GLP: A Novel Framework for Group-level Location Promotion in Geo-social Networks

Xudong Wu, Luoyi Fu, Yuhang Yao, Xinzhe Fu, Xinbing Wang and Guihai Chen
Shanghai Jiao Tong University
{xudongwu, yiluofu, yaohuhang, fxz0114, xwang8}@sjtu.edu.cn, gchen@cs.sjtu.edu.cn

Abstract—**Location-aware viral marketing is crucial in modern commercial applications for attracting customers to certain Points Of Interests (POIs). Prior works are mainly based on formulating it into a location-aware influence maximization (IM) problem in Geo-social Networks (GSNs), where K initial seed individuals are selected in hope of maximizing the number of final influenced users. In this paper, we present a first look into group-level location promotion which can potentially enhance its performance, with the phenomenon that users belonging to the same geo-community share similar moving preferences.**

We propose GLP, a new and novel framework of group-level location promotion by virtue of geo-communities, each of which is treated as a group in GSNs. Aiming to attract more users to designated locations, GLP firstly carries out user grouping through an iterative learning approach based on information extraction from massive check-ins records. The advantage of GLP is three-folded: i) by aggregating movements of group members, GLP significantly avoids the sparsity and sporadicity of individual check-ins, and thus obtains more reliable mobility models; ii) by generalizing a new group-level social graph, GLP can exponentially reduce the computational complexity of seed nodes selection that is algorithmically executed by a greedy algorithm; iii) in comparison with prior individual-level cases, GLP is theoretically demonstrated to drastically increase influence spread under the same given budget. Extensive experiments on real datasets demonstrate that GLP outperforms four baselines, with notably up to 10 times larger influence spread and 100 times faster seed selection over two individual-level cases, meanwhile verifying the impact of group numbers in final influence spread.

I. Introduction

In viral marketing, location promotion is a newly flourishing topic in both academia and industry. There has been a variety of applications of location promotion in our daily lives, including promoting newly opened restaurants, libraries and plazas. As the development of Geo-social Networks (GSNs), which provide a platform where the locations can be added as the auxiliary information of online posts, users are able to make check-in records at Points of Interest (POIs) and share the records with their friends. Many popular online social networks (e.g., Gowalla, Foursquare and Wechat) can be taken as the typical examples of GSN since the online posts over them can be tagged with the location information. Considering the fact that users are more willing to accept advertisements propagated from their friends, as the posts being propagated over the online social networks, a large number of users would be influenced through the word-of-mouth effect [1]-[4]. Thus location promotion can be modeled as a kind of location-aware Influence Maximization (IM) problem. Specifically, *given a Geo-social network G and an integer K, which K seed users should be selected to maximize the number of users eventually moving to the promoted location under probabilistic influence diffusion models [5]-[8]?*

Clearly, instead of being mainly affected by the social relationships between nodes, as is the case for traditional IM issues, the final influence of a seed set in above location-aware IM problems also depends on the moving probability of users to the promoted location. This is because in reality the users who are geographically close to the promoted POI are more likely to move to it. Hence, in previous efforts there are generally two steps in designing location promotion: (1) Obtaining the moving probability of users to the promoted location. (2) Designing effective algorithms for the location-aware IM problem which has been proved to be NP-hard [2][3][5][9]. Following these two steps, existing studies are mainly on the individual level, which proposed to compute moving probability based on users' individual mobility models and greedy algorithms are designed to select K individual seed users [1][2][10]. In contrast, in this paper we claim that it will be more practical to design location promotion on group level given the large scale GSNs and sparse individual check-ins data. Our insight is that users with common properties may share similar mobility patterns, which naturally lead to the geo-communities that are widely observed in real situations [11][12]. One typical example is the mobility pattern of residents in the same town, where there are multiple common well-visited locations like residential areas, shopping malls and churches. Another instance belongs to the students of a same university, as they may regularly move from halls to their labs, and from classrooms to restaurants or playgrounds. Thus, defining the users with similar mobility as a group in GSNs, we intend to investigate the following group-level location promotion problem: *Given the generalized group-level Geo-social graph G_{group} , which K groups should be selected to maximize the final influence of location promotion?* By elevating the location promotion from individual level to the group level, its performance could be significantly improved from the following three aspects:

(1) The mobility model on group level is far more reliable. In reality, each user is unlikely to make check-ins at every visited location, but instead prefer to make check-ins at which impresses them [13]. Such selective check-in behaviors result in the sparsity and sporadicity of individual check-ins, which brings the unreliability of determining moving probability from individual mobilities in step (1). In contrast, on group level, members of a same geo-community share significant regular movements. By aggregating their movements, we can deal with the sparsity and sporadicity of individual check-ins and generalize group-level mobility models.

(2) Computational complexity on group level can be largely reduced. Recall step (2), the state-of-the-art greedy algorithms are unscalable to very large GSNs due to the exponential computational complexity [1][2][14]. On group level, we transfer social graph from individual level to group level, over which the number of groups can be scaled as $m = \log(n)$ (where n is the number of individuals) in very large-scale networks [14]. Thus the computational complexity of solving IM problem on group level can be largely reduced.

(3) Group-level promotion can significantly increase the influence spreading. On individual level, even though the selection of K seed users can be completed through efficient heuristics [3] in step (2), it still remains very challenging to attract the selected K users to visit the promoted location where they can make check-ins, e.g., advertising via local celebrities may be extremely expensive. On group level, a far larger number of seed users can be convinced through initial propagation under the same budget.

The following motivated example illustrates the above three advantages.

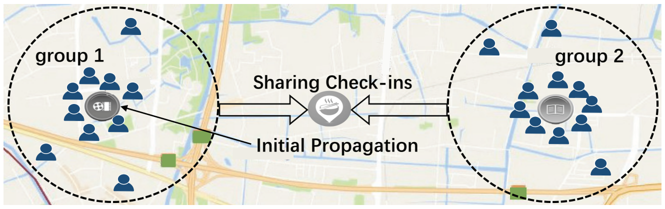


Fig. 1: A motivated example

A Motivated Example: Let us consider the case where there are two groups of users around a newly opened restaurant, with residences of a housing estate forming group 1 and students of a same university being group 2, as illustrated in Figure 1. The moving probability of the two groups can be obtained based on the distance from the restaurant to their well-visited locations, i.e., a nearby library and cinema respectively. Assuming that the moving probability is same for the two groups, the objective of step (2) is to select a more influential one to maximize the number of users eventually moving to the promoted restaurant. As shown in Figure 1, the initial propagation for the two groups can be the cinema advertising and a billboard near the library respectively. The cost for cinema advertising in North America is roughly $\$30$ per screen per week [15], while the weekly rental for a billboard is about $\$150$-$\$600$ [16]. Due to the cost-effectiveness of cinema advertising, group 1 is selected as seed and the audience of the cinema are all the potential seed users in the location promotion process. While on the individual level, the same budget may only be able to activate a few seed users. Thus the initial propagation on group level is much more effective than that on individual level.

Incorporating the three advantages brought by groups, we propose GLP, a novel framework for the location promotion on group level. While we defer the details of GLP to later sections (Sections IV and V), here we briefly summarize its main ingredients: it firstly designs an iterative approach to mine geo-communities from massive sparse individual check-ins data. Since users belonging to a same geo-community spatially transit among several well-visited locations, we characterize the common mobility models as Hidden Markov Models (HMMs). Then the Bayesian Classifier is adopted to group users sharing similar mobilities under HMMs, thus the reliable moving probabilities can be extracted from common group-level mobility models. Jointly considering the intra and inter-group influences among groups (geo-communities) which refer to influences diffused among group mates and members of different groups, we redefine the influences diffusion model among groups which enables us to efficiently select seed groups from a generated group-level social graph. Upon proving that the location-aware IM is still an NP-hard problem under this new model, we propose a greedy algorithm ($\mathcal{GLP}$)[1] that efficiently selects the K seed groups and thus leads to largely reduced computational complexity in the group-level graph. Meanwhile, $\mathcal{GLP}$ is proved to achieve an approximation ratio of nearly $(1 - \frac{1}{e})$. Finally, we theoretically prove that the GLP framework can significantly improve the expected influence diffusion size compared with that on individual level. The superiority of GLP is also empirically justified.

Our principle contributions are summarized as follows.

1. We formally introduce a novel framework, GLP, for group-level location promotion (a location-aware IM problem). To the best of our knowledge, this is the first study that designs location-aware IM problem on the group level.
2. We innovatively propose to select well visited locations of seed groups to conduct initial propagation, which leads to a larger number of influenced seed users. Furthermore, we theoretically prove that GLP can drastically improve the performance of location promotion under the same budget.
3. We empirically evaluate GLP on three real GSNs (Brightkite, Gowalla and Foursquare) datasets, where GLP is demonstrated to increase final influence up to 10 times, along with a node selection that is 100 times faster than that on individual level. The results also exhibit the heterogeneous geographical distribution of users, as well as the effect of such heterogeneous distribution on the performance of location promotion.

Paper structure: We review related studies in Section II. The problem is formulated in Section III. In Section IV, we introduce the approach for mining geo-communities and computing group-level moving probability. In Section V, we investigate the group-level location promotion algorithm. We evaluate our framework on real-life GSNs data in Section VI and conclude the paper in Section VII.

II. Related work

(1) Location-aware influence maximization. In general, location promotion can be formulated as a location-aware IM problem that has been widely studied in recent years. In [1], each user is supposed to stay at a fixed location. Given a specific area R, the authors design to select K seed users

[1] Throughout the paper, GLP refers to the proposed framework while $\mathcal{GLP}$ means the proposed algorithm under the GLP framework.

in order to maximize the influence spread to users stay in R. However, for location promotion, it is difficult to determine the appropriate target area around the promoted location. In [10], the authors proposed a distance-based location distribution and validate its effectiveness in describing users' moving probabilities by experimental results. The major shortcoming of this model is that it does not consider the social relations among users. In [2] and [3], the authors adopted distance-based location distribution to describe moving probability and simultaneously consider social relations among users. [3] calculates the influence of each user on their social neighbors and simply chooses K users with highest direct influence, while [2] adopts the bound based pruning strategies and achieves an approximating ratio of $\beta(1-\frac{1}{e})$.

(2) Group/community based influence maximization. Community, as a prevalent social network structure, refers to a group of users formed with similar social properties (e.g., students of the same university). Eftekhar *et al.* [14] and Lu *et al.* [17] proposed to select K groups/communities to maximize the influence diffusion in online social networks without considering users' locations. [14] also proved that solving IM problem on group-level graph can achieve better performance and higher speed comparing with that on individual level. Xiao *et al.* [18] and Fan *et al.* [19] found that members of a same group frequently visit common locations and the groups can be tagged by locations, i.e., geo-community. [18] and [19] focus on designing the routing algorithms for transmitters to fast spread information among groups.

We note that all the existing works on location promotion are focused on individual level. Besides, existing group-based IM models are based on predefined groups in online social networks and neglect the offline moving probability, and fail to be directly applied to the case of location promotion. In this paper, we conduct the first study for taking advantages of geo-communities to fundamentally improve the performance of location promotion on group level.

III. PRELIMINARIES

A. Individual-Level Location Promotion

For a clearer representation of the group-level case of interest, we start from introducing the individual-level location promotion, which has been previously formalized as a location-aware Influence Maximization (IM) problem over a Geo-Social Network (GSN) $G=(V,E)$, where V and E respectively denote the set of nodes and edges. Recall that traditional studies on individual level characterize the location promotion process by the location-aware Independent Cascade (IC) model with multiple steps [1]-[3]. Concretely, at step 0, a set $\mathcal{S}_{in}$ of seed users make check-ins at the promoted location L and share them over GSNs. The users who are convinced by the check-ins and then moving to location L are tagged as influenced. If user u_i is first influenced at step s, the check-ins made by him will influence his social neighbor u_j at and only at step $(s+1)$ successfully with probability I_{ij}. The process converges when there is no newly influenced user.

Under the above location-aware IC model, the individual influence of user u_i on user u_j (i.e., I_{ij}) is divided into two stages: (1) online stage, where u_i convinces u_j through the check-in records in GSNs; (2) offline stage, where u_j moves to the promoted location L after having been convinced online. Concretely, in stage (1), the online influence is modeled by the traditional IC model [3][5][14], which specifies that when user u_i is influenced, he has a single chance to influence his friend u_j successfully with probability w_{ij}. w_{ij} is the weight of the directed edge from u_i to u_j and is independent of the influence diffusion process. In stage (2), how u_j moves to the promoted location is characterized by the moving probability which decreases with the distance from his historical check-in locations to the promoted location L [1]-[3], i.e., P_j^L. Thus the influence of u_i on u_j in the individual-level location promotion can be defined as:

Definition 1. *(Location-aware individual influence.) The influence of user u_i on u_j in location promotion is equal to:*

$$I_{ij}=w_{ij}P_j^L,$$

where P_j^L is the moving probability of u_j to the location L.

Remark. In the formulation of the location-aware individual influence as defined in Definition 1, the influence from the friend u_i (denoted by w_{ij}) and the moving probability (denoted by P_j^L) respectively correspond to the influences on the online and the offline stages. The influence from the friend u_i is propagated online via the posts shared over the GSNs, and whether or not the user u_j could be convinced by the posts is independent of his offline state. Then in the case that the user u_j is convinced by the posts shared by his friend u_i online, he would decide whether to visit the promoted location according to his offline state. Thus we treat the influences on the two stages as independent events, and formulate the location-aware individual influence I_{ij} as $I_{ij}=w_{ij}*P_j^L$. The same formulation of the location-aware individual influence is also adopted in [2][3].

