From the perspective of influence maximization, the propagation probability of the edge from u_1 to u_2 is the probability of activating u_2 from u_1 [13]. Thus, in LBSNs, this probability is the probability of u_2 checks-in at the target location. If u_2 does, the information of the target location will propagate to u_2's friend, u_3. In Figure 3, u_2 and u_3 have 70% and 60% to check-in at the target location, respectively. Thus, the propagation probabilities of the edge (u_1, u_2) and (u_2, u_3) are 0.7 and 0.6, respectively. Clearly, if the target location is changed, the probabilities will be changed. To infer whether each user will checks-in at the target location, we explore the mobility model in an

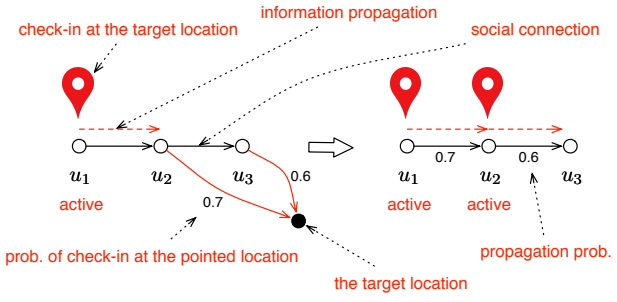
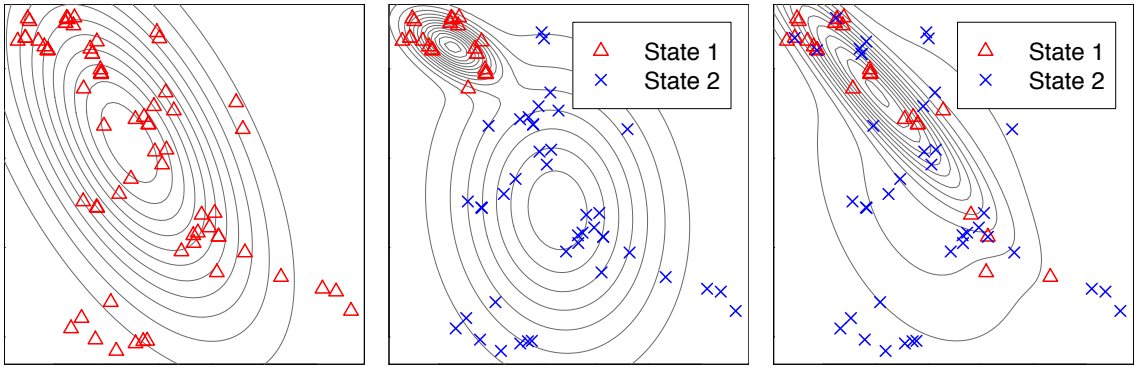


Figure 3: An example of information propagation in an LBSN



(a) GMM-Basic (b) GMM-Spatial (c) GMM-Temporal

Figure 4: Examples of Gaussian-based mobility models

LBSN to infer the propagation probability in LBSNs. There are two kinds of mobility model proposed: one is Gaussian-based models and the other is distance-based models. Based on these two models, we further derive the corresponding probability functions.

4.2.1 *Gaussian-based Mobility Models*

Prior works [5][6] utilize the Gaussian-based mobility models to model individual users' check-in behavior in LBSNs. However, in prior works [5][6], the mobility model is temporal related. In our work, we do not care about the probability of a location at a specific time since we just care about whether each user visits the target location or not. Thus, we show three major types of Gaussian-based mobility model based on existing works [5][6].

Gaussian-based mobility model (denoted as GMM-Basic): This is the basic form of Gaussian-based mobility model. Each user is modeled by one bivariate Gaussian distribution. Given the check-in records $\{(u,\ell = (x,y),t)\}$ of user u, the u's mobility model is as follows:

$$p(\ell) = \frac{1}{2\pi\sqrt{|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\ell-\mu)^T\Sigma^{-1}(\ell-\mu)\right) \quad (1)$$

and the parameters of $p(\ell)$ can be estimated by the following equations: $n = |\{(u,\ell = (x,y),t)\}|$, $\hat{\mu} = \frac{1}{n}\sum[x,y]$ and $\hat{\Sigma} = \frac{1}{n}\sum([x,y]-\hat{\mu})^T([x,y]-\hat{\mu})$.

