Unsupervised Interesting Places Discovery in Location-Based Social Sensing

Chao Huang, Dong Wang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
chuang7@nd.edu, dwang5@nd.edu

***Abstract*—This paper presents an unsupervised approach to accurately discover interesting places in a city from location-based social sensing applications, a new sensing application paradigm that collects observations of physical world from Location-based Social Networks (LBSN). While there are a large amount of prior works on personalized Point of Interests (POI) recommendation systems, they used supervised learning approaches that did not work for users who have little or no historic (training) data. In this paper, we focused on an *interesting place discovery* problem where the goal is to accurately discover the interesting places in a city that *average people* may have strong interests to visit (e.g., parks, museums, historic sites, etc.) using *unsupervised* approaches. In particular, we develop a new *Physical-Social-aware Interesting Place Discovery (PSIPD)* scheme which jointly exploits the location's physical dependency and the visitor's social dependency to solve the interesting place discovery problem using an *unsupervised approach*. We compare our solution with state-of-the-art baselines using two real world data traces from LBSN. The results showed that our approach achieved significant performance improvements compared to all baselines in terms of both estimation accuracy and ranking performance.**

***Keywords*-Social Sensing, Interesting Place Discovery, Unsupervised Learning, Physical Dependency, Social Dependency**

I. Introduction

This paper develops an unsupervised approach to accurately discover interesting places in a city from location-based social sensing applications. This work is motivated by the emergence of social sensing as a new application paradigm of collecting sensory measurements from common individuals with smart sensing devices (e.g., smartphones) [11], [12], [29]. This trend becomes more prevalent with the advent of online social media that allows the crowd to distribute their measurements in a timely and scalable way [2], [30], [33], [34]. For example, using Location-Based Social Network (LBSN) services (e.g., Foursquare, Google Place, Gowalla, etc.), people can now easily upload the "check-in" points or GPS traces from their phones to report the places they visit in a city. Alternatively, a group of citizens who care about the appearance of their neighborhood may download a geotagging app to take pictures of litter locations and share them with the community. While there are a large amount of prior works on personalized Point of Interests (POI) recommendation systems, they used *supervised learning* approaches that did not work for users who have little or no historic (training) data [9], [16], [18], [31], [41], [42]. In this paper, we focus on an *interesting place discovery* problem in social sensing applications where the goal is to accurately discover the interesting places in a city where *average people* may have strong interests to visit (e.g., parks, museums, historic sites, etc.) using *unsupervised* approaches.

Significant efforts have been made to solve the interesting place discovery problem using the *crowdsourcing* methods [3], [10], [15], [26], [38], [44]. The basic principle of those solutions is to estimate the locations of interesting places by analyzing the GPS traces from a large crowd in a given area (e.g., city) [5]. The crowdsourcing methods have a few clear advantages compared to the traditional methods (e.g., travel websites or search services) [28], [32]. *First*, crowdsorucing is cost efficient since the crowd often volunteer to share their location data through the services they use (e.g., LBSN) [20]. *Second*, the interestingness of a place may change over time and the crowdsourcing methods can track such changes by analyzing the most recent trajectory data uploaded by the crowd [39]. *Third*, the crowdsourcing traces normally have a better spatial-temporal coverage of the interesting places as the crowd are naturally distributed across the region [26].

However, several key limitations exist in the current interesting place discovery solutions using crowdsourcing methods. *First*, a large category of existing solutions developed heuristic-based models that assume *linear* relationship between the user's travel experience [1] and the number of places he/she visited [44]. Such assumption does not hold in scenarios where the relationship between a user's travel experience and the number of places he/she visited is *nonlinear* [27]. *Second*, physical dependency often exists between places that are close to each other. For example, users who visit the aquarium in Chicago may also choose to visit the planetarium a few hundred meters away. However, they may not upload two separate check-in points at the two nearby places. The current schemes often ignored such physical dependency and were shown to generate many

[1]The travel experience was shown to directly affect the user's ability to find interesting places [15]

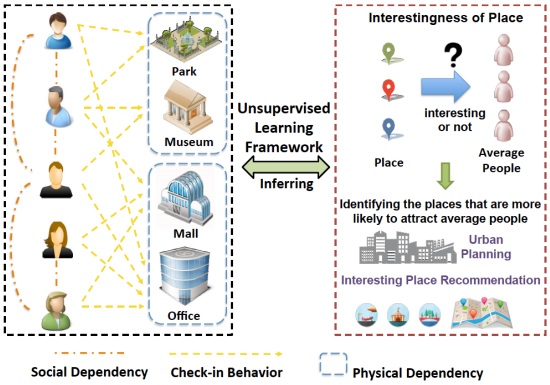


Figure 1. The Overview of PSIPD Framework

false negatives [10]. *Third*, the social dependency between users also affects their visiting behavior and the interesting place discovery results. Unfortunately, the current techniques either ignored the impact of user's social dependency or consider it using heuristic approaches, which generated a large number of false positives [26].

In sharp contrast to our previous work in [10], this paper develops a Physical-Social-aware Interesting Place Discovery (PSIPD) scheme that addresses the above limitations by explicitly exploiting both the *physical dependency* between places and the *social dependency* between users using an *unsupervised approach*. The overview of the PSIPD framework is shown in Figure 1. In particular, a maximum likelihood estimation (MLE) approach is developed to jointly estimate both the user's travel experience and the interestingness of a place. The MLE approach considers both the user's visiting behavior and the physical-social dependency information embedded in the crowdsourcing data under a rigorous analytical framework. We evaluate the PSIPD scheme using two real world data traces collected from LBSNs (i.e., Brightkite and Gowalla). The results showed that our approach achieved significant performance improvements compared to the sate-of-the-art baselines in terms of both estimation accuracy and ranking performance. The results of this paper are important because they allow crowdsourcing applications to accurately discover interesting places by explicitly considering the dependency in both physical and social spaces using a principled unsupervised approach. To summarize, the contributions of this work are as follows:

- We propose an *unsupervised* approach to solve the interesting place discovery problem by jointly exploiting both the *physical dependency* between places and the *social dependency* between users.
- We develop a new analytical framework that allows us to derive an *optimal solution* (in the sense of maximum likelihood estimation) for the physical-social-aware interesting place discovery problem.
- Our MLE solution explicitly handles the *nonlinear relationship* between the user's travel experience and the interestingness of places.
- The PSIPD scheme achieves *non-trivial performance gains* compared to the state-of-the-art baselines on real world data traces.