Let $I(\mathcal{S}_{in},v)$ denote the expected probability that a set of users $\mathcal{S}_{in}$ successfully influence user v under the location-aware IC model. The goal of individual-level location promotion is to select a set $\mathcal{S}_{in}^{opt}$ with size K to maximize the influence over G, i.e.,

$$\mathcal{S}_{in}^{opt}=\operatorname{argmax}_{\mathcal{S}_{in}\subseteq V}\sum_{v\in V}I(\mathcal{S}_{in},v),|\mathcal{S}_{in}|=K.$$

However, as noted in Section I, the inherent sparsity and sporadicity of individual check-ins brings unreliability to the moving probabilities of individual users (i.e., P_j^L). Besides, the task of selecting and activating seed users suffers from the limitations such as low efficiency and the high cost in seeding individual users. Hence, to fundamentally improve the performance of location promotion, in this paper, we focus on a novel group-level location promotion that takes advantages of the widely existing phenomenon of geo-communities. Particularly, modeling each geo-community as a group in GSNs, we can obtain the reliable moving probabilities by aggregating the users sharing similar mobility models. Based on that, the limitations incurred by the individual cases can be potentially tackled over the group-level GSNs. The rationale behind is that conducting initial propagation at well-visited locations of geo-communities can influence much more seed users and resolving location-aware IM problem on group-level GSNs is of high efficiency. Given the definitions and formulation of individual-

TABLE I: The summary of Notations

Notation	Meaning
$G = (V, E)$	Individual-level social graph
$G_{group} = (\mathcal{G}, \mathcal{B})$	Group-level social graph
$s_{i,h}$	The $h - th$ steady state of group g_i
P_{g_i}	Moving probability of users in group g_i
B_{ij}	weight of edge from group g_i to group g_j
ρ_i	closeness of groupg_i
$\mathcal{S}_k$	the selected seed groups set after kth iteration
$I(g_i, g_j)$	Influence of g_i on g_j
$I(g\|\mathcal{S}_k, v)$	marginal influence of group g on group v under selected seed set $\mathcal{S}_k$
$\mathcal{M}(g_i\|\mathcal{S}_k)$	marginal influence of group g_i on whole network under selected seed set $\mathcal{S}_k$

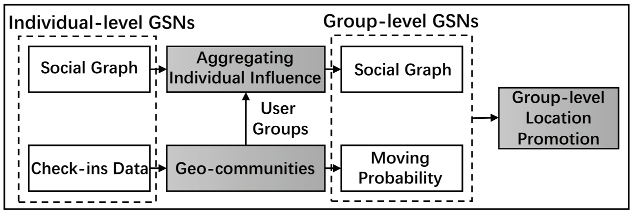


Fig. 2: The framework of GLP

level location promotion, in the sequel, we will elaborate on the formulation of group-level location promotion.

B. Group-level Location Promotion

In group-level location promotion, we consider the influences among groups instead of individual users. Different from that of individuals, the state of a group cannot be simply tagged by a binary variable, i.e., influenced or uninfluenced. Thus we quantify the state of groups, as well as the influences among groups, by percentages of influenced members. Specifically, the influence of group g_i on g_j is the aggregating location-aware individual influences of group g_i's members on g_j's, and the influenced fraction of g_j is denoted by $I(g_i, g_j)$. Particularly, $I(g_i, g_i)$ is the fraction of g_i's own influenced members, and $I(\mathcal{S}, v)$ is the influence of a set $\mathcal{S}$ of seed groups on group v. Based on the influences among groups, the goal of group-level location promotion over a given GSN is formally stated as follows.

Problem Statement: The objective of group-level location promotion is to select a set of K seed groups, i.e., S^{opt} to maximize the number of influenced members in all groups:

$$S^{opt} = \text{argmax}_{S \subseteq \mathcal{G}} \sum_{v \in \mathcal{G}} I(\mathcal{S}, v) N_v, |S| = K, \tag{1}$$

where $\mathcal{G}$ is the set of all groups in the given GSN, and N_v is the size of members in group v.

Table I summarizes the notations that will be frequently used throughout the paper. Here we note that raw input for the group-level location promotion is social relations and check-ins of individual users. However, it still remains unknown which users are the members of a same geo-community (group) and what the mobility model of each group is. Besides, the formulations of $I(g_i, g_j)$ and $I(\mathcal{S}, v)$ depend on how influences diffused among groups and remain unknown as well. Such unknown factors provide obstacles for resolving the problem in Eqn. (1). In the present work, we thus propose a novel group-level location promotion (GLP) framework to systematically resolve the difficulties above.

For ease of understanding, we present an overview of our proposed GLP framework, as illustrated in Figure 2. The input of GLP is the check-ins data of individual users, which record their historical locations, and the social relations among them. As shown in Figure 2, an iterative learning framework is firstly proposed to mine geo-communities, which are modeled as groups in GLP, from massive check-ins data. By aggregating the individual mobilities and influences of group members, we can obtain the group-level moving probabilities and social graph respectively. Then a set of seed groups is selected to maximize the expected number of users moving to the promoted location.

In next section, we will first illustrate how to mine geo-communities and determine group-level moving probabilities from massive check-ins data.

IV. Group-level moving probability

In this section, we first illustrate how to mine geo-communities from massive check-ins data, and then utilize the learnt mobility models and group membership to compute the group-level moving probability.

A. Learning Mobility Model

In daily activities, users may be driven to move both by social and spatial factors [11][20] and thus there are multiple states in their mobility models, e.g., students studying on campus (state s_1) sometimes go shopping on a commercial street (state s_2). For the mobilities of users, the observing values are users' check-in records (e.g., a post at a restaurant published on Facebook), and the states behind them are unobservable. Furthermore, the next state of a user depends on the previous one (e.g., students in classrooms will move to a restaurant after class). Therefore, we model group-level mobilities as Hidden Markov Models (HMMs), in which the state of users can be characterized as hidden contextual variables (e.g., studying, having dinner and exercising) [21][22]. Each state has a spatial probability distribution to generate check-in locations, describing the location distribution of users in a given state. [21] assumes check-in locations are generated from a bivariate Gaussian, while it is difficult to approximate the integrating area around the promoted location for computing moving probability. [10] and [13] find that the probability of moving from one location to another is proportional to negative power of the distance between them, i.e., $p \propto \Delta l^{-\alpha}$. Such probability decay of moving distance enables us to estimate the moving probabilities of users based on distances from their well-visited locations to the promoted location. Thus we adopt the distance-based Pareto distribution as the generating distribution of each state:

$$p(\Delta l) = \frac{\tau}{(\Delta l + \varepsilon)^{\tau+1}} (\varepsilon \geq 1, \tau > 0), \tag{2}$$

where Δl is the distance to the core location of a given state. [10] also experimentally proves that the distance-based Pareto distribution can effectively model moving probability because of the self-similar properties of movements distance in GSNs. Based on such distribution, we define the group-level mobility model as below.

Definition 2. *(Group-level mobility model.) The group-level mobility models are formulated as HMMs with H latent states. An H-dimensional vector π defines the steady state distribution of the H states and an $H \times H$ matrix A defines the transition probabilities among them. Given a user in state s_h, his check-in location $\widetilde{l}$ is generated based on $p(l|s_h) = \frac{\tau}{(||l-\widetilde{l}_h||_2+\varepsilon)^{\tau+1}}$, where $\widetilde{l}_h$ is the core location of state s_h.*

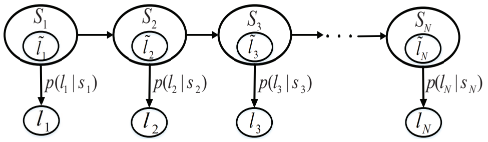


Fig. 3: Framework of Group-level mobility model

Now we introduce how to learn group-level mobility models and mine geo-communities from massive check-ins data. A geo-community is formed as a group in GSNs, and each group g_i has a HMM based mobility model with H latent states, i.e., $\pi_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,H}\}$. According to Definition 2, each HMM has a shaping parameter τ that represents the moving preference of group members, and each state respectively has a core location $\widetilde{l}_h$, e.g., a shopping mall in the commercial street. Hence, the parameter space for a HMM is $\Theta = \{\pi, A, \widetilde{l}, \tau\}$, where π is the steady distribution of states in Π_i and $\widetilde{l} = \{\widetilde{l_1}, \widetilde{l_2}, \ldots, \widetilde{l_H}\}$ is an H-dimensional vector which contains the H core locations of the H states. Let $V = \{u_1, u_2, \ldots, u_{|V|}\}$ denote the set of users. The input for mobility models learning is check-ins of individual users in V, and N consecutive check-ins locations of a user constitute a trajectory, i.e., $l_r = \{l_{r,1}, l_{r,2}, \ldots, l_{r,N}\}$ with the corresponding latent states being $D_r = \{s_{r,1}, s_{r,2}, \ldots, s_{r,N}\}$. The framework of the group-level mobility model over a given trajectory is shown in Figure 3. We use T^u to denote user u's trajectories set, where each trajectory l_r consists of N consecutive check-in locations. Given the number of check-ins of user u is N^u, we set the number of his trajectories as $floor(\frac{N^u}{N})$ ($floor(\cdot)$ is the Integer-valued function) and $|T^u| = floor(\frac{N^u}{N})$.

Given the trajectories of all users, the geo-communities mining is to iteratively train the parameters in HMMs and group users based on the posterior probability of their trajectories under each HMM. Specifically, in every loop, we first train the HMM for each group based on the Expectation-Maximization (EM) algorithm, then adopt the Bayesian Classifier to refine membership vectors which are probabilities of users belonging to each group, i.e., $p(g_i|u)$ under the refined parameters Θ_i. Notably, the initial parameters Θ of each HMM and initial membership vectors in every loop are set to those learnt from the last loop to make sure the learning process finally converge. The detail of each module is presented as below.

Initialization: The membership vector for each user u is randomly generated under the condition that $\sum_{g_i \in \mathcal{G}} p(g_i|u) = 1$. Then, for $\forall g_i \in \mathcal{G}$, we randomly initialize the parameters Θ_i for the HMM-based mobility model.

EM Algorithm: The objective in the $(t+1)$-th iteration of the EM algorithm for training Θ of a group g is to maximize the log likelihood under parameters learnt in the t-th iteration, i.e., $\Theta^{t+1} = \text{argmax}_{\Theta}\, Q(\Theta|\Theta^t)$. The objective function $Q(\Theta|\Theta^t)$ in the $(t+1)$-th iteration is

$$Q(\Theta|\Theta^t) = \sum_{r=1}^{R} \sum_{D_r} w_r P(D_r|l_r, \Theta^t) \cdot \ln p(s_{r,1}) \tag{3}$$

$$+ \sum_{r=1}^{R} \sum_{D_r} \sum_{n=2}^{N} w_r P(D_r|l_r, \Theta^t) \cdot \ln p(s_{r,n}|s_{r,n-1}) \tag{4}$$

$$+ \sum_{r=1}^{R} \sum_{D_r} \sum_{n=1}^{N} \sum_{h=1}^{H} w_r P(D_r|l_r, \Theta^t) \delta(s_{r,n}, h) \cdot \ln p(l_{r,n}|s_{r,n}), \tag{5}$$

where w_r is the mixture coefficient of l_r depends on the probability of the user of l_r belonging to group g ($\sum_{r=1}^{R} w_r = 1$), R is the total number of trajectories, and $\delta(s_{r,n}, h) = 1$ holds only when $s_{r,n} = h$. The derivation of $Q(\Theta|\Theta^t)$ and updating rules in **M-step** can be deferred to **Appendix**.

E-step: To maximize $Q(\Theta|\Theta^t)$, the E-step is to compute $P(D_r|l_r, \Theta^t)$, which is the distribution of latent states in r-th trajectory under parameters Θ^t learnt in last iteration. To this end, we first use Baum-Welch algorithm to compute two distributions, i.e., $\alpha(s_{r,n}) = P(l_{r,1}, l_{r,2}, \ldots, l_{r,n}, s_{r,n}|\Theta^t)$ and $\beta(s_{r,n}) = P(l_{r,n+1}, l_{r,n+2}, \ldots, l_{r,N}|s_{r,n}, \Theta^t)$. Based on $\alpha(s_{r,n})$ and $\beta(s_{r,n})$, we compute the following two variables of the r-th trajectory: (1) $\gamma(s_{r,n}) = p(s_{r,n}|l_r, \Theta^t)$ and (2) $\xi(s_{r,n-1}, s_{r,n}) = p(s_{r,n-1}, s_{r,n}|l_r, \Theta^t)$. Then in M-step that maximizes the objective function $Q(\Theta|\Theta^t)$, new parameters Θ^{t+1} are refined with $\gamma(s_{r,n})$ and $\xi(s_{r,n-1}, s_{r,n})$. The formulations for α, β, γ and ξ can be deferred to **Appendix**.