Figure 4(a) shows an example of GMM-Basic. Only one Gaussian distribution represents each user's check-in behavior. On the left side, GMM-Basic has higher probability density, but there is no check-in record there. Thus, only one Gaussian distribution could not describe each user's check-in behavior well.

Gaussian-based mixture mobility model based on spatial information (denoted as GMM-Spatial): Each user's check-in records can be divided into several states, and each state can be modeled by one Gaussian distribution. Given the check-in records $\{(u,\ell = (x,y),t)\}$ of user u, u's mobility model is as follows:

$$p(\ell) = \sum_{i=1}^{K} p(\ell|Z_i)p(Z_i) \quad (2)$$

where K is the number of states, $p(\ell|Z_i)$ denotes the probability density of ℓ in the state Z_i, and $p(Z_i)$ denotes the probability of state Z_i. $p(Z_i)$ and the parameters of $p(\ell|Z_i)$ can be learned by the EM algorithm.

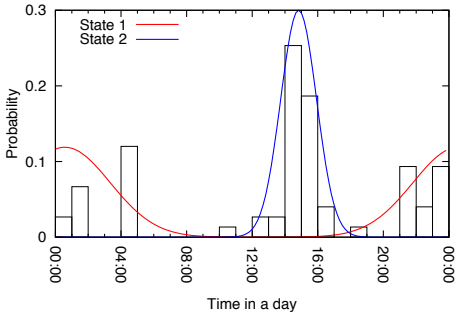


Figure 5: Histogram of the check-in timestamp of a user in a day. The states correspond to Figure 4(c)

As presented in [5], K is set to 2 and these states represent the home and work states. Figure 4(b) shows an example of GMM-Spatial with two states. The user has 75 check-in records in the area with latitude between 30.3N and 30.6N, and longitude between 97.9W and 97.7W. Note that the x-axis is longitude and the y-axis is latitude. Two Gaussian distributions are used to represent each user's check-in behavior with two different states. Obviously, GMM-Spatial is better than GMM-Basic for describing each user's check-in behavior. However, each state could not be well represented by one Gaussian distribution such as state 2 in Figure 4(b).

Gaussian-based mixture mobility model based on temporal information (denoted as GMM-Temporal): Since users may have different states of check-in behavior at different time, wrapped Gaussian distribution[1] is selected to describe the work state and home state on the timeline [5]. Given the check-in records $\{(u,\ell = (x,y),t)\}$ of user u, u's mobility model is the same as GMM-Spatial (Equation 2) and is also learned via the EM algorithm. However, in the E-step of the EM algorithm, the check-in records are classified by the Bayesian classifier on the timeline in one

[1]In [5], the authors selected the truncated Gaussian distribution for the time distribution. However, the time of a day is circular. Thus, the wrapped Gaussian distribution is better than truncated Gaussian distribution for the time distribution here.

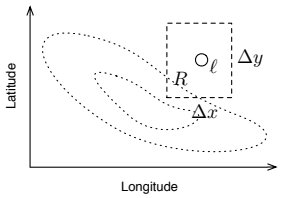


Figure 6: An example of calculating the propagation probability in GMMs

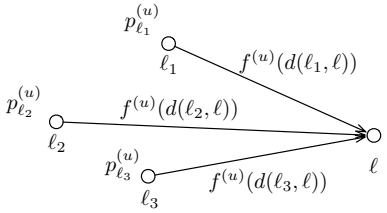


Figure 7: The concept of the distance-based mobility model

day. The PDF of a time t of each state is as follows:

$$p(t|Z_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\left(\frac{-(t-\mu_i+2\pi k)^2}{2\sigma_i^2}\right)$$

where the time t is mapped into $[0, 2\pi)$.

Figure 4(c) shows an example of GMM-Temporal with two states, and Figure 5 shows the corresponding check-in time distribution in a day. The distribution is not obvious to distinguish into two bivariate Gaussian distributions since the check-in records are classified by the timestamp but not the coordinate. Thus, the two bivariate Gaussian distributions overlap. However, for each type of check-in record, it is also hard to describe the check-in records completely.