II. Related Work

Advances in location-acquisition on mobile devices and wireless communication technologies have enabled the Location-Based Social Networking (LBSN) services [22]–[24]. Recent research works start to address interesting challenges in LBSN such as user mobility modeling [4], [19], [45], semantic analysis [17], [36] and user relationship study [8], [40]. An emerging problem of *interesting place discovery* arises in LBSN due to the proliferation of location-based crowdsourcing applications (e.g., Foursquare, Google Places, Gowalla) [1], [5]. These applications empower common individuals to easily share their location information (e.g., check-in points) with other people almost anywhere and anytime. To address this emerging problem, we develop a physical-social-aware interesting place discovery scheme to accurately identify interesting places in a city by explicitly exploiting both the *physical dependency* between places and *social dependency* between users under a *rigorous* analytical framework. The proposed framework accurately discovers the interesting places from massive check-in points contributed by the crowd.

Significant amount of work has been done in recommending Points of Interests (POI) in data mining and geographic information systems [9], [16], [18], [41], [42]. For example, Zhang et al. [41] developed a kernel density estimation method to infer the POI based on the observation that the geographical proximity significantly affects user check-in behaviors. They further integrated their model with social connections between users and category information of places [42]. Hu et al. [9] proposed a comprehensive model that explicitly considered the geographical influence and temporal activity patterns in POI recommendations. Additionally, Kurashima et al. [16] proposed a geo-topic model to estimate interested places by learning the user's activity area and various features of locations. Lian et al. [18] studied POI recommendation using a weighted matrix factorization and an augmented latent space model. However, the above solutions all used *supervised learning* approaches for personalized POI recommendation, which did not work for users who have little or no prior data to train their models. In contrast, this paper developed an *unsupervised approach* to address the interesting place discovery problem that requires no training data.

In information retrieval and data mining, there exists a good amount of work on the topic of mining geo-spatial data traces to discover interesting places for average people. For example, Zheng et al. [43], [44] proposed a *Hyperlink-Induced Topic Search (HIST)* based method to recommend

interesting places for visitors by mining their check-in patterns. Tiwari et al. [26] used the semantic features of geo-spatial regions to recommend popular and significant places. Furthermore, Khetarpaul et al. [15] used relational algebra operators combined with statistical operators to determine interesting locations from the aggregated GPS traces of multiple users. Zhang et al. [40] developed a novel method to predict links across partially aligned location-based social networks and address the data sparsity problem in interesting place finding. However, the above works either assumed *linear* correlations between the travel experience of users and the interestingness of places or ignored the *physical-social* dependency between places and users respectively. In contrast, this paper considers both the *nonlinear* relationship scenario and the physical-social dependency in the proposed model, which is shown to significantly improve the accuracy of the interesting place discovery results.

III. Problem Formulation

In this section, we formulate the physical-social-aware interesting place discovery problem as a constraint optimization problem. In particular, we consider a location-based crowdsourcing application (e.g., LBSN) where a group of M users (i.e., $U_1, U_2, ..., U_M$) visit a set of N places (i.e., $P_1, P_2, .., P_N$). For simplicity, we assume the interestingness of a place to be binary [2]. Specifically, $P_k = I$ represents that place P_k is interesting and $P_k = \bar{I}$ represents that place P_k is not interesting. We further define the following terms to be used in our model.

- UP is defined as a $M \times N$ matrix that represents the visiting behavior of all users U at all places P. It is referred to as the *User-Place Matrix*. In UP, $U_iP_k = 1$ when user U_i visits place P_k and $U_iP_k = 0$ otherwise.
- PD is defined as a set of joint probability distributions that describe the physical dependency between places in the system. Suppose we divide all N places into R independent groups (e.g., based on their geographic proximity) where places in the same group are correlated and places in different groups are independent. In particular, for a group r, we have a joint distribution PD_r to represent the dependency between places in the group. PD_r is often known from the application context or can be learned from the collected data [3].
- SD is defined as a $M \times M$ matrix to represent the social dependency between users. It is referred to as the *Social-Dependency Matrix*. In SD, $SD_{i,j} = 1$ when user U_i and U_j have a friend relationship and $SD_{i,j} = 0$ otherwise. In this paper, we consider the bi-directional friendship between users (e.g., friendship on Facebook and Foursquare) and SD is a symmetric matrix (i.e., $SD_{i,j} = SD_{j,i}$). It is trivial to extend our model to handle directional friendship as well. Based on SD, we can divide all users into C independent groups where users in the same group are dependent and users in different groups are independent.

