Towards Social-aware Interesting Place Finding in Social Sensing Applications

Chao Huang, Dong Wang, Brian Mann

Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556

Abstract

This paper develops a principled approach to accurately identify interesting places in a city through social sensing applications. Social sensing has emerged as a new application paradigm, where a crowd of social sources (humans or devices on their behalf) collectively contribute a large amount of observations about the physical world. This paper studies an *interesting place finding* problem, in which the goal is to correctly identify the interesting places in a city. Important challenges exist in solving this problem: (i) the interestingness of a place is not only related to the number of users who visit it, but also depends upon the travel experience of the visiting users; (ii) the user's social connections could directly affect their visiting behavior and the interestingness judgment of a given place. In this paper, we develop a new *Social-aware Interesting Place Finding Plus (SIPF+)* approach that addresses the above challenges by explicitly incorporating both the user's *travel experience* and *social relationship* into a rigorous analytical framework. The SIPF+ scheme can find interesting places not typically identified by traditional travel websites (e.g., TripAdvisor, Expedia). We compare our solution with state-of-the-art baselines using two real-world datasets collected from location-based social network services and verified the effectiveness of our approach.

Keywords: Interesting Place Finding, Social Dependency, Social Sensing, Crowdsourcing, Expectation Maximization

1. Introduction

This paper develops a principled approach to accurately identify interesting places in a city through social sensing applications. This work is motivated by the emergence of social sensing as a new application paradigm of collecting observations about the physical world from social sources (humans or devices on their behalf) [1]. This paradigm is enabled by a few recent technical trends: (i) the proliferation of smart devices (e.g., smartphones) owned by average individuals; (ii) the ubiquitous coverage of wireless communication (e.g., 4G, WiFi, WiMax); (iii) the advent of online social media (e.g., Twitter, Foursquare, Facebook). For example, common citizens can now easily use a Location-Based Social Network (LBSN) service (e.g., Foursquare) on their mobile phones to upload the "check-in" points of the places they visit in a city. Alternatively, a group of drivers may use a smartphone app to report traffic conditions (e.g., congestion, accidents, etc.) they experience in a given area. In this paper, we focus on an *interesting place finding* problem, where the goal is to correctly identify the interesting places in a city where people may have strong interest in visiting (e.g., parks, museums, historic sites, scenic trails, etc.). The results of this work can be used to develop future travel recommendation systems, mobile guidance applications, and user travel experience sharing applications that explore the power of social sensing data contributed by common citizens [2, 3]. For example, the results can help people find more interesting places in a city, design a better route for their travels, and share their travel experience with other users in a timely fashion.

Previous studies have adopted social sensing (in some cases referred to as crowdsourcing) to solve the interesting place finding problem. The main idea behind current solutions is to automatically infer the locations of interesting places in a given region (e.g., a city) from the check-in points or GPS traces that users share when using location-aware applications [4]. The advantages of using crowdsourcing methods compared to the traditional methods (e.g., search engine, travel websites) are threefold. *First*, the cost of data collection using crowdsorucing is low since the location data of users is already made available through the location based services (e.g., LBSN) [5]. *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 [6]. *Third*, the crowdsourcing traces normally have a better spatial-temporal coverage of the interesting places, as the crowd is naturally distributed across the region [7]. Table 1 also shows some examples of interesting places in the cities of Chicago and San Francisco that are either not recommended by traditional travel websites (e.g., TripAdvisor, Expedia, and CityPass) or have very low recommendation rankings on those websites. However, those places are identified as very interesting places by many people who visited them in person and shared their experience on social media (e.g., Twitter and Foursquare). In this paper, we develop a new social sensing based scheme that is able to find such interesting places that cannot easily be identified by traditional travel websites.

Chicago	San Francisco
North Avenue Beach	Conservatory of Flowers
Museum of Contemporary Art	Japanese Tea Garden
Oriental Theatre	Aquarium of the Bay
The Peggy Notebaert Nature Museum	Yerba Buena Gardens
Music Box Theatre	Mission Dolores Park

Table 1: Interesting Places that are Missed by Traditional Travel Websites

While previous studies in information retrieval [8, 7] , data mining [9, 4], and social sensing [10, 11] have made significant efforts to address the interesting place finding problem, two important limitations remain in current solutions. *First*, the current techniques are mostly heuristic-based and make strong assumptions when they handle users in the problem. For example, they either assume all users have exactly the same travel experience [1] or the correlation be-