Deriving propagation probability: The Gaussian-based mobility models are two-dimensional distribution. Thus, it is necessary to select a region to calculate the probability. Figure 6 shows an example of calculating the propagation probability in GMMs, where the target location ℓ, and the region R around ℓ are controlled by$\Delta\ x$ an$\Delta\ y$. The probability of u checking in at ℓ is as follows:

$$P_u(\ell) = \int_R p(\ell) \mathrm{d}A \tag{3}$$

where the $p(\ell)$ is from Equation 1 or 2. The propagation probability of edge (v, u) is the probability of u checking in at the target location ℓ. Hence the propagation probability of the edge from v to u is as follows:

$$p_{v\to u} = P_u(\ell) \tag{4}$$

where u follows v in an LBSN. However, it is hard to set the size of$\Delta\ x$ and$\Delta\ y$. Thus, we will show another approach to derive the propagation probability in an LBSN.

Estimation error: The estimation error describe the error between the mobility model from observations and real mobility model. Here we select Mean Square Error (MSE) to describe the situation. The details are as follows:

$$\mathrm{MSE}(\hat{\mu}) = \mathrm{Var}(\hat{\mu}) = \frac{1}{n}[\sigma_x^2 + \sigma_y^2]$$

and

$$\begin{aligned}\mathrm{MSE}(\hat{\Sigma}) &= \mathrm{Var}(\hat{\Sigma}) = trace(\mathrm{Var}(S^2)) \\ &= \frac{2}{n-1}\sigma_x^2 + \frac{2}{n-1}\sigma_y^2 = \frac{4}{n-1}(\sigma_x^2 + \sigma_y^2)\end{aligned}$$

Thus, the MSE of $\hat{\mu}$ and $\hat{\Sigma}$ are $O(\frac{1}{n})$, where n denotes the number of check-in records.

4.2.2 Distance-based Mobility Models

Since it is hard to calculate the probability from Gaussian-based Mobility Models, we use the distance instead of coordination of location to describe each user's mobility model. Moreover, using a Gaussian-based mobility model can not capture the order of the check-in records. We argue that different orders of check-in records have different mobility models. However, the distance-based mobility model takes the order of check-in records into account. Thus, we select the distance as an important feature to capture each user's check-in behavior.

Distance-based mobility model (denoted as DMM-Basic): The idea of distance-based mobility models is that it estimates the probability of a user moving from their visited locations to the target location. Therefore, distance-based mobility models have two layers. The first layer is the stationary distribution of visited locations, and the second layer is the probability density of moving from the visited locations to the target location. The stationary distribution is from random walk with restart since it is used to simulate the movements of moving objects [11]. Moreover, we select the Pareto distribution for the distance distribution since the movement distance in LBSNs has a self-similar property [23]. Based on the concept above, the probability density of a location ℓ, $p_u(\ell)$, can be shown as follows:

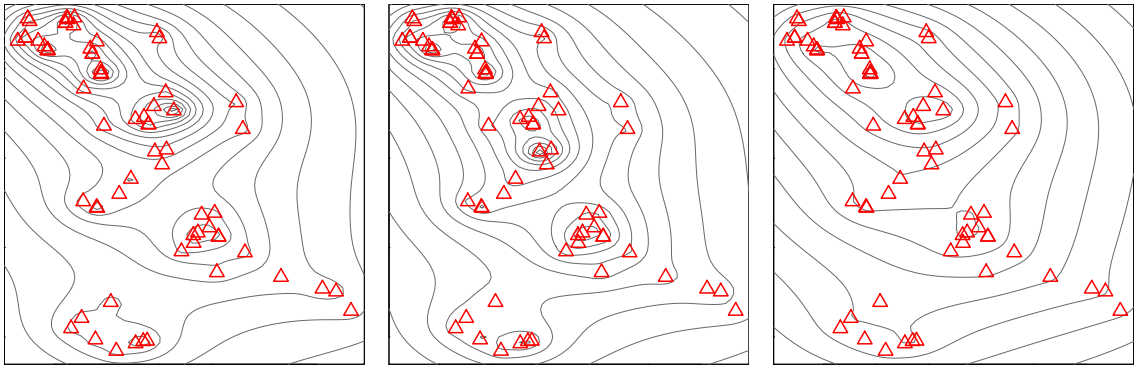
$$\begin{aligned}p_u(\ell) &= \sum_l P(u \text{ moves from } l \text{ to } \ell) \\ &= \sum_l P(u \text{ is at } l) P(u \text{ moves distance } d(l,\ell) \text{ from } l) \\ &= \sum_l p_l^{(u)} f^{(u)}(d(l,\ell))\end{aligned} \tag{5}$$

where $p_l^{(u)}$ denotes the stationary probability of user u at visited location l, $f^{(u)}(d(l,\ell))$ denotes the probability density of user u moving from a visited location l to the target location ℓ, and $d(l,\ell)$ denotes the distance between l and ℓ.