We formulate our physical-social-aware interesting place discovery problem as follows. First, we define several important probability terms to be used in the problem formulation: if a user U_i is independent (i.e., U_i does not have any social connection with other users), Te_i is defined as the U_i's *independent travel experience*, which is the probability that a place P_k is interesting given that the user visits P_k. If U_i is dependent (i.e., user U_i has social connections with other users), we define $Te_{i,j}$ as the user's *dependent travel experience*, which is the probability that a friend of U_i (i.e., U_j) visits an interesting place P_k given that U_i has visited P_k. Formally, Te_i and $Te_{i,j}$ are defined as follows:

$$\begin{aligned} Te_i &= \Pr(P_k = I | U_iP_k = 1) \\ Te_{i,j} &= \Pr(P_k = I, U_jP_k = 1 | U_iP_k = 1) \end{aligned} \tag{1}$$

We further define a few relevant conditional probabilities: if U_i is independent, E_i and F_i are defined as the probability that U_i visits a place P_k given that P_k is interesting (or not) respectively. If U_i is dependent, $E_{i,j}$ and $F_{i,j}$ are defined as the probability that user U_i visits a place P_k given that the place is interesting (or not) and U_i's friend U_j also visits P_k respectively. Formally, E_i, $E_{i,j}$, F_i and $F_{i,j}$ are defined as:

$$\begin{aligned} E_i &= \Pr(U_iP_k = 1 | P_k = I) \\ E_{i,j} &= \Pr(U_iP_k = 1 | U_jP_k = 1, P_k = I) \\ F_i &= \Pr(U_iP_k = 1 | P_k = \bar{I}) \\ F_{i,j} &= \Pr(U_iP_k = 1 | U_jP_k = 1, P_k = \bar{I}) \end{aligned} \tag{2}$$

Observing that users may visit different numbers of places, we denote the probability that user U_i visits a place by p_i (i.e., $p_i = \Pr(U_iP_k = 1)$) where P_k is a randomly chozen place. We further denote d as the prior probability that a randomly chosen place is interesting (i.e., $d = \Pr(P_k = I)$). Using the Bayes' theorem, we can obtain the relationship between the items defined above:

$$\begin{aligned} E_i &= \frac{Te_i \times p_i}{d}, \quad F_i = \frac{(1 - Te_i) \times p_i}{(1 - d)} \\ E_{i,j} &= \frac{Te_{i,j} \times p_i}{Te_j \times p_j}, \quad F_{i,j} = \frac{(1 - Te_{i,j}) \times p_i}{(1 - Te_j) \times p_j} \end{aligned} \tag{3}$$

Using the above definitions, we formally formulate the physical-social-aware interesting place discovery problem as a constraint maximum likelihood estimation (MLE) problem: given the User-Place matrix UP, the joint distribution of physical dependency between places PD and the Social-Dependency matrix SD, the objective is to estimate both the *interestingness of each place* and the *travel experience*

[2] It turns out our solution presented in the next section could also provide a quantitative metric to evaluate exactly how interesting a place would be.

[3] In the evaluation, we showed how to obtain the PD from a real world crowdsourcing application using an open map service.

of each user without knowing either of them *a priori*. Formally, it is given as follows:

$$\forall k, 1 \leq k \leq N : \Pr(P_k = I | UP, PD, SD)$$
$$\forall i, 1 \leq i \leq M : \Pr(P_k = I | U_i P_k = 1) \quad (4)$$

IV. Solution

In this section, we develop a Physical-Social-aware Interesting Place Discovery (PSIPD) scheme to solve the optimization problem formulated in the previous section.

A. Likelihood Function Formulation

It turns out that our constraint MLE problem lends itself nicely to an expectation maximization (EM) solution [6]. In particular, given a set of observations, EM can estimate both the parameters and the hidden variables of a MLE model, which is most consistent (in MLE sense) with the observed data. To develop an EM solution, let us first define a likelihood function $L(\Theta; X, Z) = p(X, Z|\Theta)$, where X is the observed data, Θ is a set of estimation parameters and Z denotes a set of hidden variables. EM computation contains two iterative steps: *E-step* and *M-step*. The E-step maximizes the likelihood function w.r.t. Z and the M-step maximizes the likelihood function w.r.t. Θ. Formally, they are given as:

$$\text{E-step: } Q(\Theta|\Theta^{(n)}) = E_{Z|x,\Theta^{(n)}}[logL(\Theta; x, Z)] \quad (5)$$
$$\text{M-step: } \Theta^{(n+1)} = \arg\max_{\Theta} Q(\Theta|\Theta^{(n)}) \quad (6)$$

In our physical-social-aware interesting place discovery problem, the observed data include the User-Place Matrix UP, the joint distribution of physical dependency between places PD and the Social-Dependency Matrix SD. The estimation parameter $\Theta = (E_1, ..., E_M; F_1, ..., F_M; E_{1,j}, ..., E_{M,j}; F_{1,j}, ..., F_{M,j}; d)$, where E_i, F_i, $E_{i,j}$, $F_{i,j}$ and d are defined in Equation (2). Furthermore, we define a vector of latent variables Z to indicate the interestingness of places. Specifically, we have a corresponding variable z_k for each place P_k (i.e., $z_k = 1$ if $P_k = I$ and $z_k = 0$ otherwise). Hence, the likelihood function of the physical-social-aware interesting place finding problem can be written as:

$$\begin{aligned} L(\Theta; X, Z) &= \Pr(X, Z|\Theta) \\ &= \prod_{r \in R} \Pr(X_r, Z_r|\Theta) = \prod_{r \in R} \Pr(Z_r) \times \Pr(X_r|Z_r, \Theta) \\ &= \prod_{r \in R} \left\{ \sum_{r_1,...,r_h \in \Psi_r} \Pr(z_{r_1}, ..., z_{r_h}) \prod_{k \in r} \prod_{g \in C} \prod_{i \in g} \eta_{k,g,i} \right\} \end{aligned} \quad (7)$$

where $\eta_{k,g,i}$ is defined in Table I and $\Pr(z_{r_1}, ..., z_{r_h})$ represents the joint probability distribution of places in an independent group r. We let Ψ_r represent all possible combinations of r_1,...,r_h in the group (e.g., for a group of two places, $\Psi_r = [(0,0), (0,1), (1,0), (1,1)]$). $|g|$ denotes the size of a user independent group g. Figure 2 illustrates the key parameters and E and M steps of the PSIPD scheme.