[1] The travel experience of a user is highly correlated with the user's ability to find interesting places in a city [12]

tween a user's travel experience and the number of places he/she visited is simply linear [8]. However, these assumptions cannot be applied in real-world scenarios where the relationship between a user's travel experience and the number of places he/she visited is *nonlinear* [13]. The interesting place finding problem becomes more challenging when neither the user's travel experience nor the interestingness of a place is known *a priori* [14]. Hence, we need to develop a new framework that can accurately model both the user's travel experience and the interestingness of places based on the social sensing data observed. *Second*, the social connections between users could easily affect their visiting behavior and the judgment on the interestingness of places they visited. For example, a group of colleagues who work in the same company are more likely to visit the same building every day; however, the building of their company may not necessarily be interesting to the general public. Unfortunately, current interesting place finding techniques completely ignore the impact of a user's social dependency, which can easily lead to suboptimal solutions, as we observed in our experiments.

In this paper, we develop a Social-aware Interesting Place Finding (SIPF+) scheme that addresses the above limitations by explicitly incorporating both the user's *travel experience* and *social dependency* into a Maximum Likelihood Estimation (MLE) framework. In particular, a principled, unsupervised learning approach based on Expectation Maximization (EM) is developed to jointly estimate both the user's travel experience and the interestingness of a place without prior knowledge on either. We evaluate SIPF+ using two real-world datasets collected from location-based social network services. The evaluation results show that our approach significantly outperforms the state-of-the-art baselines by correctly identifying more interesting places in a city while minimizing the number of false positives. The results of this paper are important because they allow social sensing applications to accurately identify interesting places by taking into account the user's travel experience and social dependency under a principled framework. To summarize, the contributions of this work are as follows:

- To the best of our knowledge, we are among the first to develop a principled, unsupervised learning framework that allows us to derive an optimal solution (in the sense of maximum likelihood estimation) for the social-aware interesting place finding problem.
- We explicitly consider both the user's travel experience and social dependency in the interesting place finding solution.
- Our MLE solution handles the nonlinear relationship between the user's travel experience and the interestingness of places.
- We perform extensive experiments on two real-world datasets, comparing the performance of our scheme to that of the state-of-the-art baselines.

A preliminary version of this work has been published in [11]. We refer to the previous version as SIPF. The current paper is a significant extension of the previous work in the following aspects. First, we extend our previous model in [11] by addressing more complex social dependency (i.e., arbitrary user dependency graph that includes cycles) between users and improve the interesting place finding results (Section 4). Second, we compare our scheme with more state-of-the-art baselines from Point of Interests (POI) recommendation systems and added more evaluation metrics (e.g., mean average precision) to evaluate the performance of our scheme (Section 5.2). Third, we perform a new set of experiments on a second city (i.e., Chicago) to further evaluate the robustness of our scheme (Section 5.3). Finally, we add a few real-world examples to demonstrate that our algorithms can identify interesting places more accurately than other baselines (Section 5.3).

2. Related Work

There exists a good amount of work on the topic of Points of Interests (POI) recommendation [15, 16, 8, 17, 18]. For example, Zhang et al. developed a kernel density estimation method (i.e., iGSLR) to infer the POI based on the geographical proximity [15] and social connections between users and the categorical information of places (i.e., GeoSoCa) [16]. Zheng et al. proposed an iterative approach (i.e., HITS) to explore users' travel experience and discover the interesting places in a simple linear way by using the GPS trajectories generated by users [8]. A geo-topic model (i.e., GTM) was proposed to estimate interesting places by learning the user's activity area and various features of locations [17]. Hu et al. [18] proposed a comprehensive model (i.e., STT) that explicitly considered the geographical influence and temporal activity patterns in POI recommendations. However, most of the above solutions used supervised learning approaches for personalized POI recommendation, which requires a significant amount of labeled data to train their models. In contrast, this paper develops an unsupervised approach to address the interesting place finding problem that requires no training data.