Figure 7 shows an example of the distance-based mobility model. The user u has three visited locations ℓ_1, ℓ_2 and ℓ_3. The stationary probabilities of ℓ_1, ℓ_2 and ℓ_3 are $p_{\ell_1}^{(u)}$, $p_{\ell_2}^{(u)}$ and $p_{\ell_3}^{(u)}$, respectively. Assume that u has five check-in records (u,ℓ_1,t_1), (u,ℓ_2,t_2), (u,ℓ_3,t_3), (u,ℓ_1,t_4) and (u,ℓ_3,t_5), where $t_i < t_j$ if $i < j$. Thus, the transition probabilities are $P(\ell_1 \to \ell_2) = 0.5$, $P(\ell_1 \to \ell_3) = 0.5$, $P(\ell_2 \to \ell_3) = 1$ and $P(\ell_3 \to \ell_1) = 1$. Then the stationary distribution of locations is $p_{\ell_1}^{(u)} = 0.39$, $p_{\ell_2}^{(u)} = 0.21$ and $p_{\ell_3}^{(u)} = 0.40$, where the restart probability is 0.15 and the initial distribution of ℓ_1, ℓ_2 and ℓ_3 is uniform.

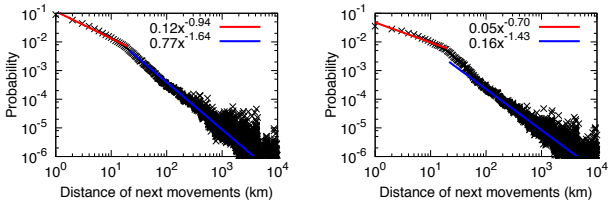
The distance distribution is described by the Pareto distribution. The form of Pareto distribution is as follows:

$$f(x;\alpha,\beta) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}} \tag{6}$$



(a) DMM-Basic (b) DMM-Basic (different order) (c) DMM-Social

Figure 8: Examples of distance-based mobility models, where the user is the same as in Figure 4



(a) gowalla (b) brightkite

Figure 9: The distribution of distance of next movement in two different datasets

where α is the shape parameter and β is the minimum of x. Since β cannot be 0, we fix $\beta = 1$. Thus, the Equation 6 can be written as follows:

$$f(x;\alpha) = \frac{\alpha}{x^{\alpha+1}} \tag{7}$$

To utilize Pareto distribution to describe the movement distance of each user, we have to evaluate a suitable α for each user from their distance observations. Assume there are n distance observations $x_1, x_2, \cdots, x_n$. Since β is fixed to 1 and the minimum of x_i is 0, x_i is shifted 1 to y_i where $y_i = x_i + 1$. Based on MLE, the estimator of α is as follows:

$$\hat{\alpha} = \frac{n}{\sum_i \ln(y_i)} = \frac{n}{\sum_i \ln(x_i + 1)} \tag{8}$$

Figure 8(a) shows an example of DMM-Basic. DMM-Basic is more representative than GMMs for individual check-in behavior. If we modify the order of the check-in records, Figure 8(b) shows the result. If the order is modified, the model is different in DMM. It is different from GMMs since GMMs focus on the coordinates of the check-in records but not on the relation between check-in records. There are two differences if the order is modified. First, the distribution of higher probability density in Figure 8(a) is different from the distribution of higher probability density in Figure 8(a). It shows that the stationary distribution is changed if the order is modified. Second, the distribution probability density in Figure 8(b) is smoother. It shows that the user is more likely to move to distant locations.

Distance-based mobility model with social information (denoted as DMM-Social): Users' check-in behavior can be influenced by their friends in LBSNs [5][22]. Therefore, the moving distances are divided into two parts, self movement behavior and social influenced moving behavior. Self movement behavior means that the movement is based on their movement behavior, and social influenced moving behavior means the movement is influenced by their friends. Figure 9 shows the distribution of distance of next movement in two different datasets. The distributions can be divided into two parts. Similar observations are also found in [5]. The distance-based with social mobility model can be shown as follows:

$$p(\ell) = \sum_l p_l^{(u)} [f_{\mathrm{M}}^{(u)}(d(l,\ell))p(\mathrm{M}) + f_{\mathrm{S}}^{(u)}(d(l,\ell))p(\mathrm{S})] \tag{9}$$