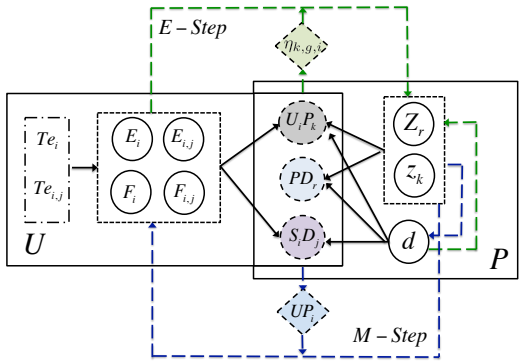


Figure 2. The E and M Steps of PSIPD Scheme

Table I
Notation for PSIPD

$\eta_{k,g,i}$	Constrains
E_i	$\|g\| == 1, U_iP_k = 1, z_k = I$
$1 - E_i$	$\|g\| == 1, U_iP_k = 0, z_k = I$
$\prod_{j \in g} E_{i,j}$	$\|g\| > 1, U_iP_k = 1, U_jP_k = 1, S_iD_j = 1, z_k = I$
$\prod_{j \in g} 1 - E_{i,j}$	$\|g\| > 1, U_iP_k = 0, U_jP_k = 1, S_iD_j = 1, z_k = I$
F_i	$\|g\| == 1, U_iP_k = 1, z_k = \bar{I}$
$1 - F_i$	$\|g\| == 1, U_iP_k = 0, z_k = \bar{I}$
$\prod_{j \in g} F_{i,j}$	$\|g\| > 1, U_iP_k = 1, U_jP_k = 1, S_iD_j = 1, z_k = \bar{I}$
$\prod_{j \in g} 1 - F_{i,j}$	$\|g\| > 1, U_iP_k = 0, U_jP_k = 1, S_iD_j = 1, z_k = \bar{I}$

B. PSIPD Scheme

Given the above likelihood function of our problem, we can derive the corresponding E-step and M-step of the PSIPD scheme. First, we derive the Q function for the E-step based on Equation (5):

$$\begin{aligned} Q(\Theta|\Theta^{(n)}) &= E_{Z|X,\Theta^{(n)}}[logL(\Theta; X, Z)] \\ &= \sum_{r \in R} \Pr(z_{r_1}, ..., z_{r_h}|X_r, \Theta^{(n)}) \\ &\times \left\{ \sum_{k \in r} \sum_{g \in C} \sum_{i \in g} log(\eta_{i,k,g}) + log\Pr(z_{r_1}, ..., z_{r_h}) \right\} \end{aligned} \quad (8)$$

where $\Pr(z_{r_1}, ..., z_{r_h}|X_r, \Theta^{(n)})$ is the conditional joint probability of all places in the independent group r (i.e., r_1,...,r_h). Given the observed data regarding these places X_r and the current estimates of the parameters $\Theta^{(n)}$, $\Pr(z_{r_1}, ..., z_{r_h}|X_r, \Theta^{(n)})$ can be computed as:

$$\begin{aligned} &\Pr(z_{r_1}, ..., z_{r_h}|X_r, \Theta^{(n)}) \\ &= \frac{\Pr(z_{r_1}, ..., z_{r_h}; X_r, \Theta^{(n)})}{\Pr(X_r, \Theta^{(n)})} \end{aligned} \quad (9)$$

where $\Pr(z_{r_1}, ..., z_{r_h}; X_r, \Theta^{(n)})$ and $\Pr(X_r, \Theta^{(n)})$ can be further expressed as follows:

$$\begin{aligned}
&\Pr(z_{r_1}, ..., z_{r_h}; X_r, \Theta^{(n)}) \\
&= \Pr(X_r, \Theta^{(n)} | z_{r_1}, ..., z_{r_h}) \times \Pr(z_{r_1}, ..., z_{r_h}) \\
&= \prod_{k \in r} \prod_{g \in C} \prod_{i \in g} \eta_{k,g,i} \times \Pr(z_{r_1}, ..., z_{r_h}) \\
&\Pr(X_r, \Theta^{(n)}) \\
&= \sum_{r_1, ..., r_h \in \Psi_r} \left[\Pr(X_r, \Theta^{(n)} | z_{r_1}, ..., z_{r_h}) \times \Pr(z_{r_1}, ..., z_{r_h}) \right] \\
&= \sum_{r_1, ..., r_h \in \Psi_r} \left[(\prod_{k \in r} \prod_{g \in C} \prod_{i \in g} \eta_{k,g,i}) \times \Pr(z_{r_1}, ..., z_{r_h}) \right]
\end{aligned} \tag{10}$$

For simplicity, we further denote $\Pr(z_k = I | X_k, \Theta^{(n)})$ as $Y(n,k)$. It is the probability that place P_k is interesting given the observed data and the current estimation parameters. $Y(n,k)$ can be computed as *marginal distribution* of the joint probability of all places in the independent group r to which place P_k belongs (i.e., $\Pr(z_{r_1}, ..., z_{r_h} | X_r, \Theta^{(n)})$, $k \in r$).