M-step: In the M-step, the updating rule for parameters in the $(t+1)$-th iteration Θ^{t+1} is given by

$$\pi_h^{t+1} = \sum_{r=1}^{R} w_r \gamma(s_{r,1} = h), \tag{6}$$

$$a_{ij}^{t+1} = \frac{\sum_{r=1}^{R} \sum_{n=2}^{N} w_r \xi(s_{r,n-1} = i, s_{r,n} = j)}{\sum_{r=1}^{R} \sum_{n=1}^{N-1} w_r \gamma(s_{r,n} = i)}, \tag{7}$$

$$\widetilde{l}_h^{t+1} = \frac{\sum_{r=1}^{R} \sum_{n=1}^{N} w_r \gamma(s_{r,n} = h) l_{r,n}}{\sum_{r=1}^{R} \sum_{n=1}^{N} w_r \gamma(s_{r,n} = h)}, \tag{8}$$

$$\tau^{t+1} = \frac{\sum_{r=1}^{R} \sum_{n=1}^{N} \sum_{h=1}^{H} w_r \gamma(s_{r,n} = h)}{\sum_{r=1}^{R} \sum_{n=1}^{N} \sum_{h=1}^{H} w_r \gamma(s_{r,n} = h) \ln(||l_{r,n} - \widetilde{l}_h||_2 + \varepsilon)}. \tag{9}$$

Bayesian Classifier: In every loop, we suppose the EM algorithm for training each HMM continues to the $(T+1)$-th iteration. After obtaining the refined parameters in $(T+1)$-th iteration Θ^{T+1}, we adopt the Bayesian classifier to compute the probability that users belong to each group $p(g_i|u, \Theta_i^{T+1})$. Based on the Bayesian theory, we have

$$p(u|g_i; \Theta_i^{T+1}) = \prod_{r=1}^{|T^u|} p(l_r^u|g_i, \Theta_i^{T+1}),$$

where $p(l_r^u|g_i, \Theta_i^{T+1}) = \sum_{h=1}^{H} \alpha(s_{r,N} = h|g_i, \Theta_i^{T+1})$, and the probability that u belongs to g_i is $p(g_i)p(u|g_i; \Theta_i^{T+1})/p(u)$. Here, $p(g_i)$ is the prior probability of group g_i, which is computed based on $p(g_i|u)$ in previous loop, i.e., $p(g_i) = \sum_{u \in V} p^{last}(g_i|u)/|V|$, and $p(u) = \sum_{i=1}^{|\mathcal{G}|} p(g_i)p(u|g_i, \Theta_i^{T+1})$ is the normalization divisor in the computation. Thus the probability of a user belonging to group g_i can be updated as

$$p(g_i|u) = \frac{p(g_i)p(u|g_i;\Theta_i^{T+1})}{\sum_{i=1}^{|\mathcal{G}|} p(g_i)p(u|g_i,\Theta_i^{T+1})}.$$

Then, we update w_r of each trajectory for each group and enter next loop until $p(g_i|u)(g_i \in \mathcal{G}, u \in V)$ converges or the learning process reaches preset loop times. Completing the learning process, we obtain the mobility model of each group which is an HMM parameterized by $\Theta = \{\pi, A, \widetilde{l}, \tau\}$. At the same time, users are grouped by their membership vectors $p(g_i|u)(g_i \in \mathcal{G})$, i.e., user u belongs to group g_i with the probability $p(g_i|u)$, belongs to group g_j with the probability $p(g_j|u)$, and so on.

B. Computing Moving Probability

From the learnt group-level mobility models, we extract the state matrix for each group,

$$\mathbf{g}_i^s = \begin{bmatrix} \pi_{i,1} & \pi_{i,2} & \dots & \pi_{i,H} \\ \widetilde{l}_{i,1} & \widetilde{l}_{i,2} & \dots & \widetilde{l}_{i,H} \end{bmatrix}, \quad 1 \le i \le |\mathcal{G}|.$$

The state matrix represents the steady-state distribution of the members belonging to a geo-community, and the corresponding core location of each state. The membership vector of group g_i is the probability of each user belonging to it, i.e., $\mathbf{g}_i^m = \{p(g_i|u)\}, u \in \mathcal{U}$. The number of members in each group is $N_{g_i} = \sum_{u\in\mathcal{U}} p_{u,i}$, where $p_{u,i} = p(g_i|u)$ denotes the probability of user u belonging to group g_i. Based on the above parameters of each group, we can determine the moving probability of their members as follows.

Group-level moving probability. The mobility model of a geo-community lies on an HMM with steady distribution being $\{\pi_{i,1}\, \pi_{i,2}\, \dots\, \pi_{i,H}\}$. Under the distance-based movements distribution in Eqn. (2), the probability of users starting from state $s_{i,h}$ to the promoted location L, i.e., $P(L|s_{i,h})$ decreases with the distance between L and the core location $l_{i,h}$, i.e., $||L - \widetilde{l}_{i,h}||_2$. Notably, if a user in state $s_{i,h}$ tends to move a larger distance than that between his location to the promoted location L (i.e., $||L - l_{i,h}||_2$), there is a higher chance that he is willing to move a distance equal to $||L - \widetilde{l}_{i,h}||_2$ and visit location L. Hence, the moving probability of users in state $s_{i,h}$ is equivalent to the probability that users moves a farther or a same distance of $||L - l_{i,h}||_2$, i.e., $P(\Delta l \ge ||L - l_{i,h}||_2)$. Thus the moving probability $P(L|s_{i,h})$ in state $s_{i,h}$ is given by

$$P(L|s_{i,h}) = \int_{||L-l_{i,h}||_2}^{\infty} \frac{\tau}{(x+\varepsilon)^{\tau+1}}\, dx = \frac{1}{(||L-l_{i,h}||_2 + \varepsilon)^{\tau}}.$$

Under the HMM based mobility model, the group-level moving probability defined in Definition 3 is determined as the expectation of moving probability in each steady state.

Definition 3. *(Group-level moving probability.) The moving probability of members in group g_i to the promoted location L can be formulated as $P_{g_i} = \sum_{h=1}^{H} \pi_{i,h} P(L|s_{i,h})$, where $P(L|s_{i,h})$ is the moving probability in state $s_{i,h}$.*

The group-level moving probability of a geo-community quantifies the possibility of its members moving to the promoted location after having been convinced by check-ins of their social neighbors. By modeling each geo-community as a group in the GSN, group-level location promotion focuses on selecting K seed groups to maximize the influence of location promotion among groups under the group-level moving probabilities as shown in Eqn. (1), whose solution is detailedly studied in the following section.

V. Group-level Location promotion

In this section, we investigate the solution for group-level location promotion problem in Eqn. (1). On the basis of the mined geo-communities, location promotion is elevated from individual level to group level by modeling geo-communities as groups in GSNs. However, the problem of selecting K seed groups is still NP-hard, as proved in Lemma 1.

Lemma 1. *Finding a set of K seed groups $\mathcal{S}_K \subseteq \mathcal{G}$ that has the maximum influence on the whole network (i.e., $\sum_{v\in\mathcal{G}} I(\mathcal{S}_K, v)N_v$) among all the subsets of $\mathcal{G}$ with size K is an NP-hard problem.*

Proof. We first consider an instance of the NP-hard set cover problem. Given a universal $U = \{x_1, x_2, \dots, x_n\}$ and a collection $\mathcal{C} = \{C_1, C_2, \dots, C_m\}$ of subsets of the universal U, the goal of the set over is to determine whether there is a cover $S \subset \mathcal{C}$ of size K whose union equals U. We then show that the group-level location promotion problem in Lemma 1 can be concluded as the set cover problem.

We construct a corresponding bipartite graph $\mathcal{T}$. The left part of $\mathcal{T}$ is m nodes representing the subsets in collection $\mathcal{C}$, and the right partition consists of the elements in universal U. Here, we set $n = |\mathcal{G}|$, and U represents the set of groups in G_{group} (i.e., $U = \mathcal{G}$). There exist an edge with weight 1 from C_j to x_i whenever $x_i \in C_j$. The set cover over $\mathcal{T}$ is equivalent to finding a set of K nodes in $\mathcal{T}$ who can finally influence $(K+n)$ nodes over the bipartite graph $\mathcal{T}$ under condition that both the moving probability and the size of each group equal 1. Since the set cover problem is NP-hard, the group-level location promotion problem in Lemma 1 is NP-hard. □

For effectively solving the above NP-hard problem, the seed selection in our proposed Group-level Location Promotion framework (GLP) is conducted over the group-level GSNs, which takes each group (geo-community) as a node in its corresponding graph. However, the formulation of the influences among groups (i.e., $I(\mathcal{S}, v)$ in Eqn. (1)) still remains to be derived. To tackle the issue, in the sequel, we first show how the influences are diffused over the group-level graph and then present the formulation of influences among groups.

A. Graph Structure in GLP

GLP is designed over a group-level social graph, in which each node represents a group of users. Different from that in individual-level graphs, the state of a node in influence diffusion cannot be simply defined as influenced or uninfluenced. We shall define the state of a node as the percent of its influenced members and the influence among nodes is also measured as percents. In location promotion, members of a group are both influenced by their group mates (intra-group influence) and members of other groups (inter-group influence). Specially, the intra-group influence depends on the closeness of g_i.

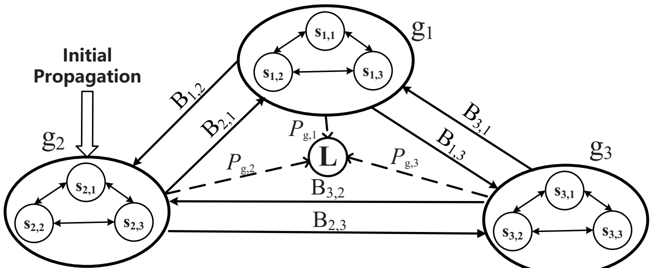


Fig. 4: The sketch of GLP

Definition 4. *The closeness of group g_i can be expressed as*

$$\rho_i = \frac{\sum_{m,n \in g_i} w_{mn} p_{n,i}}{N_{g_i}},$$

where $p_{m,i}$ is the probability that user m belongs to group g_i, N_{gi} is the number of members in g_i. w_{mn} is the individual influence of user m on user n under IC model.

The closeness ρ_i quantifies that given the newly influenced fraction of g_i's members in step s, what percentage of g_i's members can be expectedly influenced in step $(s+1)$. In addition, the inter-group influence, which is the aggregating influence of a group's members on others, depends on the weights of edges among groups as defined in Definition 5.

Definition 5. *The weight of the edge from group g_i to group g_j in group-level location promotion is*

$$B_{ij} = \frac{B\{g_i - g_j, g_j - g_i\} + B\{g_i \cap g_j\}}{N_{gj}},$$

where $g_i \cap g_j$ denotes the users concurrently belong to group g_i and g_j, $g_i - g_j$ denotes those who belong to group g_i while do not belong to group g_j, $B\{g_i - g_j, g_j - g_i\} = \sum_{m \in g_i - g_j, n \in g_j - gi} w_{mn} p_{n,j} P_{g_j}$, and $B\{g_i \cap g_j\} = \sum_{m \in g_i \cap g_j} p_{m,j}$.

In Definition 5, we consider the following two cases:
(1) $m \in g_i - g_j, n \in g_j - g_i$: The condition for user m successfully influencing user n is: m belonging to g_i makes a check-in record at L and the record convinces n through GSNs, then n belonging to g_j moves to location L after been convinced. Hence, in this case, the aggregating influence $B\{g_i - g_j, g_j - g_i\} = \sum_{m \in g_i - g_j, n \in g_j - gi} w_{mn} p_{n,j} P_{g_j} / N_{g_j}$.
(2) $m \in g_i \cap g_j$: In case that a user m simultaneously belonging to group g_i and g_j is influenced, then a $p_{m,j}/N_{gj}$ fraction of members in g_j is certainly influenced. Thus in this case, the aggregating influence is proportional to the number of members in both groups, i.e., $B\{g_i \cap g_j\} = \sum_{m \in g_i \cap g_j} p_{m,j} / N_{g_j}$.

Combining the above two cases, the influence of group g_i on g_j is equal to $B_{ij} = \frac{B\{g_i - g_j, g_j - g_i\} + B\{g_i \cap g_j\}}{N_{g_j}}$, which quantifies that given the influenced fraction of g_i's members, what percentage of g_j's members can be expectedly influenced in the diffusion from g_i to g_j.

In summary, we transfer the social graph from the individual level, i.e., $G = (V, E)$ to the group level, i.e., $G_{group} = (\mathcal{G}, \mathcal{B})$, where $\mathcal{G}$ denotes the set of all groups and $\mathcal{B}$ denotes the set of edges among them. The nodes (groups) in G_{group} have four key attributes: state matrix $\mathbf{g}^s$, membership vector $\mathbf{g}^m$, moving probability P_g and closeness ρ.

Over the group-level graph G_{group}, Figure 4 shows a mini technical example of group-level location promotion. There is a network of three nodes in Figure 4, each node represents a geo-community that has three steady states (i.e., $s_{i,s}, 1 \leq i \leq 3, 1 \leq s \leq 3$), and $P_{g_i} (1 \leq i \leq 3)$ denotes the group-level moving probability of group $g_i (1 \leq i \leq 3)$. For group g_1 and g_2, $B_{1,2}$ and $B_{2,1}$ respectively represent the weight of edge from g_1 to g_2 and that from g_2 to g_1. The goal of GLP in this mini example is to select a group who can maximize the influences among the three networks and one of its three core locations is selected for conducting initial promotion. In following context of the section, we will illustrate the formulation of influences among groups and then design the seed groups selection algorithm in the general cases.