where $p(\mathrm{M})$ and $p(\mathrm{S})$ denote the probability of a distance contributed by mobility and social influence, respectively. Moreover, $f_{\mathrm{M}}^{(u)}(d(l,\ell))$ and $f_{\mathrm{S}}^{(u)}(d(l,\ell))$ denote the probability of moving $d(l,\ell)$ contributed by mobility and social influence, respectively. $p(\mathrm{M})$, $p(\mathrm{S})$ and the parameters of $f_{\mathrm{M}}^{(u)}(d(l,\ell))$ and $f_{\mathrm{S}}^{(u)}(d(l,\ell))$ can be learned by the EM algorithm. The Pareto distribution is also utilized for the distance distribution $f_{\mathrm{M}}^{(u)}(\cdot)$ and $f_{\mathrm{S}}^{(u)}(\cdot)$.

Figure 8(c) shows an example of DMM-Social. DMM-Social is smoother than the DMM-Basic in Figure 8(a) and 8(b). It reflects the distance which will be influenced by friends in Figure 9. Moreover, the coordinate of check-in records is also considered.

Deriving propagation probability: If a user tends to travel a larger distance than that between the user and the target location, there is a higher chance that the user will visit the target location. Thus, to calculate the probability based on a distance-based mobility model, we just calculate the probability of moving the distance which exceeds the distance between the visited locations and the target location. Thus, the probability of u check-in at ℓ is as follows:

$$P_u(\ell) = \sum_l p_l^{(u)} \int_{d(l,\ell)+1}^{\infty} f^{(u)}(x)\mathrm{d}x \tag{10}$$

Furthermore, in DMM-Basic, the probability is as follows:

$$P_u(\ell) = \sum_l \frac{p_l^{(u)}}{(d(l,\ell)+1)^{\hat{\alpha}}}$$

Table 1: The number of users in the different numbers of training records

	# of training records										total users
	10	20	30	40	50	60	70	80	90	100	
gowalla	1,568	10,648	444	4,604	244	2,560	137	1,614	94	1,079	22,992
brightkite	636	315	215	143	111	80	62	64	40	31	1,697

where $\hat{\alpha}$ is from Equation 8. Then, in DMM-Social, the probability is as follows:

$$P_u(\ell) = \sum_l p_l^{(u)} [\frac{p(\mathrm{M})}{(d(l,\ell)+1)^{\hat{\alpha}_\mathrm{M}}} + \frac{p(\mathrm{S})}{(d(l,\ell)+1)^{\hat{\alpha}_\mathrm{S}}}]$$

where $p(\mathrm{M})$, $p(\mathrm{M})$, $\hat{\alpha}_\mathrm{M}$ and $\hat{\alpha}_\mathrm{S}$ are from Equation 9. Furthermore, the propagation probability on edges is also the same as in Equation 4.

Estimation error: The estimation error of the distance-based mobility model is as follows:

$$\begin{aligned}\mathrm{MSE}(\hat{a}) = \mathrm{Var}(\hat{a}) &= \mathrm{Var}(\frac{n}{\sum_i \ln x_i}) \\ &= n^2 \,\mathrm{Var}(\frac{1}{\sum_i y_i}), \text{ where } y = \ln x \\ &\approx n^2 \frac{a^4}{n^3} = \frac{a^4}{n}\end{aligned}$$

The MSE of $\hat{\alpha}$ is $O(\frac{1}{n})$. The results[2] show that the estimation error of the distance-based mobility models is similar to the estimation error of the Gaussian-based mobility models.

5. PERFORMANCE EVALUATION

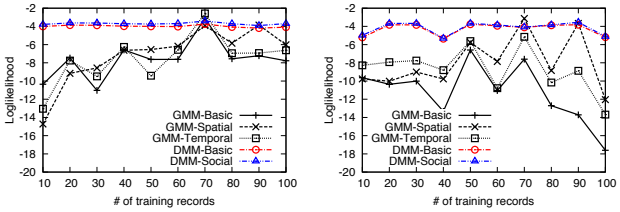
In this section, extensive experiments are conducted to evaluate the effectiveness of our proposed mobility models, GMMs and DMMs. Moreover, we also show the results of static propagation probability and location-aware propagation probability in LBSNs. We implemented the proposed models in Python.