For the M-step, in order to get the optimal Θ^* that maximizes Q function, we set partial derivatives of $Q(\Theta | \Theta^{(n)})$ given by Equation (8) with respect to Θ to 0. In particular, we get the solutions of $\frac{\partial Q}{\partial E_i} = 0$, $\frac{\partial Q}{\partial F_i} = 0$, $\frac{\partial Q}{\partial E_{i,j}} = 0$, $\frac{\partial Q}{\partial F_{i,j}} = 0$, $\frac{\partial Q}{\partial d} = 0$ in each iteration, we can get expressions of the optimal E_i^*, F_i^*, $E_{i,j}^*$, $F_{i,j}^*$ and d^*:

$$\begin{aligned}
E_i^{(n+1)} = E_i^* &= \frac{\sum_{k \in UP_i} Y(n,k)}{\sum_{k=1}^{N} Y(n,k)} \\
F_i^{(n+1)} = F_i^* &= \frac{\sum_{k \in UP_i} (1 - Y(n,k))}{\sum_{k=1}^{N} (1 - Y(n,k))} \\
E_{i,j}^{(n+1)} = E_{i,j}^* &= \frac{\sum_{k \in UP_{i,j}} Y(n,k)}{\sum_{k \in UP_j} Y(n,k)} \\
F_{i,j}^{(n+1)} = F_{i,j}^* &= \frac{\sum_{k \in UP_{i,j}} (1 - Y(n,k))}{\sum_{k \in UP_j} (1 - Y(n,k))} \\
d^{(n+1)} = d^* &= \frac{\sum_{k=1}^{N} Y(n,k)}{N}
\end{aligned} \tag{11}$$

where UP_i is the set of places user U_i visits and $UP_{i,j}$ is the set of places both user U_i and U_j visit.

V. Evaluation

In this section, we evaluate the performance of the PSIPD scheme and compare it with the state-of-the-art baselines on two real world data traces. To better understand the effects of exploiting physical and social dependency on the final results, we consider three variants of the PSIPD scheme in our evaluation: (i) *PSIPD-P*: a simplified version of PSIPD that only considers the physical dependency between places; (ii) *PSIPD-S*: a simplified version of PSIPD that only considers the social dependency between users; (iii) *PSIPD-PS*: the full version of PSIPD that considers both physical and social dependency.

In the rest of this section, we first describe the experimental setups and data pre-processing steps. Then we introduce the state-of-the-art baselines and evaluation metrics used in the experiments. Finally, we present the evaluation results and demonstrate the performance improvements achieved by our proposed scheme.

A. Experiment Settings

1) Data Trace Statistics: we use two real world data traces to evaluate our proposed schemes. These traces are collected from location-based social network services, namely, Brightkite[4] and Gowalla[5] [5]. In these location-based social network services, users share their location information (i.e., check-in records) at different places. Each check-in record is formatted as: (user ID, latitude, longitude, timestamp). The Brightkite trace was collected from April 2008 to October 2010 and the Gowalla trace was collected from February 2009 to October 2010. Additionally, the social dependency information is specified as bi-directional friendship in the traces. Other statistics of the traces are presented in Table II.

Table II
Trace Statistics

Description	Brightkite	Gowalla
Number of Users	58,228	107,092
Number of Friendships	214,078	950,327
Number of Check-ins	4,491,143	6,442,890

2) Data Pre-Processing: To evaluate our methods in real-world settings, we conducted the following data pre-processing steps: (i) clustering all raw geographical check-in points into meaningful clusters that represent places in the physical world; (ii) identifying independent groups of places based on their physical locations; (iii) identifying independent groups of users based on their social connections. The goal of the above pre-processing steps is to generate the inputs to the PSIPD scheme: the User-Place Matrix UP, the joint probability distributions of place dependency PD, and the Social-Dependency Matrix SD we defined in the Model Section.

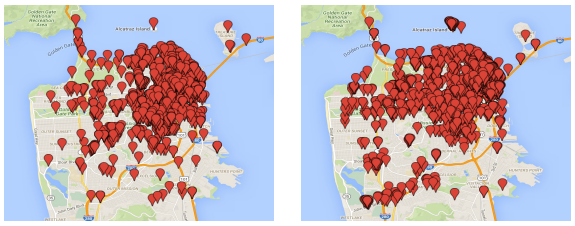
In our real-world evaluation, we select San Francisco as our target city. Figure 3 shows the check-in points of two data traces in San Francisco [6]. We also plotted the distributions of the check-in points per user in Figure 4. The figure suggest power-law-like distributions on both data

[4]http://snap.stanford.edu/data/loc-brightkite.html

[5]http://snap.stanford.edu/data/loc-gowalla.html

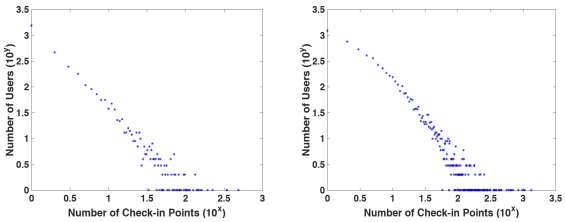
[6]We did plot duplicate check-in points (i.e., check-in points from the same user at the same location)

traces which are consistent with typical observations in social networks [14].



(a) Brightkite Trace (b) Gowalla Trace

Figure 3. Maps of San Francisco Check-in Points



(a) Brightkite Trace (b) Gowalla Trace

Figure 4. Distribution of Check-in Points Per User

Clustering: we applied the K-means clustering algorithm [25] to first cluster the raw check-in records (with duplicated ones removed) into intermediate clusters without any semantic meanings (i.e., the clustering process only considered the physical distance between locations). We then re-organized the intermediate clusters into meaningful places using the *City Point-of-Interest* service from Google Map [7]. As a result, we found 53 places in total for the Brightkite trace, of which 24 places are interesting and 29 places are not interesting. For the Gowalla trace, we found 56 places in total, of which 25 places are interesting and 31 places are not interesting. After the clustering step, we generate the User-Place Matrix UP by associating each user with the places the user visited.

Identifying Physical Independent Groups: we manually examined the places returned by the previous clustering step and organized them into physical independent groups. Places within the same independent group are so close to each other that users are more likely to visit them together (however, users may not check in at each individual place.). In particular, we identified 26 independent groups in the Brightkite trace and 29 in the Gowalla trace. Furthermore, we empirically estimate the joint probability distribution of places in the same independent group based on their locations and historical visiting records [5]. After this step, we generate the set of joint probability distributions PD.