B. Formulation of Group-level Influence

Now, we proceed to give the formulation of influences among groups over the group-level graph (i.e., $I(\mathcal{S}, v)$ in Eqn. (1)). We first show the case when $|\mathcal{S}| = 1$, and then extend it to general values.

1) Influence of a single seed group.: For a selected seed group, we first consider the intra-group diffusion among group mates after the initial propagation. Similar to the influence process under classical IC model, the intra-group influence is also characterized by steps. Due to the budget constraints, we assume that only one steady state is chosen for conducting initial propagation to a seed group. If group g_i is chosen as a seed group, the core location $\widetilde{l^*_{i,h}} = \mathrm{argmax}\ \pi_{i,h} \cdot P(L|s_{i,h})$ is chosen as the initial propagation location. Then the initial influenced fraction of g_i in step 0 can be given by

$$q_i = \pi^*_{i,h} P(L|s^*_{i,h}) p,$$

where p is the probability that initial propagation can convince a user. Recall the motivated example shown in Fig. 1, the cinema is selected as the site for the initial propagation, and thus, the $\pi^*_{i,h}$ and the $P(L|s^*_{i,h})$ respectively denote the steady-state probability of users around the cinema and the probability of users in this steady state moving to the promoted location. Since the cinema can consistently give influence, the fraction of users who are activated by the initial propagation is estimated as $q_i = \pi^*_{i,h} P(L|s^*_{i,h}) p$. Then in the subsequent steps of the IC model, the influence diffusion is via the check-ins shared over the GSNs where each user only has one single chance to influence their inactivated friends. The subsequent steps of the intra-group influence diffusion is concretely presented as follows.

In step 1, the fraction q_i of group members move to the promoted location L and make check-ins records there. Then these records spread through GSNs and activate the intra-group propagation, which leads to an expected fraction of $q_i \rho_i P_{gi}$ additional influenced members. Then in step 2, the fraction of $q_i \rho_i P_{gi}$ newly influenced members will further incur an additional fraction of $q_i (\rho_i P_{gi})^2$ influenced members in g_i. Accordingly, we obtain the final influenced fraction of the seed group g_i through intra-group propagation:

$$\phi^0_i = \sum_{n=0}^{\infty} q_i (\rho_i P_{g_i})^n = \frac{q_i}{1 - \rho_i P_{g_j}}. \qquad (10)$$

Now, we consider the influence of the seed group on the other groups. Since the seed group g_i has a fraction of ϕ^0_i

influenced members, the aggregating influence of this fraction of members on group g_j is $\phi_i^0 B_{ij}$. The fraction of $\phi_i^0 B_{ij}$ is the initial influenced fraction of group g_j, and such influenced members will activate intra-group propagation in group g_j with steps as well. Similar to Eqn. (10), the influence of group g_i on group g_j is given by

$$I(g_i, g_j) = \sum_{n=0}^{\infty} \phi_i^0 B_{ij} (\rho_j P_{g_j})^n = \frac{\phi_i^0 B_{ij}}{1 - \rho_j P_{g_j}}. \quad (11)$$

In GLP, the influence of the seed group on the whole network is the sum of its own influenced members, i.e., $\phi_i^0 N_{g_i}$ and the influence it has on all the other groups, i.e.,$\sum_{g_j \in \mathcal{G} \backslash g_i} I(g_i, g_j) N_{g_j}$. Hence, if group g_i is selected as the single seed group, its influence on the whole network is

$$\mathcal{M}(g_i) = \phi_i^0 N_{g_i} + \sum_{g_j \in \mathcal{G} \backslash g_i} I(g_i, g_j) N_{g_j}. \quad (12)$$

Thus when $K = 1$, the idea of selecting seed group is to find the group with the maximum influence on the whole network, i.e., : $s_1 = \mathrm{argmax}_{g_i \in \mathcal{G}} \mathcal{M}(g_i)$.

2) Influence of a set of $K (K > 1)$ seed groups.: We consider a set of K seed groups, i.e., $\mathcal{S}_K = \{s_1, s_2, \ldots, s_K\}$. For one of the groups s in $\mathcal{S}_K$, its influence on itself is $I(s, s) = \phi_s^0$, and the influence it has on one of the other groups v is $I(s, v) = \frac{\phi_s^0 B_{sv}}{1 - \rho_v P_v}$. To compute the influence of the seed set $\mathcal{S}_K$ on any group in $\mathcal{G}$, we need to consider the intersections of seed groups' influence on a same group. Thus the influence of seeds set $\mathcal{S}_K$ on one of the groups in $\mathcal{G}$ is given by

$$I(\mathcal{S}_K, v) = \sum_{i=1}^{K} (-1)^{i-1} \sum_{\substack{n1, \ldots, ni: \\ 1 \le n1 < \cdots < ni \le i}} \prod_{m=1}^{i} I(s_{nm}, v) \quad (13)$$

$$= 1 - \prod_{s \in \mathcal{S}_k} (1 - I(s, v)). \quad (14)$$

For Eqn. (14), taking $K = 2$ as an example, we have $I(\mathcal{S}_2, v) = I(s_1, v) + I(s_2, v) - I(s_1, v) I(s_2, v)$. Similar to Eqn. (12), the influence of S_K on the whole network is equal to $\mathcal{M}(S_K) = \sum_{g_j \in \mathcal{G}} I(S_K, g_j) N_{g_j}$. Thus the idea of selecting a seeds set with size K is to maximize the influence $\mathcal{M}(S_K)$, i.e., $\mathcal{S}_K = \mathrm{argmax}_{S \subseteq \mathcal{G}} \sum_{v \in \mathcal{G}} I(\mathcal{S}, v) N_v, |S| = K$, which is algorithmically elaborated as follows.

C. Algorithm of Seed Group Selection

1) Algorithm design. Based on the influences among groups as illustrated above, we proceed to present our group-level location promotion algorithm $\mathcal{GLP}$ whose pseudo code is shown in Algorithm 1. The input for $\mathcal{GLP}$ includes the learnt groups set $\mathcal{G}$, the state matrices $\mathbf{g}^s$, the transition probability matrices A and the membership vectors $\mathbf{g}^m$ of each group. Overall, $\mathcal{GLP}$ is composed of both offline and online phases. In the offline phase, $\mathcal{GLP}$ computes another two key attributes for each group to generate the group-level graph G_{group}, i.e., P_{g_i} and ρ_i, and the weight of edges among groups, i.e., B_{ij}. In the online phase, $\mathcal{GLP}$ greedily selects K seed groups through K iterations. In the $(k+1)$-th iteration, the selecting rule is maximizing the marginal influence under the having been selected seed groups $\mathcal{S}_k$, i.e., $s_{k+1} = \mathrm{argmax}_{g_i \in \mathcal{G} \backslash \mathcal{S}_k} \mathcal{M}(g_i | \mathcal{S}_k)$. At last, the output of $\mathcal{GLP}$ is the seed groups set with a size of K. In lines 17-19, the marginal influence of seed groups in each iteration is computed as follows.

Algorithm 1: $\mathcal{GLP}$

```
Input: group set $\mathcal{G}$, group state $\mathbf{g}_i^s$, transition probability
       $A_i$, promoted location $L$, group membership $\mathbf{g}_i^m$,
       individual graph $G = (V, E)$;
Output: $\mathcal{S}$: $K$ seed groups set
1  // Offline-Precomputing
2  for each $g_i$ in $\mathcal{G}$ do
3      Compute $P_{g_i}$ (Theorem 3) and $\rho_i$(Definition 4) ;
4  end
5  for each group couple $(g_i, g_j)$ in $\mathcal{G}$ do
6      Compute $B_{ij}$ for group couple $(g_i, g_j)$ (Definition 5);
7  end
8  // Online-Searching
9  Initialize seed set $\mathcal{S} = \Phi$;
10 for $k$ : 1 to $K$ do
11     for each $v \in \mathcal{G}$ do
12         $I(\mathcal{S}_k, v) = 1 - \prod_{s \in \mathcal{S}} (1 - I(s, v))$;
13     end
14     for each $g \in \mathcal{G} \backslash \mathcal{S}_k$ do
15         $\mathcal{M}(g|\mathcal{S}_k) = 0$;
16         for each $v \in \mathcal{G}$ do
17             $I(\mathcal{S}_k \cup g, v) = 1 - \prod_{s \in \mathcal{S}_k \cup g} (1 - I(s, v))$;
18             $I(g|\mathcal{S}_k, v) = I(\mathcal{S}_k \cup g, v) - I(\mathcal{S}_k, v)$;
19             $\mathcal{M}(g|\mathcal{S}_k) = \mathcal{M}(g|\mathcal{S}_k) + I(g|\mathcal{S}_k, v) N_v$;
20         end
21     end
22     $s_k = \mathrm{argmax}_{g \in \mathcal{G} \backslash \mathcal{S}_k} M(g|\mathcal{S}_k)$;
23     $\mathcal{S}_k = \mathcal{S}_k \cup s_k$;
24 end
25 return seed set $\mathcal{S}_K$.
```

Marginal influence. If group g_i is selected as the seed group in the $(k+1)$-th iteration, its marginal influence on any group v is $I(g_i | \mathcal{S}_k, v) = I(\mathcal{S}_k \cup g_i, v) - I(\mathcal{S}_k, v)$. For itself, the initial propagation can bring a fraction of ϕ_i^0 influenced members. Since there are a fraction of $I(\mathcal{S}_k, g_i)$ members that have been influenced by $\mathcal{S}_k$, its marginal influence on itself can be expressed as

$$\phi^0(g_i | \mathcal{S}_k) = \phi_i^0 - \phi_i^0 \cdot I(\mathcal{S}_k, g_i). \quad (15)$$

Now, we consider its marginal influence on other groups. Assuming g_i is added to the seed set, the influence of the new seed set $\mathcal{S}_k \cup g_i$ on one of the groups in $\mathcal{G}$ is given by

$$I(\mathcal{S}_k \cup g_i, v) = 1 - \prod_{s \in \mathcal{S}_k \cup g_i} (1 - I(s, v)). \quad (16)$$

Thus the marginal influence of g_i on any other group in $\mathcal{G}$ is equal to $I(g_i | \mathcal{S}_k, v) = I(\mathcal{S}_k \cup g_i, v) - I(\mathcal{S}_k, v)$. Similar to Eqn.(12), the marginal influence of g_i on the whole network can be given by

$$\mathcal{M}(g_i | \mathcal{S}_k) = \phi^0(g_i | \mathcal{S}_k) N_{g_i} + \sum_{g_j \in \mathcal{G} \backslash g_i} I(g_i | \mathcal{S}_k, g_j) N_{g_j}.$$

The idea of finding the $(k+1)$-th seed group is to select the group with maximum marginal influence on the whole network under seed set $\mathcal{S}_k$. Consequently, the $(k+1)$-th seed group is: $s_{k+1} = \mathrm{argmax}_{g_i \in \mathcal{G} \backslash \mathcal{S}_k} M(g_i | \mathcal{S}_k)$.

Remark. As shown in $\mathcal{GLP}$, we only consider the direct influence (single hop) of seed groups. The reason behind is that the uninfluenced fraction of each group dynamically and randomly changes in the influence diffusion process, which hinders us from accurately determining the influence of seed groups if multi-hop diffusion is considered. However, in the group-level location location promotion, we prove that the influence on a group through multi-hop diffusion can be approximated by the direct influence within a factor of $\left(1-\Theta\left(\frac{1}{\Delta l^{\tau}\log^2 n-1}\right)\right)$, as stated in Lemma 2. Here, Δl scales the distance from well-visited locations to the promoted location L, and n equals the number of individual users in G.

Lemma 2. *In the group-level location promotion, the direct influence of seed groups is a tight lower bound of the influence through multi-hop diffusion with an approximating ratio of more than* $\left(1-\Theta\left(\frac{1}{\Delta l^{\tau}\log^2 n-1}\right)\right)$.