5.1 Datasets Description

In this paper, we have selected the gowalla and brightkite datasets [5] for observation and evaluation[3]. There are 196,591 users, 950,327 social connections and 6,442,890 check-ins during February 2009 - October 2010 in the gowalla dataset. Moreover, there are 58,228 users, 214,078 social connections and 4,491,143 check-ins during April 2008 - October 2010 in the brightkite dataset.

5.2 Comparison of Different Mobility Models

To compare the performance of different mobility models, log-likelihood is selected to measure the mobility models for LBSNs since the mobility model is a two-dimensional probability density distribution [5][6][17]. If the value of log-likelihood is higher, the mobility model represents the real check-in behavior of each user well. Due to the spatial and temporal sparsity issue of check-in records, most users seldom perform check-in records ($10-100$ records). Therefore, we observe the relation between the number of training data and log-likelihood of each mobility model here. The proportion of training and testing of check-in records are 80% and 20%, and the number of training records are 10, 20, $\cdots$, 100 in the two datasets. The details of the number of users are given in Table 1.



(a) gowalla (b) brightkite

Figure 10: The results of GMMs and DMMs with different numbers of training records in different LBSNs

Figure 10 shows the results of comparison with different numbers of training records in the two datasets. In the gowalla dataset, DMMs present higher log-likelihood value than GMMs. Since the state-of-the-art approaches [5][6][17] proposed Gaussian-based mobility models, our proposed GMMs further consider spatial and temporal information. As shown in Figure 10, DMMs can better reflect individual check-in behavior than GMMs, which demonstrates the advantage of DMMs. However, when the number of training data is 70, all approaches have higher log-likelihood value. It can be considered as a bias since the number of data which contain 70 training records is lower than others (Table 1). On the other hand, DMM-Social has higher log-likelihood value than DMM-Basic. By exploring social information in LBSNs, DMM-Social is closer to real individual check-in behavior. Furthermore, DMMs can capture individual check-in behavior by only 10 training records. It shows that DMMs deal with the spatial and temporal sparsity issue of check-in records. The results in brightkite are similar to the results in gowalla. However, the amplitude of GMMs is large since the number of data we selected is lower.

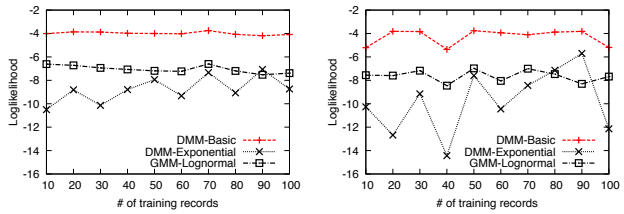
5.3 Comparison of Different Distributions for Distance

In DMMs, the Pareto distribution is selected to represent the distance of the next movement of each user. Some related distributions (such as exponential and log-normal distribution) are listed. The domain of these two distributions is $[0, \infty)$, and they are related to Pareto and Gaussian distribution, respectively. To compare different distance distributions in DMMs, DMM-Basic is selected since it is the simplest form in DMMs. Then, the distance distribution in DMM-Basic is replaced with exponential and log-normal distribution, respectively. The modified DMM-Basic is called DMM-Exponential and DMM-Lognormal, respectively.

Figure 11 shows the results of the comparison of DMM-Basic, DMM-Exponential and DMM-Lognormal. DMM-

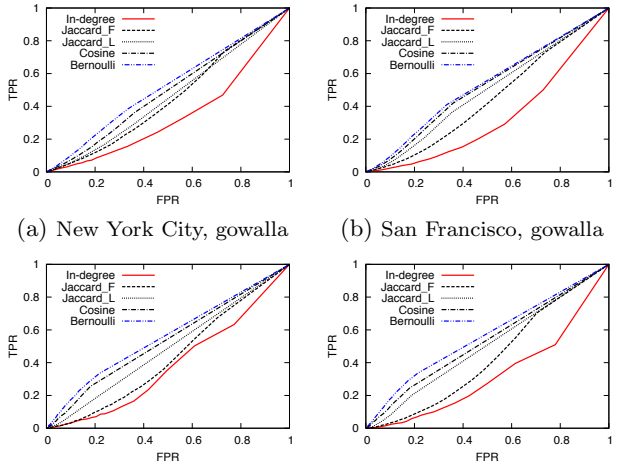
[2]The approximation is based on the delta method.

[3]Both of these datasets are public datasets which can be downloaded from https://snap.stanford.edu/data/.