Identifying Social Independent Groups: we used SPLA [37] (a community detection algorithm) to identify independent groups of users. We first obtain the social connections between users from the friendship information included in the trace. In particular, we generated the social dependency graph as an undirected graph $G = (V, E)$ where V and E represents the set of users and their friendship respectively: if user u is a friend of user v in the trace, we have a link between u and v. We then applied the SPLA algorithm on the graph G to partition the whole set of users into different independent groups. The users in the same independent group form a clique in graph G. Using the output of this step, we generate the Social-Dependency Matrix SD.

[7]https://www.google.com/maps

B. Baselines and Evaluation Metrics

1) Baselines: In the evaluation, we compare the performance of our new schemes (i.e., *PSIPD-P*, *PSIPD-S* and *PSIPD-PS*) with the following state-of-the-art baselines from current literature. The first (and simplest) baseline is *Voting*, which computes the interestingness of a place by counting the number of times the place is visited. The second baseline is the *HITS* [44], which assumes linear relationship between the user's travel experience and the interestingness of a place. The third baseline is *Regular-EM* which is shown to outperform four state-of-the-art techniques in identifying interesting entities from noisy crowdsourcing data [35]. Note that the above baselines all assume that both places and users are *independent*. Additionally, we also compare the performance of PSIPD with four recent baselines from Point-of-Interest recommendation literature: the first baseline is called *iGSLR* [41], which explored the geographical proximity influence on users' check-in behaviors in computing the interestingness of a place. The second baseline is *GeoSoCa* [42], which explored geographical, social and category information for Point-of-Interest recommendations. The third baseline is *STT* [9], which captured the spatial and temporal aspects of check-ins to recommend locations. The fourth baseline is *GTM* [16], which developed a geo-topic model to incorporate the user's activity area into the estimation process of interesting places.

2) Evaluation Metric: In the experiments, we use two sets of evaluation metrics. The first set of metrics are used to evaluate the accuracy of different techniques in terms of discovering interesting places. These metrics are: *precision*, *recall*, *F1-measure* [21] and receiver operating characteristics (ROC) curves [7]. The second set of metrics are used to evaluate the ranking performance of different schemes. [8] These metrics are *normalized discounted cumulative gains (NDCG)* [13]. NDCG is an indicator of the average ranking performance of all compared schemes. Given a query, NDCG at position n is calculated as:

$$NDCG(n) = Nr(n) \times \sum_{l=1}^{n} \frac{2^{r(l)} - 1}{log(1 + l)} \qquad (12)$$

[8]To evaluate the ranking performance, we ranked all places using the estimated interestingness scores of places returned by different schemes.

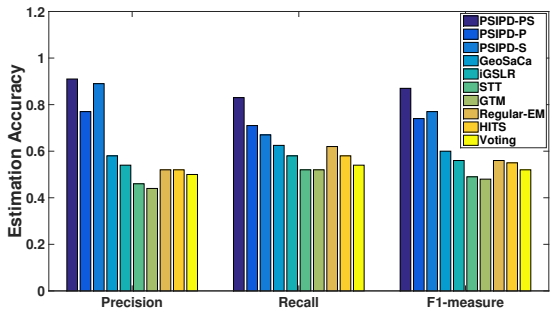


Figure 5. Estimation Accuracy on Brightkite Trace

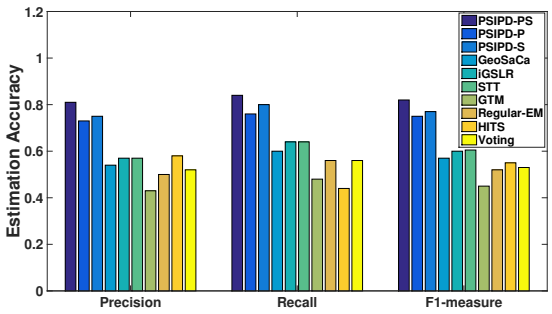


Figure 6. Estimation Accuracy on Gowalla Trace

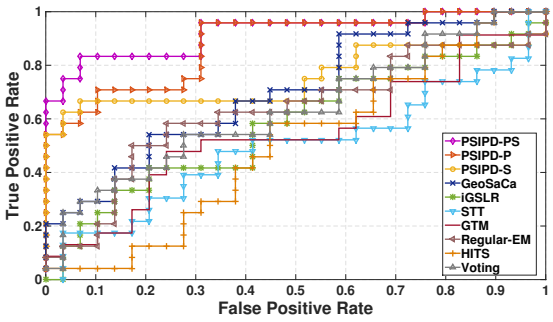


Figure 7. ROC Curves on Brightkite Trace

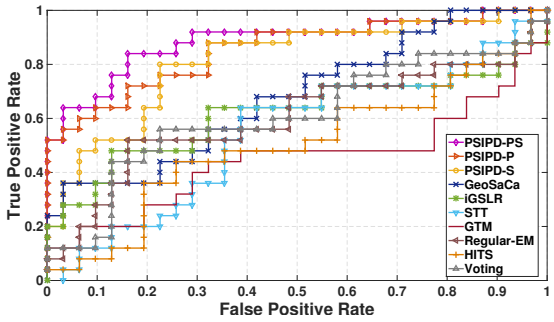


Figure 8. ROC Curves on Gowalla Trace

where $r(l)$ indicates the score for rank l. In our case, $r(l)$ is equal to 1 if the l-th place is interesting and $r(l) = 0$ otherwise. $Nr(n)$ is a normalization factor that guarantees the NDCG of the perfect ranking scheme is equal to 1.