Proof. In group-level location promotion, the influence of seed groups through multi-hop diffusion can be approximated by the direct influence. Without loss of generality, we take the first seed, i.e., $\mathcal{S}_1=\{g_i\}$ as an example. Based on Eqn.(11), the direct influence of g_i on other groups is $I(g_i,g_j)=\frac{\phi_i^0 B_{ij}}{1-\rho_j P_{g_j}}, g_j\in\mathcal{G}\backslash g_i$. Then the newly influenced members in these groups will attempt to influence inactive members in g_i through the second hop. The marginal influence of such groups on g_i through the second hop is given by

$$I(g_j|g_i,g_i)=\frac{I(g_i,g_j)(1-\phi_i^0)B_{ji}}{1-\rho_i P_{g_i}}.$$

We set $\mathcal{Z}=\mathcal{G}\backslash g_i$, based on Eqn.(13), we have

$$\begin{aligned}I(\mathcal{Z}|g_i,g_i)=&\sum_{h=1}^{|\mathcal{G}|-1}(-1)^{h-1}\sum_{\substack{n1,\ldots,nh:\\1\le n1<\cdots<nh\le h}}\prod_{m=1}^{h}I(z_{nm}|g_i,g_i)\\ \le&\sum_{g_j\in\mathcal{Z}}I(g_j|g_i,g_i)\le(|\mathcal{G}|-1)I(v|g_i,g_i),\end{aligned}$$

where $I(v|g_i,g_i)=\max_{g_j\in\mathcal{Z}}I(g_j|g_i,g_i)$. Since the direct influence on g_i equals ϕ_i^0, we have

$$\begin{aligned}&\frac{I(\mathcal{Z}|g_i,g_i)}{\phi_i^0}\le\frac{(|\mathcal{G}|-1)I(v|g_i,g_i)}{\phi_i^0}\\ =&\frac{(|\mathcal{G}|-1)(1-\phi_i^0)}{(1-\rho_v P_v)(1-\rho_i P_{g_i})}\cdot\left(B_{iv}B_{vi}-\frac{B^2\{g_i\cap g_j\}}{N_{g_i}N_v}\right).\end{aligned}\quad(17)$$

Since $|\mathcal{G}|\le\Theta(\log n)$ (n is the number of individual users and $|\mathcal{G}|=\Theta(\log n)$ only occurs in very large networks) [14], and we use B'_{iv} as the abbreviation of $B\{g_i-v,v-g_i\}$ and B''_{iv} as the abbreviation of $B\{g_i\cap v\}$, then Eqn.(17) becomes

$$\begin{aligned}\text{Eqn.(17)}=&\ \Theta(\log n)\cdot\frac{B'_{iv}B'_{vi}+B''_{iv}(B'_{iv}+B'_{vi})}{N_{g_i}N_v}\\ =&\frac{\Theta(\log n)\left\{\Theta(\frac{n}{\log^3 n\Delta l^{\tau}})^2+\Theta(\frac{n}{\log^2 n})\Theta(\frac{n}{\log^3 n\Delta l^{\tau}})\right\}}{\Theta(\frac{n^2}{\log^2 n})}\\ =&\Theta\left(\frac{1}{\Delta l^{\tau}\log^2 n}\right).\end{aligned}$$

For the influence on g_i through next $|\mathcal{G}|-2$ hops, since the inactive fraction of group g_i is less than $(1-\phi_i^0)$, the upper bound of the influence of group g_i on itself is expressed as

$$\begin{aligned}&I(g_i,g_i)\\ <&\phi_i^0+\phi_i^0\cdot\Theta\left((\Delta l^{\tau}\log^2 n)^{-1}+\cdots+(\Delta l^{\tau}\log^2 n)^{-(|\mathcal{G}|-1)}\right)\\ =&\phi_i^0+\phi_i^0\cdot\Theta\left(\frac{(\Delta l^{\tau}\log^2 n)^{-1}(1-(\Delta l^{\tau}\log^2 n)^{-(|\mathcal{G}|-1)})}{1-(\Delta l^{\tau}\log^2 n)^{-1}}\right)\\ <&\phi_i^0+\phi_i^0\cdot\Theta\left(\frac{1}{\Delta l^{\tau}\log^2 n-1}\right).\end{aligned}$$

Thus the upper bound of $I(g_i,g_i)$ is

$$I^U(g_i,g_i)=\phi_i^0+\phi_i^0\cdot\Theta\left(\frac{1}{\Delta l^{\tau}\log^2 n-1}\right).$$

As $\frac{I^U(g_i,g_i)-\phi_i^0}{\phi_i^0}=\Theta\left(\frac{1}{\Delta l^{\tau}\log^2 n-1}\right)\ll 1$, the influence through the next $|\mathcal{G}|-1$ hops is much smaller than direct influence. Therefore, the direct influence of seed groups is an tight lower bound of their influence through multi-hop with an approximating ratio more than $\left(1-\Theta\left(\frac{1}{\Delta l^{\tau}\log^2 n-1}\right)\right)$. □

2) Performance analysis. Based on the approximating ratio of influence estimation in Lemma 2, we prove that $\mathcal{GLP}$ can return a $\left(1-\Theta\left(\frac{1}{\Delta l^{\tau}\log^2 n-1}\right)\right)\left(1-\frac{1}{e}\right)$-approximate solution to the group-level location promotion problem. The fundament of performance guarantee lies on the monotonicity and submodularity of objective function as described below.

Lemma 3. *In GLP model, the influence function, i.e.,* $\mathcal{M}(\mathcal{S})=\sum_{v\in\mathcal{G}}I(\mathcal{S},v)N_v$ *is monotone and submodular.*

Proof. We use I_n as the abbreviation of $I(s_n,v)$, and the monotonicity and submodularity of the objective function are respectively proved as follows.

Monotonicity: Without loss of generality, for $\mathbb{A}=\{s_1,s_2,s_3\}$ and $\mathbb{B}=\{s_1,s_2\}$, $\mathbb{A}\subseteq\mathbb{B}$ and $N_v=1$, we have:

$$\begin{aligned}I(\mathbb{A},v)-I(\mathbb{B},v)&=I_3-I_2I_3-I_1I_3+I_1I_2I_3\\&=I_3(1-I_1)(1-I_2)>0.\end{aligned}$$

Hence, the influence function, i.e., $I(\mathbb{S},v)$ is monotone.

Submodularity: Without loss of generality, for $\mathbb{A}=\{s_1,s_2,s_3\}$, $\mathbb{B}=\{s_1,s_2\}$, $\mathbb{C}=\{s_1,s_3\}$ and $\mathbb{D}=\{s_1\}$, $\mathbb{D}\subseteq\mathbb{B}$, we have:

$$\begin{aligned}&(I(\mathbb{C},v)-I(\mathbb{D},v))-(I(\mathbb{A},v)-I(\mathbb{B},v))\\=&(I_3-I_1I_3)-(I_3-I_1I_3-I_2I_3+I_1I_2I_3)=I_3(I_2-I_1I_2)>0.\end{aligned}$$

Hence, the the influence function, i.e., $I(\mathcal{S},v)$ is submodular. Since $\mathcal{M}(\mathcal{S})=\sum_{v\in\mathcal{G}}I(\mathcal{S},v)N_v$, influence function $\mathcal{M}(\mathcal{S})$ is also monotone and submodular. □

With the monotonicity and submodularity of the objective function, referring the property proved by Nemhauser et al. [23], the exact greedy algorithm which iteratively selects the node with maximum marginal gain can achieve a $(1-\frac{1}{e})$-approximate solution to the IM problem. Then combining the approximating ratio of influence estimation stated in Lemma 2, we draw the following theorem for $\mathcal{GLP}$.

Theorem 1. *$\mathcal{GLP}$ returns a* $\left(1-\Theta\left(\frac{1}{\Delta l^{\tau}\log^2 n-1}\right)\right)\left(1-\frac{1}{e}\right)$*-approximate solution to the group-level location promotion problem in Eqn (1).*

Theorem 1 proves that $\mathcal{GLP}$ nearly achieves an approximating ratio of $(1-\frac{1}{e})$, which theoretically justifies its effectiveness. Next, we give the complexity of $\mathcal{GLP}$ to show the efficiency of group-level location promotion.

Lemma 4. (Complexity of $\mathcal{GLP}$.) *The computational complexity for $\mathcal{GLP}$ is $\mathcal{O}(|E|+|\mathcal{G}|+K^2(|\mathcal{G}|+|\mathcal{G}|^2))$.*

Proof. As shown in Algorithm 1, $\mathcal{GLP}$ is consisted of the online and offline phases, and we respectively give the complexity of the two phases as below.

In the offline precomputing phase, the social group on individual level $G=(V,E)$ is transferred to that on group level $G_{group} = (\mathcal{G}, B)$. Since the number of groups is $|\mathcal{G}|$, the process for calculating moving probability P_i of each group takes $\mathcal{O}(|\mathcal{G}|)$. From Definition 4 and Definition 5, we can see that closeness ρ_i and weights B_{ij} of all groups both correlate with the weight of edges among individual users w_{uv}. Thus ρ_i and B_{ij} of all groups can be calculated simultaneously by traversing all edges in E. This process costs $\mathcal{O}(|E|)$.

In the online searching phase, the K seed groups are selected through K iterations. In the $(k+1)$-th iteration, the selecting idea is to maximize the marginal influence under the selected groups set $\mathcal{S}_k$. We first compute the influence of $\mathcal{S}_k$ on each group and this process consists of $|\mathcal{G}|$ rounds. Since $|\mathcal{S}_k| = k$ in the $(k+1)$-th iteration, the polynomial in line 13 cost $\mathcal{O}(\frac{K(K-1)}{2})$. Then we compute the marginal influence of each group in $\mathcal{G}\backslash\mathcal{S}$ on the whole network, and this process takes $\mathcal{O}(|\mathcal{G}|^2\frac{K(K+1)}{2})$. Thus the computation complexity for online phase is $\mathcal{O}(K^2(|\mathcal{G}| + |\mathcal{G}|^2))$.

Therefore, combining the two phases, the computation complexity for $\mathcal{GLP}$ is $\mathcal{O}(|E| + |\mathcal{G}| + K^2(|\mathcal{G}| + |\mathcal{G}|^2))$. □

At the end of this section, we further prove that given the same budget, location promotion on group level can significantly improve the expected size of influenced users, which refers to the benefits gaining that will be discussed in details in the sequel.

D. Benefits Gaining on Group-level

As illustrated in Sections IV and V-C, there are two major advantages of GLP: (1) reliable moving probability; (2) reasonable computation complexity. Now we proceed to demonstrate the third advantage of GLP, i.e., benefits gaining. Here the benefits gaining means that given the same budget, location promotion on group level can expectedly influence much more users over the whole network compared with that on individual level. In the proposed GLP, the benefits gaining over the whole network is fundamentally brought by the gaining of the expected size of initial influenced users in each seed group, as given in Lemma 5.

Lemma 5. *The benefits gaining of group g_i can be expressed as $F_i = \frac{\pi^*_{i,h} N_{g_i} p b_i}{B_i}$, which is the expected gaining of the size of initial influenced users in seed group g_i. Here, B_i is the budget for initial propagation of group g_i and b_i is the budget for convincing an individual user.*

Proof. As described in section V-B, the core location $\widetilde{l}^*_{i,h} = \text{argmax}\ \pi_{i,h} \cdot P(L|\widetilde{l}_{i,h})$ is chosen as the initial propagation location for each seed group g_i with the corresponding budget being B_i. The budget B_i depends largely on the initial advertising methods, i.e., the cost for purchasing the advertising time of a cinema or the rent for billboards in a campus. Through initial propagation, the expected number of influenced users in g_i at step 0 is equal to $\pi^*_{i,h} N_{g_i} P_{g_i} p$, where $\pi_{i,h}$ is the steady-state probability of the chosen state, N_{g_i} and P_{g_i} is the size and moving probability of the group and p is the probability that initial propagation can convince a user. On individual level, let us assume the budget for convincing an individual user is b_i. If budget B_i for group g_i is used to convince users directly, the number of influenced seed users can be given by $\frac{B_i}{b_i} P_{g_i}$. Thus we obtain the benefits gaining for group g_i, i.e., $F_i = \frac{\pi^*_{i,h} N_{g_i} p b_i}{B_i}$. □

Based on the benefits gaining of a single seed group, Theorem 2 describes the gaining of K seed groups' influence on the whole network as follows.

Theorem 2. *In GLP, the benefits gaining of influence on the whole network scales as $\Theta\left(\frac{1-(1-I)^K}{1-(1-\frac{I}{F})^K}\right)$, where I approximates the influence among groups and F approximates the benefits gaining of a single seed group.*

Proof. Based on Eqn. (13), the influence of K seed groups on group v is $I(\mathcal{S}_K, v) = \sum_{i=1}^{K}(-1)^{i-1}\sum_{\substack{n1,\ldots,ni:\\ 1\leq n1<\cdots<ni\leq i}}\prod_{m=1}^{i} I(s_{nm}, v)$. If the same amount of budget are spent for convincing individuals directly, according to Lemma 5 and Eqn.(13), their influence on group v can be expressed as

$$I(\mathcal{S}_{in}, v) = \sum_{i=1}^{K}(-1)^{i-1}\sum_{\substack{n1,\ldots,ni:\\ 1\leq n1<\cdots<ni\leq i}}\prod_{m=1}^{i}\frac{1}{F_{nm}} I(s_{nm}, v).$$

Then the benefits gaining on group level is given by

$$\frac{I(\mathcal{S}_K, v)}{I(\mathcal{S}_{in}, v)} = \frac{\sum_{i=1}^{K}(-1)^{i-1}\sum_{\substack{n1,\ldots,ni:\\ 1\leq n1<\cdots<ni\leq i}}\prod_{m=1}^{i} I(s_{nm}, v)}{\sum_{i=1}^{K}(-1)^{i-1}\sum_{\substack{n1,\ldots,ni:\\ 1\leq n1<\cdots<ni\leq i}}\prod_{m=1}^{i}\frac{I(s_{nm}, v)}{F_{nm}}}$$

$$=\Theta\left(\frac{\sum_{i=1}^{K}(-1)^{i-1}\binom{K}{i}I^i}{\sum_{i=1}^{K}(-1)^{i-1}\binom{K}{i}\frac{I^i}{F^i}}\right) = \Theta\left(\frac{1-(1-I)^K}{1-(1-\frac{I}{F})^K}\right). \quad (18)$$

Since $M(\mathcal{S}_K) = \sum_{v\in\mathcal{G}} I(\mathcal{S}_K, v)N_v$, and $M(\mathcal{S}_{indv}) = \sum_{v\in\mathcal{G}} I(\mathcal{S}_{in}, v)N_v$, we have $\frac{M(\mathcal{S}_K)}{M(\mathcal{S}_{in})} = \Theta\left(\frac{1-(1-I)^K}{1-(1-\frac{I}{F})^K}\right)$. □

From Theorem 2, we further conclude the relationship between benefits gaining of GLP and the cost for seeding individual users as below:

(1) $\frac{I}{F} = \Theta(1)$: In this case, the budget for seeding a group can seed individual users with the size equals a $\Theta(1)$ fraction of group size. Thus we have $1-(1-I)^K = \Theta(1)$ and $1-(1-\frac{I}{F})^K = \Theta(1)$, then the right hand side of equal sign in Eqn.(18)=$\Theta(1)$, meaning the benefits gaining on group level in this case scales as $\Theta(1)$;

(2) $\frac{I}{F} < \Theta(1), K\cdot\frac{I}{F} \geq \Theta(1)$: The budget for seeding a group can only seed individual users with a size far smaller than group size, while the budget for K groups can seed a size equals a $\Theta(1)$ fraction of group size. Since $1-\exp(-\frac{KI}{F}) < 1-(1-\frac{I}{F})^K < 1-\exp(-\frac{2KI}{F})$ and $\frac{KI}{F} \geq \Theta(1)$, thus $1-(1-\frac{I}{F})^K = \Theta(1)$ and $1-(1-I)^K = \Theta(1)$, then the right hand side of equal sign in Eqn.(18)=$\Theta(1)$. Thus in this case the benefits gaining also scales as $\Theta(1)$.