(a) gowalla (b) brightkite

Figure 11: The results of three distance distributions in DMM-Basic with different numbers of training records in different LBSNs



(a) New York City, gowalla (b) San Francisco, gowalla

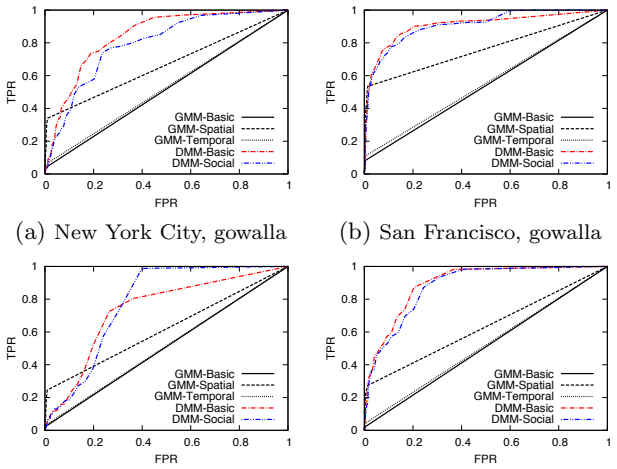
(c) New York City, brightkite (d) San Francisco, brightkite

Figure 12: The results of five approaches for static propagation probability in different LBSNs

Basic has higher log-likelihood value in two datasets. DMM-Exponential has the lowest log-likelihood value such that the distance of next movement is not memoryless. Finally, the results show that the Pareto distribution is more suitable for the distance of next movements in LBSNs than the exponential and log-normal distribution in both datasets.

5.4 Comparison of Methods for Propagation Probability

To compare different approaches to set propagation probability, the ROC curve is selected to show the results of different global activation thresholds [1][9]. Two target locations are selected: one is San Francisco Caltrain Station, San Francisco (37.776430N 122.394318W), and the other is Central Park, New York City (40.780606N 73.968088W). Moreover, we only selected users who have 10 check-in records or above. There are 67,653 users and 629,031 edges selected in the gowalla dataset, and 24,100 users and 243,922 edges selected in the brightkite dataset. To determine whether a user is active or not, we set the active range of the target location. Note that the active range is also used to calculate the propagation probability of GMMs. If one of the user's check-in records is in the active range, the user is active in the ground truth. In the experiments, the radius of the active range is set to 500m. For the approaches in static propagation probability, five approaches are selected in the



(a) New York City, gowalla (b) San Francisco, gowalla

(c) New York City, brightkite (d) San Francisco, brightkite

Figure 13: The results of five approaches for location-aware propagation probability in different LBSNs

experiments, where In-degree denotes in-degree of nodes, Jaccard_F denotes jaccard index of friends, Jaccard_L denotes jaccard index of locations, Cosine denotes cosine of locations, and Bernoulli denotes Bernoulli estimator.

Figure 12 shows the results of five approaches for static propagation probability. In the gowalla dataset (Figure 12(a) and 12(b)), Bernoulli performs well than other approaches of static probability but not significant. Although Bernoulli learns from history records, it can not adapt to the information propagation in LBSNs. The ROC curves of Jaccard_F, Jaccard_L and Cosine are closest to the diagonal. Thus, the number of common friends and locations are not related to the information propagation in LBSNs. Furthermore, In-degree, the most common used method in social networks [4][13][16], is worse than random guess. Thus, the in-degree of nodes and the degree of information propagation in LBSNs are not helpful in the location promotion problem. Finally, the results show that the ROC curves of five approaches are closed to the diagonal in both target locations. From the above experimental results, the static propagation probability is not able to truly reflect the information propagation in LBSNs. The results in the brightkite dataset are shown in Figure 12(c) and 12(d), and these results depict the similar phenomenon.