Natural Language Processing (NLP) has received a significant amount of attention with the advent of Social Web and online social media, in particular. Computational Intelligence models and approaches have been developed to achieve a deeper understanding of natural languages by leveraging semantic features and context that are not explicitly expressed in the text [19, 20, 21]. For example, Gangemi et al. [20] developed a new model to perform holder and topic detection in opinion sentences based on the neo-Davidsonian assumption. Lau et al. [21] proposed the design, development and evaluation of a weakly-supervised cybercriminal network mining approach to facilitate cybercrime forensics. In contrast, this paper focuses on structured and easy-to-process check-in data (i.e., GPS coordinates) extracted from LBSN. However, the NLP approaches can be integrated with our tool to mitigate data sparsity problems in social sensing where people might only use text to describe the location they visited rather than share the actual GPS coordinates. NLP techniques can be used to reliably transfer the unstructured, human-generated text information to structured, machine-processible data (e.g., GPS locations) and help achieve more accurate interesting place finding results.

Maximum likelihood estimation (MLE) approaches have been widely used in the domain of big data analysis [13, 22, 23, 24, 25]. For example, Vatsavai et al. [13] applied an MLE scheme to estimate the parameters of a Gaussian mixture model that deals with big spatial data. Denoeux et al. [22] developed an MLE framework to estimate the parameters in parametric statistical models using uncertain data. Banbura et al. [23] presented an MLE approach to infer a dynamic factor model on large datasets with an arbitrary pattern of missing data. Additionally, an MLE-based approach that considers the relative likelihood has been proposed to handle low-quality data [24]. However, the above work focused on the estimation of continuous variables. In contrast, this paper focuses on a set of binary variables that represent the interestingness of places. In our MLE framework, we develop an Expectation Maximization (EM) solution that explores the discrete nature of the estimation variables and social dependency of users in order to solve the interesting place finding problem in social sensing.

Our work is also related to event-based social network analysis. For example, Li et al. defined a social event organization problem and assigned users to events so as to maximize the overall happiness of the users [26]. She et al. proposed an event-planning problem by exploring both spatial and temporal constraints and maximizing the overall satisfaction of participants [27]. Liu et al. identified an event-based social network as a co-existence between both online and offline social interactions and studied its properties [28]. In addition, a new arrangement scheme has been proposed to solve a more general event-participant arrangement problem by considering the conflicts of different events [29]. Tong et al. defined the problem of bottleneck-aware social event arrangement and developed greedy heuristic algorithms to solve the problem [30]. Our work differs from the above works in two aspects: (i) this paper develops a new rigorous analytical framework to discover interesting places for average people; (ii) this work aims at jointly estimating both the user's travel experience and the interestingness of places using an unsupervised approach.

3. Problem Formulation

In this section, we formulate the social-aware interesting place finding problem as a maximum likelihood estimation problem. Consider a scenario where a group of M users, denoted by $U_1, U_2, ..., U_M$, who visit a set of N places, denoted by $P_1, P_2, .., P_N$. For simplicity, we focus on the binary case on the interestingness of a place (i.e., a place is either interesting or not) [2]. In particular, we let U_i denote the i^{th} user and P_k denote the k^{th} place. Furthermore, we let $P_k = I$ denote that place P_k is interesting and $P_k = \bar{I}$ denote that place P_k is not interesting. In our work, we define the interesting places as those in which average people may have a strong interest in visiting and sharing their check-in trace on social media (e.g., parks, museums, historic sites, scenic trails, etc.). Additionally, we define a *User-Place Matrix* H to reflect the visiting behavior of users. In particular, the element $H_{i,k} = 1$ when user U_i visits place P_k and $H_{i,k} = 0$ otherwise.

Furthermore, we explicitly consider the *social dependency* between users in our model. This is motivated by the observation that the visiting behavior of users is highly correlated with their social connections. For example, classmates are likely to visit the same school they attend and friends are likely to go to the same restaurant or bar together. Simply counting the visits from nonindependent users in the same way as independent users could easily lead to many false positives in the interesting place identification results. To address this problem, we need to explicitly model the user's social dependency in our interesting place finding problem. Therefore, we define a *User-Dependency Matrix* D to represent the social dependency between users. In particular, the elements $D_{ij} = 1$ if user U_i and user U_j are friends and $D_{ij} = 0$ otherwise. Note that D is a symmetric matrix as we only consider bi-directional friendship (e.g., friendship on Facebook) in this paper. It is trivial to extend our model and solution to handle

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

directional friendship as well. Using the D matrix, we can divide the whole set of users into C independent groups where users in the same independent groups have non zero components in D and users in different independent groups have zero components in D.