C. Evaluation Results

In this section, we conducted experiments on two real-world data traces to compare the performance of *PSIPD-P*, *PSIPD-S* and *PSIPD-PS* schemes with 7 state-of-the-art baselines (i.e., *GeoSoCa*, *iGSLR*, *STT*, *GTM*, *Regular-EM*, *Sums-Hubs* and *Voting*) in terms of *estimation accuracy* and *ranking performance*. Independently from two data traces we used in evaluation, we collected ground truth values (i.e. whether a place is interesting or not) from three widely used travel recommendation websites: *TripAdvisor*, *Planet Aware* and *San Francisco Travel*. We then decide whether a place is interesting using the following rubric:

- *Interesting places:* Places that have been recommended by at least two of these travel recommendation websites.
- *Unconfirmed places:* Places that do not satisfy the requirement of interesting places.

Note that "unconfirmed places" may include both places that are not interesting or potentially interesting places that cannot be independently verified by using the above rubric. Hence, our evaluation results present *pessimistic* bounds on the performance.

1) Estimation Performance: We first conduct experiments to evaluate the estimation performance of all schemes in terms of *precision*, *recall* and *F1-measure*. The results on Brightkite trace are shown in Figure 5. We observe that all our proposed schemes (i.e., *PSIPD-P*, *PSIPD-S* and *PSPID-PS*) outperform all the compared baselines in terms of precision, recall and F1-measure. We also observe that *PSPID-PS* performs the best among all schemes compared. The largest performance gain achieved by *PSIPD-PS* on precision over the best performed baseline (*GeoSoCa*) is 33%. The largest performance gain achieved by *PSIPD-PS* on recall is 21%. The results on Gowalla trace are shown in Figure 6. We observe similar results: all our proposed schemes continue to outperform the baselines and the largest performance gain achieved by *PSIPD-PS* on precision and recall (compared to the best performed baseline) is 23% and 20% respectively.

We also drew the ROC curves of all compared schemes on the two data traces in Figure 7 and Figure 8 respectively. We observe that PSIPD schemes (i.e., *PSIPD-P*, *PSIPD-S* and *PSIPD-PS*) perform better than the baselines while *PSIPD-PS* achieves the best performance at different true positive and false positive rates. This performance improvement of *PSIPD-PS* is achieved by explicitly exploiting both physical dependency between places and social dependency between users under a rigorous analytical framework.

2) Ranking Performance: We also evaluate the ranking performance of all schemes and use the *NDCG@n* [13]

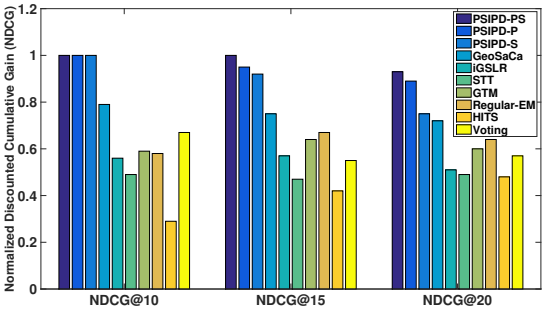


Figure 9. NDCG@n Evaluation on Brightkite Trace

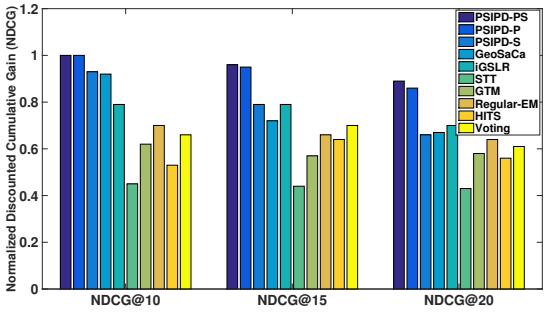


Figure 10. NDCG@n Evaluation on Gowalla Trace

metrics we introduced earlier. In Figure 9 and Figure 10, we compare the performance of the PSIPD schemes to all baselines in terms of NDCG@n on two data traces respectively. We observe PSIPD schemes continue to outperform the baselines at different values of n. Also, *PSIPD-PS* achieves the perfect NDCG score when $n \leq 15$.

VI. Conclusion

This paper develops an unsupervised social-physical-aware approach to solve the interesting place discovery problem using location-based social network data feeds. The proposed PSIPD scheme explicitly exploits both the physical dependency between places and social dependency between users under a maximum likelihood estimation framework. We evaluated the performance of the PSIPD scheme on two real world data traces collected from LBSN and our results showed that the PSIPD scheme significantly outperformed the state-of-the-art baselines in terms of both estimation accuracy and ranking performance. The result of the paper is important because it allows crowdsourcing applications to systematically investigate the physical-social dependency in the interesting place discovery problem using a principled approach.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1566465 and IIS-1447795.

References

[1] T. Abdelzaher and D. Wang. Analytic challenges in social sensing. In *The Art of Wireless Sensor Networks*, pages 609–638. Springer, 2014.

[2] C. C. Aggarwal and T. Abdelzaher. Social sensing. In *Managing and Mining Sensor Data*, pages 237–297. Springer, 2013.

[3] M. T. Al Amin, T. Abdelzaher, D. Wang, and B. Szymanski. Crowd-sensing with polarized sources. In *Distributed Computing in Sensor Systems (DCOSS), 2014 IEEE International Conference on*, pages 67–74. IEEE, 2014.

[4] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. *ICWSM*, 2011:81–88, 2011.

[5] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[7] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[8] H. Gao, J. Tang, X. Hu, and H. Liu. Modeling temporal effects of human mobile behavior on location-based social networks. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1673–1678. ACM, 2013.

[9] B. Hu, M. Jamali, and M. Ester. Spatio-temporal topic modeling in mobile social media for location recommendation. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 1073–1078. IEEE, 2013.