Figure 13 shows the results of five approaches for location-aware propagation probability. In the gowalla dataset (Figure 13(a) and 13(b)), DMM-Basic has the largest AUC (Standing for Area Under Curve), where the target location is set to NYC. On the other hand, DMM-Basic and DMM-Social have similar AUC when the target location is set to San Francisco. Both DMM-Basic and DMM-Social has a better performance with different target locations. Moreover, GMMs have the lowest AUC and these curves of GMMs are close to the diagonal, where the diagonal curve represents the results of random guess. The reason is that the location-aware propagation probability from GMMs is controlled by the area around the target location. If the area is large, and then the probability is higher. It reflects the disadvantage that it is hard to derive the location-aware propagation

probability from GMMs. Particularly, The GMM-Spatial has the best performance in GMMs. It shows individual check-in behavior can be divided into many states by spatial information. However, GMM-Spatial is limited by the disadvantage of calculating location-aware propagation of GMMs. As such, GMM-Spatial has lower AUC than DMMs. Finally, DMMs have larger AUC than GMMs, which indicates that DMMs are suitable to reflect the information propagation and derive the location-aware propagation in LBSNs. The results shown in Figure 13(c) and 13(d) are experiments in the brightkite dataset, which indicate the similar results in the gowalla dataset.

6. CONCLUSION

In this paper, we addressed on the location promotion problem in LBSNs. Explicitly, the location promotion problem is formulated as an influence maximization problem. Given a target location and an LBSN, we aimed at deriving a set of k seed users to maximize the number of influenced users to visit the target location. The most challenging is to derive the propagation probability with different target locations in LBSNs. By referring to prior works on influence maximization, we developed some baselines to determine the propagation probabilities. Note that we claimed that user mobility should be considered in deriving propagation probabilities in LBSNs. Users should reach/visit the target location and thus will spread the target location information to their friends. Since the information propagation is triggered by check-in in LBSNs, we propose Gaussian-based mobility models, GMMs, and distance-based mobility models, DMMs, to capture individual check-in behavior in LBSNs. DMMs have the advantage of not only capturing individual check-in behavior but also deriving the propagation probability in LBSNs. The experimental results show that DMMs perform better than other state-of-the-art approaches in terms of capturing individual check-in behavior and dealing with the spatial and temporal sparsity issue of the check-in data since it considers not only corrdinates of check-in records but also the relations between check-in records. Moreover, DMMs really reflects the dynamic propagation probability with different target locations in LBSNs.

Acknowledgments

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 102-2221-E-009-171-MY3 and MOST 101-2628-E-001-004-MY3), Academia Sinica (AS-102-TP-A06), and Taiwan MoE ATU Program.

7. REFERENCES

[1] N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. In *IEEE ICDM*, 2012.

[2] P. Bouros, D. Sacharidis, and N. Bikakis. Regionally influential users in location-aware social networks. In *ACM GIS*, 2014.

[3] S. Chen, J. Fan, G. Li, J. Feng, K.-l. Tan, and J. Tang. Online topic-aware influence maximization. In *VLDB*, 2015.

[4] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *ACM KDD*, 2010.

[5] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *ACM KDD*, 2011.

[6] H. Gao, J. Tang, X. Hu, and H. Liu. Modeling temporal effects of human mobile behavior on location-based social networks. In *ACM CIKM*, 2013.

[7] F. Giannotti, M. Nanni, and D. Pedreschi. Efficient mining of temporally annotated sequences. In *SIAM SDM*, 2006.

[8] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *ACM KDD*, 2007.

[9] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. Learning influence probabilities in social networks. In *ACM WSDM*, 2010.

[10] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. A data-based approach to social influence maximization. In *VLDB*, 2012.

[11] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd edition, 2011.

[12] P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. In *SSTD*, 2005.

[13] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *ACM KDD*, 2003.

[14] D. Kempe, J. Kleinberg, and E. Tardos. Influential nodes in a diffusion model for social networks. In *ICALP*, 2005.

[15] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *ACM KDD*, 2007.

[16] G. Li, S. Chen, J. Feng, K.-l. Tan, and W.-S. Li. Efficient location-aware influence maximization. In *ACM SIGMOD*, 2014.

[17] M. Lichman and P. Smyth. Modeling human location data with mixtures of kernel densities. In *ACM KDD*, 2014.

[18] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *ACM CIKM*, 2010.

[19] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *AAAI ICWSM*, 2011.

[20] K. Saito, R. Nakano, and M. Kimura. Prediction of information diffusion probabilities for independent cascade model. In *KES*, 2008.

[21] H.-H. Wu and M.-Y. Yeh. Influential nodes in one-wave diffusion model for location-based social networks. In *PAKDD*, 2013.

[22] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *ACM SIGIR*, 2011.

[23] W.-Y. Zhu, W.-C. Peng, and L.-J. Chen. Exploiting mobility for location promotion in location-based social networks. In *IEEE DSAA*, 2014.