(3) $\frac{I}{F} < \Theta(1), K\cdot\frac{I}{F} < \Theta(1)$: The budget for seeding K groups can only seed individual users with a size far smaller than group size. Since $x-\frac{x^2}{2} < 1-\exp(-x) < x(x>0)$, thus $\frac{KI}{F} - \frac{(\frac{KI}{F})^2}{2} < 1-\exp(-\frac{KI}{F}) < \frac{KI}{F}$, and $1-(1-\frac{I}{F})^K = \Theta(\frac{KI}{F})$, then:
   (a) $I \leq \Theta(1), KI \geq \Theta(1)$: $1-(1-I)^K = \Theta(1)$, then the right hand side of equal sign in Eqn.(18)=$\Theta(\frac{F}{KI})$;
   (b) $I < \Theta(1), KI < \Theta(1)$: $1-(1-I)^K = \Theta(KI)$, then the right hand side of equal sign in Eqn.(18)=$\Theta(F)$.

Thus in this case, group-level location promotion can largely improve the expected influence diffusion size with the benefits gaining larger than $\Theta(1)$.

The benefits gaining shown above implies that for the promotion mission in tough scenarios, where the cost for convincing an individual user is great, GLP can largely improve the expected number of influenced users under the same budget. While in easy scenarios where the promoted location is popular and users are easily to be influenced, GLP can also improve the influence as long as $F_i > 1$.

In summary, together with Theorem 1, Lemma 4 and Theorem 2, we theoretically show that our proposed framework GLP largely improves the performance of location promotion in terms of influence diffusion size and efficiency. In the next section, we evaluate the performance of GLP on real geo-social networks to justify our theoretical results.

VI. Experiments

In this section, we present the experimental results of a comprehensive performance study of GLP. All the experiments were implemented in Python and conducted on a computer running Ubuntu 16.04 LTS with 40 cores 2.30 GHz (Intel Xeon E5-2650) and 126 GB memory.

A. Experimental Datasets and Settings

1) Datasets. We evaluated GLP on three real location-based social networks, Brightkite, Gowalla and Foursquare. The datasets of Brightkite and Gowalla are downloaded from an open dataset website SNAP [2], and the dataset of Foursquare is downloaded from the check-ins visualization website Weeplaces [3]. In the experiments, we set $w_{mn} = \frac{|adj(m) \cap adj(n)|}{|adj(n)|}$, where $adj(m)$ is the set of friends of user m. The rationale behind is that the more co-friends of a pair of users, the closer their relationship. The statistics of the datasets are shown in Table II.

TABLE II: Statistics of Datasets

Datasets	Brightkite	Gowalla	Foursquare
# of Nodes	58,228	196,591	15,799
# of Edges	214,078	950,327	59,970
# of Check-ins	4,491,143	6,442,890	7,658,368

2) Settings. We first implement the iterative approach in Section IV-A to group users and learn group-level mobility models. Users in three GSNs (Brightkite, Gowalla and Foursquare) are all grouped into $|\mathcal{G}|$ ($|\mathcal{G}|$=30, 60, 100, 150) groups and each group has a HMM based mobility model. For Brightkite, we select the users whose daily activity area is near San Francisco (location: ($37.8°N$, $122.4°W$), population: 900 thousands). For Gowalla and Foursquare, we select San Jose (location:($37.3°N$, $121.8°W$), population: 950 thousands) and Los Angeles (location: ($34.1°N$, $118.3°W$), population: 3.9 millions) respectively. In the initialization, for each HMM model, the steady-state distribution and the transition probabilities are randomly initialized under the condition $\sum_{h=1}^{H} \pi_h = 1$ and $\sum_{j=1}^{H} a_{ij} = 1$, respectively. The core locations (i.e., $\widetilde{l}_h$) are randomly initialized by the coordinates around the location of the corresponding city for each dataset. Then the moving preference parameter τ is randomly sampled as a positive number. After obtaining mobility models and user groups, we then evaluate the performance of location promotion. To evaluate the expected influence diffusion size of selected groups / individuals, our method is simulating the diffusion process round by round. In each round, the newly influenced users in last round attempt to influence those who are uninfluenced. The evaluating metric for effectiveness is the number of final influenced users in the whole network, and for efficiency, is the running time of algorithms.

B. Performance Evaluation of GLP

1) Effectiveness study.: We define $N_{g_i}\pi_{i,k}$ as the steady number of users at each core location and the steady-state population distributions of three GSNs are shown in Figures 5, 6 and 7. The users in the three datasets are treated as samples of residents in the three cities, and we accordingly enlarge the size of each group to make the total number of users equal the three urban populations. Taking Brightkite, in which the users are from San Francisco, as an example, the sum of the members in all the groups is equal to $\sum_{g_i \in \mathcal{G}} N_{g_i}$, given the population in San Francisco is N, when plotting Figure 5, the size of the members in each group is enlarged by a multiple $\frac{N}{\sum_{g_i \in \mathcal{G}} N_{g_i}}$ to present the steady-state distribution of the residents in San Francisco.

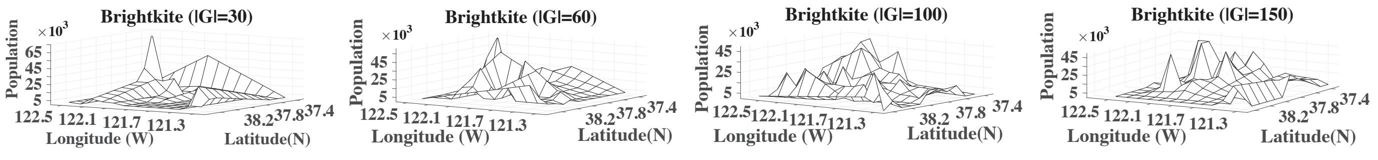
From Figures 5-7, it can be observed that when $|\mathcal{G}|$ is smaller, users are centralized at several discrete areas, and the steady distributions tend to be smooth with the increase of $|\mathcal{G}|$. This is because there are more geographical distinctions of the geo-communities when $|\mathcal{G}|$ is smaller, and the heterogeneous distribution of users results in the larger difference among the sizes of different groups. Based on the steady-state distributions, we select two locations with distinctly different population density for each dataset as the promoted locations in our experiments as shown in Figures 8-13. The promoted locations are:

(a) Brightkite: ($38.053°N$, $121.611°W$) and ($38.289°N$, $122.602°W$);
(b) Gowalla: ($38.039°N$, $122.379°W$) and ($37.604°N$, $122.047°W$);
(c) Foursquare: ($33.932°N$, $118.340°W$) and ($34.251°N$, $118.439°W$).

We compare GLP with three location-based influence maximization algorithms, i.e., Greedy [2], TPH [3] and EBA [1], and two group-level baseline algorithms, Largest and Nearest:

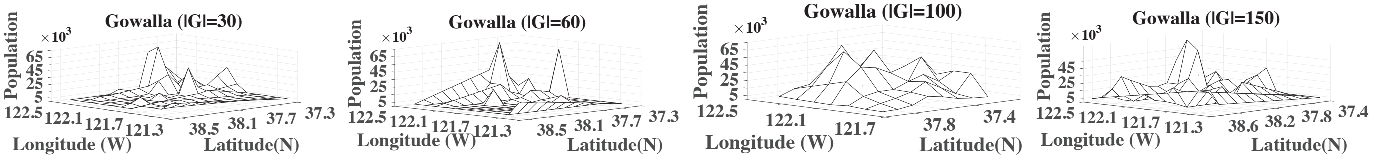
- Greedy [2]: Iteratively selecting K most influential individuals with maximum marginal influence in the location-based influence maximization, where the expected influences of the users are computed based on the location-aware IC model as described in Section III-A.
- TPH [3]: The TPH first computes a heuristic parameter $H_i = \sum_{j \in adj(i)} w_{ij} P_j^L$ for each user, and then selects K users with the largest H_i. Here $P_j^L = 1 - \prod_{c_i \in C}(1 - p_i)$, C is the set of user j's all historical locations and p_i is the moving probability to the promoted location from c_i.

[2] https://snap.stanford.edu/data/index.html
[3] http://www.yongliu.org/datasets/



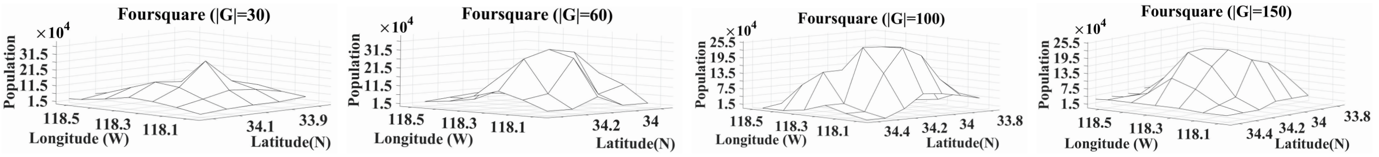
(a) $|\mathcal{G}|$=30 (b) $|\mathcal{G}|$=60 (c) $|\mathcal{G}|$=100 (d) $|\mathcal{G}|$=150

Fig. 5: Steady-state population distribution of Brightkite



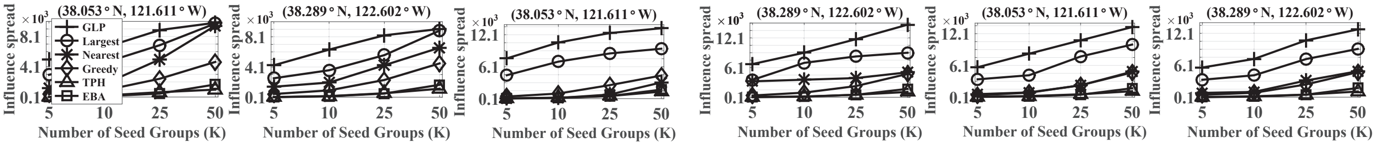
(a) $|\mathcal{G}|$=30 (b) $|\mathcal{G}|$=60 (c) $|\mathcal{G}|$=100 (d) $|\mathcal{G}|$=150

Fig. 6: Steady-state population distribution of Gowalla



(a) $|\mathcal{G}|$=30 (b) $|\mathcal{G}|$=60 (c) $|\mathcal{G}|$=100 (d) $|\mathcal{G}|$=150

Fig. 7: Steady-state population distribution of Foursquare



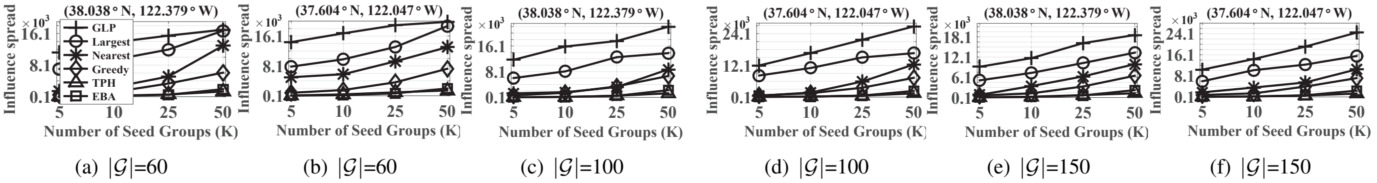
(a) $|\mathcal{G}|$=60 (b) $|\mathcal{G}|$=60 (c) $|\mathcal{G}|$=100 (d) $|\mathcal{G}|$=100 (e) $|\mathcal{G}|$=150 (f) $|\mathcal{G}|$=150

Fig. 8: Effectiveness study on Brightkite with K

- EBA [1]: The objective of the EBA [1] is maximizing the influence diffused to users who are in a given geographical range around the promoted location, and EBA [1] assumes each user is always in a fixed location. In the experiments section of [1], EBA selects a given number of users who are geographically nearest to a promoted location as the objective of influence diffusion. Given such selected objective users, EBA selects seed users through a greedy manner similar to the traditional IM.
- Largest: Select K groups having most members.
- Nearest: Select K groups whose daily activity areas are most close to the promoted location (highest P_{g_i}).