[10] C. Huang and D. Wang. On interesting place finding in social sensing: An emerging smart city application paradigm. In *2015 IEEE International Conference on Smart City 2015 (SmartCity 2015)*. IEEE, 2015.

[11] C. Huang and D. Wang. Spatial-temporal aware truth finding in big data social sensing applications. In *Trustcom/BigDataSE/ISPA, 2015 IEEE*, volume 2, pages 72–79. IEEE, 2015.

[12] C. Huang, D. Wang, and N. Chawla. Towards time-sensitive truth discovery in social sensing applications. In *Mobile Ad Hoc and Sensor Systems (MASS), 2015 IEEE 12th International Conference on*, pages 154–162. IEEE, 2015.

[13] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[14] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, page 8. ACM, 2013.

[15] S. Khetarpaul, R. Chauhan, S. Gupta, L. V. Subramaniam, and U. Nambiar. Mining gps data to determine interesting locations. In *Proceedings of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011*, page 8. ACM, 2011.

[16] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya, and K. Fujimura. Geo topic model: joint modeling of user's activity area and interests for location recommendation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 375–384. ACM, 2013.

[17] K. Lee, K. Y. Kamath, and J. Caverlee. Combating threats to collective attention in social media: An evaluation. In *ICWSM*, 2013.

[18] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui. Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 831–840. ACM, 2014.

[19] M. Lichman and P. Smyth. Modeling human location data with mixtures of kernel densities. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 35–44. ACM, 2014.

[20] A. Majid, L. Chen, H. T. Mirza, I. Hussain, and G. Chen. A system for mining interesting tourist locations and travel sequences from public geo-tagged photos. *Data & Knowledge Engineering*, 95:66–86, 2015.

[21] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[22] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1038–1043. IEEE, 2012.

[23] E. Papalexakis, K. Pelechrinis, and C. Faloutsos. Spotting misbehaviors in location-based social networks using tensors. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 551–552. International World Wide Web Conferences Steering Committee, 2014.

[24] J. Shi, N. Mamoulis, D. Wu, and D. W. Cheung. Density-based place clustering in geo-social networks. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 99–110. ACM, 2014.

[25] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. 2005.

[26] S. Tiwari and S. Kaushik. Mining popular places in a geo-spatial region based on gps data using semantic information. In *Databases in Networked Information Systems*, pages 262–276. Springer, 2013.

[27] R. R. Vatsavai, A. Ganguly, V. Chandola, A. Stefanidis, S. Klasky, and S. Shekhar. Spatiotemporal data mining in the era of big spatial data: algorithms and applications. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 1–10. ACM, 2012.

[28] D. Wang, T. Abdelzaher, and L. Kaplan. Surrogate mobile sensing. *Communications Magazine, IEEE*, 52(8):36–41, 2014.

[29] D. Wang, T. Abdelzaher, and L. Kaplan. *Social sensing: building reliable systems on unreliable data*. Morgan Kaufmann, 2015.

[30] D. Wang, T. Abdelzaher, L. Kaplan, R. Ganti, S. Hu, and H. Liu. Exploitation of physical constraints for reliable social sensing. In *The IEEE 34th Real-Time Systems Symposium (RTSS'13)*, 2013.

[31] D. Wang, H. Ahmadi, T. Abdelzaher, H. Chenji, R. Stoleru, and C. C. Aggarwal. Optimizing quality-of-information in cost-sensitive sensor data fusion. In *Distributed Computing in Sensor Systems and Workshops (DCOSS), 2011 International Conference on*, pages 1–8. IEEE, 2011.

[32] D. Wang, M. T. Al Amin, T. Abdelzaher, D. Roth, C. R. Voss, L. M. Kaplan, S. Tratz, J. Laoudi, and D. Briesch. Provenance-assisted classification in social networks. *Selected Topics in Signal Processing, IEEE Journal of*, 8(4):624–637, 2014.

[33] D. Wang and C. Huang. Confidence-aware truth estimation in social sensing applications. In *Sensing, Communication, and Networking (SECON), 2015 12th Annual IEEE International Conference on*, pages 336–344. IEEE, 2015.

[34] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal. On scalability and robustness limitations of real and asymptotic confidence bounds in social sensing. In *The 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON 12)*, June 2012.

[35] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *The 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN 12)*, April 2012.

[36] F. Wu, Z. Li, W.-C. Lee, H. Wang, and Z. Huang. Semantic annotaion of mobility data using social media. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1253–1263. International World Wide Web Conferences Steering Committee, 2015.

[37] J. Xie, B. K. Szymanski, and X. Liu. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 344–349. IEEE, 2011.

[38] H. Yoon, Y. Zheng, X. Xie, and W. Woo. Smart itinerary recommendation based on user-generated gps trajectories. In *Ubiquitous Intelligence and Computing*, pages 19–34. Springer, 2010.

[39] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong. Discovering urban functional zones using latent activity trajectories. *Knowledge and Data Engineering, IEEE Transactions on*, 27(3):712–725, 2015.

[40] J. Zhang, X. Kong, and P. S. Yu. Transferring heterogeneous links across location-based social networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 303–312. ACM, 2014.

[41] J.-D. Zhang and C.-Y. Chow. igslr: personalized geo-social location recommendation: a kernel density estimation approach. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 334–343. ACM, 2013.

[42] J.-D. Zhang and C.-Y. Chow. Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 443–452. ACM, 2015.

[43] Y. Zheng and X. Xie. Learning travel recommendations from user-generated gps traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):2, 2011.

[44] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800. ACM, 2009.

[45] W.-Y. Zhu, W.-C. Peng, L.-J. Chen, K. Zheng, and X. Zhou. Modeling user mobility for location promotion in location-based social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1573–1582. ACM, 2015.