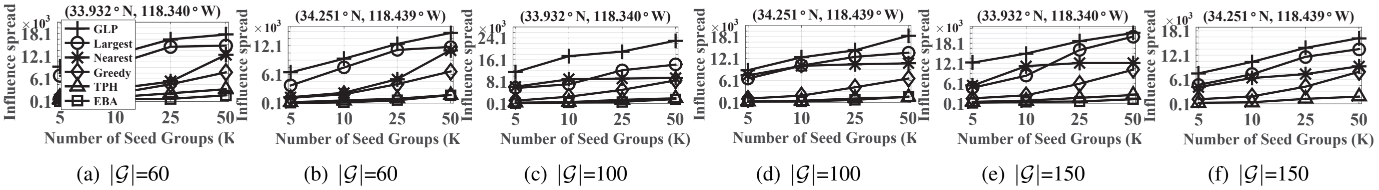
Note that the two individual baseline algorithms Greedy [2] and TPH [3] cannot be applied in the group-level promotion framework directly, since moving probabilities of individual users in the two frameworks are determined from individual mobility models. Thus, in the experiments, we compute the moving probability of each individual user as $p_u^L = \sum_{g \in \mathcal{G}} p(g|u) P_g^L$, where p_u^L denotes the moving probability of individual user u. In conducting EBA [1], similar to its original settings, we set the objective users as the top 10% users in each dataset who are geographically nearest to the promoted location. Then the number of seed individual users in the three individual-level location-aware IM algorithms are set to $\frac{B_i}{b_i} \cdot K$ (K is the number of seed groups).

Effect of the seeds set size K. For each selected seed group, the initial influenced fraction is set to $q_i = \pi^*_{i,h} P(L|\widetilde{l^*_{i,h}})p$ as described in Section V-B. We set $p = 0.05, 0.07, 0.1$ for $|\mathcal{G}| = 60, 100, 150$ respectively considering the size and centrality of groups. In Figures 8-10, we fix $\frac{B_i}{b_i} = 20$ to show the effect of K, and the effect of $\frac{B_i}{b_i}$ is investigated in Figures 11-13. Figures 8-10 plot the influence spread over three GSNs with $K = 5, 10, 25, 50$. As we can see, designing location promotion on group level can significantly improve the effectiveness compared with that on individual level as expected. Selecting the well-visited locations of the group members for initial propagation can achieve much more initial influenced users, which contributes to the improvements of influence spread. The increase of K, representing the increase of the initial propagation budget, can raise the final influence spread for all nine scenes. From Figures 8-10 , we find that improvements of influence spread are more significant when K is smaller, due to the submodularity of influence function. The influence spread curves of GLP, Largest, Greedy TPH and EBA all grow stably with increasing K, while the curves for Nearest have much more fluctuations. The reason behind is that some nearby groups may have small size, only when $K = 25, 50$, the seed set of Nearest may contain a few influential large groups. This insight demonstrates that only selecting several nearby locations for initial propagation may



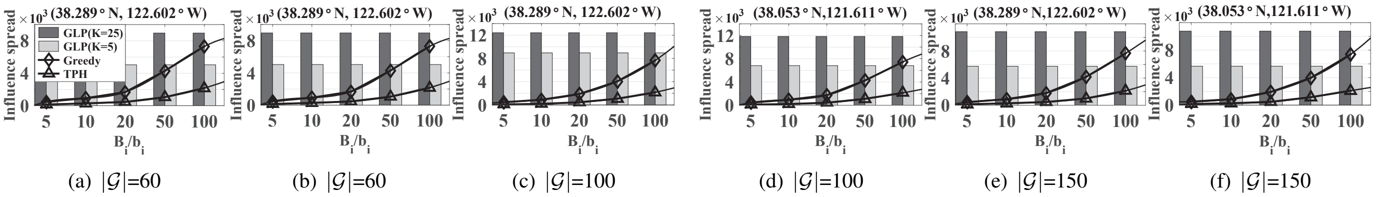
(a) $|\mathcal{G}|$=60 (b) $|\mathcal{G}|$=60 (c) $|\mathcal{G}|$=100 (d) $|\mathcal{G}|$=100 (e) $|\mathcal{G}|$=150 (f) $|\mathcal{G}|$=150

Fig. 9: Effectiveness study on Gowalla with K



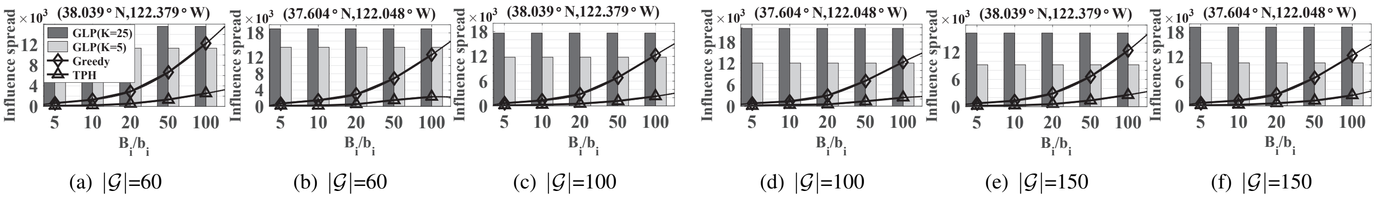
(a) $|\mathcal{G}|$=60 (b) $|\mathcal{G}|$=60 (c) $|\mathcal{G}|$=100 (d) $|\mathcal{G}|$=100 (e) $|\mathcal{G}|$=150 (f) $|\mathcal{G}|$=150

Fig. 10: Effectiveness study on Foursquare with K



(a) $|\mathcal{G}|$=60 (b) $|\mathcal{G}|$=60 (c) $|\mathcal{G}|$=100 (d) $|\mathcal{G}|$=100 (e) $|\mathcal{G}|$=150 (f) $|\mathcal{G}|$=150

Fig. 11: Effectiveness study on Brightkite with $\frac{B_i}{b_i}$



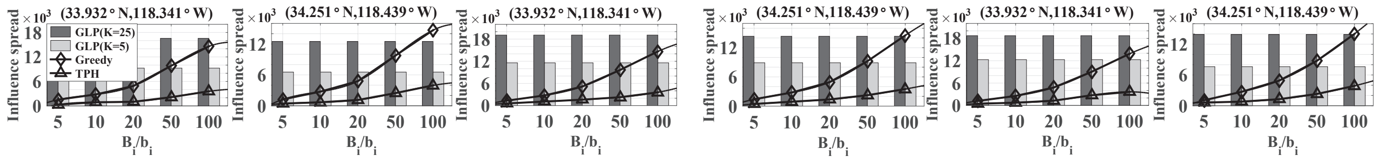
(a) $|\mathcal{G}|$=60 (b) $|\mathcal{G}|$=60 (c) $|\mathcal{G}|$=100 (d) $|\mathcal{G}|$=100 (e) $|\mathcal{G}|$=150 (f) $|\mathcal{G}|$=150

Fig. 12: Effectiveness study on Gowalla with $\frac{B_i}{b_i}$

not achieve good performance especially for the POIs not in bustling regions. Comparing with Largest, GLP is 50% larger in influence spread, this superiority arises from the consideration of both social ties and co-influences among groups in selection. The curves for TPH grow far more slowly than others due to the loss of consideration of co-influence of individual seed users in seed selection.

Effects of group size $|\mathcal{G}|$ and cost ratio $\frac{B_i}{b_i}$. From Figures 8-10, it can be found that the influence spread of $|\mathcal{G}| = 100$ is larger than that of $|\mathcal{G}| = 60$ and $|\mathcal{G}| = 150$. When $|\mathcal{G}| = 60$, there are several very large groups, and considering the low centrality of large groups, the propagation naturally decreases, resulting in a lower final influenced size compared with that of $|\mathcal{G}| = 100$. When $|\mathcal{G}| = 150$, the fraction of initial influenced users is lower than that of $|\mathcal{G}| = 100$ because the number of seed groups is the same for the two cases. Such results show the effect of group numbers on GLP. Then we evaluate the influence of the cost for conducting initial propagation. Figures 11-13 show the influence spread with $\frac{B_i}{b_i}$=100, 50, 20, 10, 5, respectively, and K is fixed as 25. $\frac{B_i}{b_i}$ is the ratio of the cost for seeding a group to the cost for seeding an individual user. For example, $\frac{B_i}{b_i} = 100$ means the number of seed users in Greedy and TPH is the 100 times that of seed groups in GLP. As expected, GLP can significantly improve the performance of location promotion especially in tough scenarios, where the cost for convincing an individual seed user is extremely expensive. From Figures 11-13, it can also be observed that, with the increase of $\frac{B_i}{b_i}$, influence spread on individual-level grows as well and the relative gain of GLP decreases since the same budget can convince more individual seed users. Even in case $\frac{B_i}{b_i}$=100, the influence spread of GLP is still 30% more than that on individual level.

2) Efficiency Study.: In the efficiency study, we present the running time of the seed selection in GLP and two individual-level baselines Greedy [2] and TPH [3]. The seed selection in the framework GLP is specifically conducted by the algorithm $\mathcal{GLP}$. Since users' social behavior patterns usually remain stable within a relatively long interval [19], learning group-level mobility models and users grouping can be pre-conducted offline. Thus we compare the running time of the selecting algorithm $\mathcal{GLP}$ with the seed selection time in both Greedy and TPH to demonstrate the efficiency for supporting real-time applications. Figure 14 shows the running time of the seed selection in $\mathcal{GLP}$, Greedy [2] and TPH [3] with $K = 5, 10, 25, 50$, respectively. From Figure 14, it can be observed that the running time of $\mathcal{GLP}$ is as long as about 1% of the two individual algorithms. The Greedy needs to sample the whole network to estimate the marginal influence of all remaining individuals in each iteration, which is a time consuming task. For the TPH, it needs to recompute moving probability to new promoted locations for each individual user



(a) $|\mathcal{G}|$=60 (b) $|\mathcal{G}|$=60 (c) $|\mathcal{G}|$=100 (d) $|\mathcal{G}|$=100 (e) $|\mathcal{G}|$=150 (f) $|\mathcal{G}|$=150

Fig. 13: Effectiveness study on Foursquare with $\frac{B_i}{b_i}$

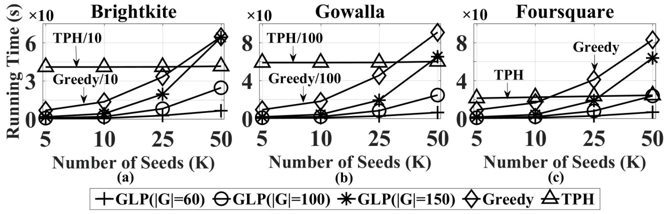


Fig. 14: Efficiency study

and thus all heuristic parameters H has to be recomputed. Since TPH considers all historical check-ins records of each user, the recomputing process brings high time complexity. While $\mathcal{GLP}$ is conducted on group-level graphs, and the number of nodes in which is far smaller than that of individual-level graphs. Meanwhile, the running time of $\mathcal{GLP}$ increases superlinearly with K, due to the fact that the time complexity of $\mathcal{GLP}$ is quadratic with K as shown in Lemma 4.

VII. Conclusion

In this paper, we propose GLP, a novel framework for group-level location promotion. Our insight is that users belonging to a same geo-community may have a few well visited locations and conduct initial propagation at such locations can influence much more seed users. We design an iterative learning approach to group them and extract common mobility models from massive check-ins data. Then GLP generates a new group-level graph considering social relationship and group-level moving probability of each group, and a greedy algorithm is proposed to effectively select K seed groups over the graph. Furthermore, we theoretically prove that GLP can largely improve the influence spreading under the same budget. Finally, the extensive experiments on real datasets justify the high performance of GLP in location promotion.

Acknowledgement

This work was supported by National Key R&D Program of China 2018YFB1004705, NSF China under Grant 61822206, Grant 61532012, Grant 61602303, Grant 61829201, CCF-Tencent RAGR 20180116, and Open Research Fund of Tianjin Key Laboratory of Advanced Networking (TANK).

References

[1] G. Li, S. Chen, J. Feng, K. Tan, and W. Li. Efficient location-aware influence maximization. In *Proc. SIGMOD*, pages 87–98. ACM, 2014.

[2] X. Wang, Y. Zhang, W. Zhang, and X. Lin. Distance-aware influence maximization in geo-social network. In *Proc. ICDE*. IEEE, 2016.

[3] T. Zhou, J. Cao, B. Liu, S. Xu, Z. Zhu, and J. Luo. Location-based influence maximization in social networks. In *Proc. CIKM*, pages 1211–1220. ACM, 2015.

[4] J. Yuan and S. Tang. Adaptive discount allocation in social networks. In *Proc. MobiHoc*. ACM, 2017.

[5] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proc. SIGKDD*, pages 137–146. ACM, 2003.

[6] G. Tong, W. Wu, S. Tang, and D. Du. Adaptive influence maximization in dynamic social networks. *IEEE/ACM Transactions on Networking (TON)*, 25(1):112–125, 2017.

[7] H. Zhang, D. Nguyen, H. Zhang, and M. Thai. Least cost influence maximization across multiple social networks. *IEEE/ACM Transactions on Networking (TON)*, 24(2):929–939, 2016.

[8] X. Li, J. Smith, T. Dinh, and M. Thai. Why approximate when you can get the exact? optimal targeted viral marketing at scale. In *Proc. INFOCOM*. IEEE, 2017.

[9] Y. Wang, G. Cong, G. Song, and K. Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proc. SIGKDD*, pages 1039–1048. ACM, 2010.

[10] W. Zhu, W. Peng, L. Chen, K. Zheng, and X. Zhou. Modeling user mobility for location promotion in location-based social networks. In *Proc. SIGKDD*, pages 1573–1582. ACM, 2015.

[11] J. Toole, C. Herrera-Yaqüe, C. Schneider, and M. González. Coupling human mobility and social ties. *Journal of The Royal Society Interface*, 12(105):20141128, 2015.

[12] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences (PNAS)*, 107(52):22436–22441, 2010.

[13] P. Deville, C. Song, N. Eagle, V. Blondel, A. Barabási, and D. Wang. Scaling identity connects human mobility and social interactions. *Proceedings of the National Academy of Sciences (PNAS)*, page 201525443, 2016.

[14] M. Eftekhar, Y. Ganjali, and N. Koudas. Information cascade at group scale. In *Proc. SIGKDD*, pages 401–409. ACM, 2013.

[15] Brad Goodman. Commercials in movie theaters. https://www.videouniversity.com/articles/commercials-in-movie-theaters/, 2011.

[16] Gaebler Ventures. Business advertising. http://www.gaebler.com/Business-Advertising.htm, 2017.

[17] Z. Lu, Y. Wen, W. Zhang, Q. Zheng, and G. Cao. Towards information diffusion in mobile social networks. *IEEE Transactions on Mobile Computing*, 15(5):1292–1304, 2016.

[18] M. Xiao, J. Wu, and L. Huang. Community-aware opportunistic routing in mobile social networks. *IEEE Transactions on Computers*, 63(7):1682–1695, 2014.

[19] J. Fan, J. Chen, Y. Du, W. Gao, J. Wu, and Y. Sun. Geocommunity-based broadcasting for data dissemination in mobile social networks. *IEEE Transactions on Parallel and Distributed Systems*, 24(4):734–743, 2013.

[20] E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proc. SIGKDD*, pages 1082–1090. ACM, 2011.

[21] C. Zhang, K. Zhang, Q. Yuan, L. Zhang, T. Hanratty, and J. Han. Gmove: Group-level mobility modeling using geo-tagged social media. In *Proc. SIGKDD*, pages 1305–1314. ACM, 2016.

[22] W. Mathew, R. Raposo, and B. Martins. Predicting future locations with hidden markov models. In *Proc. Ubicomp*, pages 911–918. ACM, 2012.

[23] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

Appendix A
Derivations for EM algorithm in Section IV

1. Derivations for objective function $Q(\Theta|\Theta^t)$. Here, we take the trajectories as observing values and user states as the hidden variables. For the r-th trajectory, the expected value of log likelihood is $Q_r(\Theta|\Theta^t) = \mathbb{E}_{D_r|l_r,\Theta^t} LL(\Theta|D_r, l_r)$. Since $LL(\Theta|D_r, l_r) = \ln P(D_r, l_r|\Theta)$, we have

$$Q_r(\Theta|\Theta^t) = \sum_{D_r} P(D_r|l_r, \Theta^t) \ln P(D_r, l_r|\Theta). \tag{19}$$

In training HMM for group g, different trajectories have different weights because the users of them have different probabilities belonging to group g. Hence, the mixture likelihood is expressed as

$$Q(\Theta|\Theta^t) = \sum_{i=1}^{R} \sum_{D_r} w_r P(D_r|l_r, \Theta^t) \ln P(D_r, l_r|\Theta). \tag{20}$$

Since $P(D_r, l_r|\Theta)$ represents the joint distribution of a hidden Markov process, we have

$$P(D_r, l_r|\Theta) = P(l_{r,1}, s_{r,1}, l_{r,2}, s_{r,2}, \ldots, l_{r,N}, s_{r,N})$$

$$= p(s_{r,1}) p(l_{r,1}|s_{r,1}) \prod_{n=2}^{N} p(s_{r,n}|s_{r,n-1}) p(l_{r,n}|s_{r,n}). \tag{21}$$

Taking Eqn. (21) into Eqn. (20), we obtain the objective function $Q(\Theta|\Theta^t)$.

2. Formulations for α, β, ξ and γ. $\alpha(s_{r,n}) = P(l_{r,1}, l_{r,2}, \ldots, l_{r,n}, s_{r,n}|\Theta^t)$ is the probability distribution of the n-th state with the observing sequence $\{l_{r,1}, l_{r,2}, \ldots, l_{r,n}\}$, and can be calculated by the forward algorithm. Specially,

$$\alpha(s_{r,n} = h) = P(l_{r,1}, l_{r,2}, \ldots, l_{r,n}, s_{r,n} = h|\Theta^t)$$

$$= p(l_{r,n}|s_{r,n} = h) \sum_{i=1}^{H} \alpha(s_{r,n-1} = i) p(s_{r,n} = h|s_{r,n-1} = i),$$

where the initial value is $\alpha(s_{r,1} = h) = \pi_h p(l_{r,1}|s_{r,1} = h)$; $\beta(s_{r,n}) = P(l_{r,n+1}, l_{r,n+2}, \ldots, l_{r,N}|s_{r,n}, \Theta^t)$ is the probability of the observable sequence $\{l_{r,n+1}, l_{r,n+2}, \ldots, l_{r,N}\}$ conditioned on the n-th state and can be calculated via the backward algorithm. Hence, we have

$$\beta(s_{r,n} = h) = P(l_{r,n+1}, l_{r,n+2}, \ldots, l_{r,N}|s_{r,n} = h, \Theta^t)$$

$$= \sum_{i=1}^{H} \beta(s_{r,n+1} = i) p(l_{r,n+1}|s_{r,n+1} = i) p(s_{r,n+1} = i|s_{r,n} = h),$$

where the initial value is $\beta(s_{r,N} = h) = 1$. Based on $\alpha(s_{r,n})$ and $\beta(s_{r,n})$, $\gamma(s_{r,n}) = p(s_{r,n}|l_r, \Theta^t)$ is the probability distribution of the n-th latent state, and $\xi(s_{r,n-1}, s_{r,n}) = p(s_{r,n-1}, s_{r,n}|l_r, \Theta^t)$ is the joint distribution of two consecutive latent states $p(s_{r,n-1}, s_{r,n}|l_r, \Theta^t)$. Specifically,

$$\gamma(s_{r,n} = h) = \frac{\alpha(s_{r,n} = h)\beta(s_{r,n} = h)}{\sum_{i=1}^{H} \alpha(s_{r,n} = i)\beta(s_{r,n} = i)},$$

$$\xi(s_{r,n-1} = i, s_{r,n} = j)$$

$$= \frac{\alpha(s_{r,n-1} = i) p(l_{r,n}|s_{r,n} = j) a_{ij} \beta(s_{r,n} = j)}{\sum_{i=1}^{H} \sum_{j=1}^{H} \alpha(s_{r,n-1} = i) p(l_{r,n}|s_{r,n} = j) a_{ij} \beta(s_{r,n} = j)}.$$

3. Derivations for updating rules in M-step.
1) From Eqn. (3), the updating object for π_h is to maximize

$$Q(\pi) = \sum_{r=1}^{R} \sum_{D_r} \sum_{h=1}^{H} w_r P(D_r|l_r, \Theta^t) \delta(s_{r,1}, h) \ln p(s_{r,1})$$

$$= \sum_{r=1}^{R} \sum_{h=1}^{H} w_r \gamma(s_{r,1} = h) \ln \pi_h,$$

with the constraining condition $\sum_{h=1}^{H} \pi_h = 1$, we obtain the updating rule in Eqn. (6) with Lagrange multiplier algorithm.
2) From Eqn. (4), the updating object for a_{ij} is to maximize

$$Q(a_{ij}) = \sum_{r=1}^{R} \sum_{n=2}^{N} \sum_{i=1}^{H} \sum_{j=1}^{H} w_r P(D_r|l_r, \Theta^t) \delta(s_{r,n-1}, i) \delta(s_{r,n}, j) \ln a_{ij}$$

$$= \sum_{r=1}^{R} \sum_{n=2}^{N} \sum_{i=1}^{H} \sum_{j=1}^{H} w_r \xi(s_{r,n-1} = i, s_{r,n} = j) \ln a_{ij}.$$

Since $\sum_{j=1}^{H} a_{ij} = 1$, we obtain the updating rule in Eqn. (7) based on the Lagrange multiplier algorithm.
3) When the user of l_r is in state h, the probability that she is at location $l_{r,n}$ equals $f(l_{r,n}|s_{r,n} = h) = \frac{\tau}{(||l_{r,n} - \widetilde{l}_h||_2 + \varepsilon)^{\tau+1}}$. Thus from Eqn.(5), the updating object for $\widetilde{l}_h$ is to maximize

$$Q(\ \widetilde{l}_h\ ) = -\sum_{r=1}^{R} \sum_{n=1}^{N} \sum_{h=1}^{H} w_r \gamma(s_{r,n} = h) \ln ||l_{r,n} - \widetilde{l}_h||_2,$$

and we obtain the updating rule in Eqn.(8).
4) From Eqn.(5), the updating objective for τ is to maximize

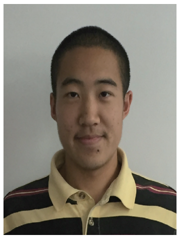
$$Q(\tau) = \sum_{r=1}^{R} \sum_{n=1}^{N} \sum_{h=1}^{H} w_r \gamma(s_{r,n} = h) \left\{ \ln \tau - \tau \ln(||l_{r,n} - \widetilde{l}||_2 + \varepsilon) \right\},$$

and we obtain the updating rule in Eqn.(9).

Xudong Wu received his B. E. degree in Information and Communication Engineering from Nanjing Institute of Technology, China, in 2015. His is currently pursuing his Ph.D. degree in Department of Computer Science and Engineering in Shanghai Jiao Tong University. His research of interests are in the area of social networking and big data, machine learning and combinatorial optimization.

Luoyi Fu received her B. E. degree in Electronic Engineering from Shanghai Jiao Tong University, China, in 2009 and Ph.D. degree in Computer Science and Engineering in the same university in 2015. She is currently an Assistant Professor in Department of Computer Science and Engineering in Shanghai Jiao Tong University. Her research of interests are in the area of social networking and big data, scaling laws analysis in wireless networks, connectivity analysis and random graphs. She has been a member of the Technical Program Committees of several conferences including ACM MobiHoc 2018-2019, IEEE INFOCOM 2018-2019.

Yuhang Yao is currently pursuing the bachelor's degree with the Department of Computer Science, Shanghai Jiao Tong University, Shanghai, China. He is currently a Research Intern supervised by Prof. Xinbing Wang. His research interests include privacy protection in blockchain and asymptotic analysis in social networks and scholar networks.



Xinzhe Fu received his B. E. degree in Department of Computer Science and Engineering at Shanghai Jiao Tong University, China, 2017. During his undergraduate study, he was working as an research intern supervised by Dr. Luoyi Fu. His research interests include combinatorial optimization, asymptotic analysis and privacy protection in social networks. He is pursuing Ph. D. degree in the Massachusetts Institute of Technology (MIT), Massachusetts, USA.

Xinbing Wang received the B.S. degree (with hons.) from the Department of Automation, Shanghai Jiaotong University, Shanghai, China, in 1998, and the M.S. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2001. He received the Ph.D. degree, major in the Department of electrical and Computer Engineering, minor in the Department of Mathematics, North Carolina State University, Raleigh, in 2006. Currently, he is a professor in the Department of Electronic Engineering, Shanghai Jiaotong University, Shanghai, China. Dr. Wang has been an associate editor for IEEE/ACM Transactions on Networking and IEEE Transactions on Mobile Computing, and the member of the Technical Program Committees of several conferences including ACM MobiCom 2012, 2018-2019, ACM MobiHoc 2012-2014, IEEE INFOCOM 2009-2017.

Guihai Chen received the B.S. degree from Nanjing University, the M.E. degree from Southeast University, and the Ph.D. degree from The University of Hong Kong. He visited the Kyushu Institute of Technology, Japan, in 1998, as a Research Fellow, and the University of Queensland, Australia, in 2000, as a Visiting Professor. From 2001 to 2003, he was a Visiting Professor with Wayne State University. He is currently a Distinguished Professor and a Deputy Chair with the Department of Computer Science, Shanghai Jiao Tong University. He has published over 200 papers in peer-reviewed journals and refereed conference proceedings in the areas of wireless sensor networks, high-performance computer architecture, peer-to-peer computing, and performance evaluation. He is a member of the IEEE Computer Society. He has served on technical program committees of numerous international conferences.