We formulate the social-aware interesting place finding problem as follows. First, we define a few important items that will be used in the problem formulation. If user U_i is an independent user (i.e., U_i has no social connections with other users), we denote the *travel experience* of user U_i by t_i, which is the probability that a place is interesting given that user U_i visits it. If user U_i is a non-independent user (i.e., U_i has social connections with other users), for a friend user U_j of U_i, we denote the *dependent travel experience* of U_i by $t_{i,j}$ where $t_{i,j}$ is the probability that a place is interesting and the friend U_j visits this place given that U_i visits it. Formally, t_i and $t_{i,j}$ are defined as:

$$\begin{aligned} t_i &= \Pr(P_k = I | H_{i,k} = 1) \\ t_{i,j} &= \Pr(P_k = I, H_{j,k} = 1 | H_{i,k} = 1) \end{aligned} \tag{1}$$

For independent users, let us further define T_i to be the probability that user U_i visits the place P_k given that the place is interesting, and let F_i be the probability that user U_i visits the place P_k given that the place is not interesting. For non-independent users, we define $T_{i,j}$ as the probability that user U_i visits the place P_k given that the place is interesting and his/her friend U_j also visits the place. Similarly, we also define $F_{i,j}$ as the probability that user U_i visits the place P_k given that the place is not interesting and his/her friend U_j also visits the place. Formally, T_i, $T_{i,j}$, F_i and $F_{i,j}$ are defined as follows:

$$\begin{aligned} T_i &= \Pr(H_{i,k} = 1 | P_k = I) \qquad & T_{i,j} &= \Pr(H_{i,k} = 1 | H_{j,k} = 1, P_k = I) \\ F_i &= \Pr(H_{i,k} = 1 | P_k = \bar{I}) \qquad & F_{i,j} &= \Pr(H_{i,k} = 1 | H_{j,k} = 1, P_k = \bar{I}) \end{aligned} \tag{2}$$

Additionally, we denote the prior probability that user U_i visits a place by s_i (i.e., $s_i = \Pr(H_{i,k} = 1)$) and denote d as the prior probability that a randomly chosen place is interesting (i.e., $d = \Pr(P_k = I)$). Based on Bayes' theorem, we

Table 2: The Summary of Notations

Description	Notation
Set of Users	U
Set of Places	P
User-Place Matrix	H
Set of Independent Group	C
User-Dependency Matrix	D
User Visit Probability	$s_i = \Pr(H_{i,k} = 1)$
Travel Experience (Independent Users)	$t_i = \Pr(P_k = I \mid H_{i,k} = 1)$
Travel Experience (Non-independent Users)	$t_{i,j} = \Pr(P_k = I, H_{j,k} = 1 \mid H_{i,k} = 1)$
Conditional Interesting Place Visiting Probability (Independent Users)	$T_i = \Pr(H_{i,k} = 1 \mid P_k = I)$
Conditional Interesting Place Visiting Probability(Non-independent Users)	$T_{i,j} = \Pr(H_{i,k} = 1 \mid H_{j,k} = 1, P_k = I)$
Conditional Non-interesting Place Visiting Probability (Independent Users)	$F_i = \Pr(H_{i,k} = 1 \mid P_k = \bar{I})$
Conditional Non-interesting Place Visiting Probability (Non-independent Users)	$F_{i,j} = \Pr(H_{i,k} = 1 \mid H_{j,k} = 1, P_k = \bar{I})$

have:

$$T_i = \frac{t_i \times s_i}{d}, \quad F_i = \frac{(1 - t_i) \times s_i}{(1 - d)}$$
$$T_{i,j} = \frac{t_{i,j} \times s_i}{t_j \times s_j}, \quad F_{i,j} = \frac{(1 - t_{i,j}) \times s_i}{(1 - t_j) \times s_j} \tag{3}$$

Table 2 summarizes the introduced notations.

Therefore, the social-aware interesting place finding problem can be formulated as a Maximum Likelihood Estimation (MLE) problem: given the User-Place Matrix H, the User-Dependency Matrix D, our goal is to estimate both the *interestingness of each place* and the *travel experience of each user*. Formally, we compute:

$$\forall k, 1 \le k \le N : \Pr(P_k = I \mid H, D)$$
$$\forall i, 1 \le i \le M : \Pr(P_k = I \mid H_{i,k} = 1) \tag{4}$$

4. Social-aware Interesting Place Finding

In this section, we solve the interesting place finding problem formulated in Section 3 by developing a Social-aware Interesting Place Finding Plus (SIPF+) scheme.

4.1. Likelihood Function Development

The Expectation Maximization (EM) algorithm is a commonly used optimization technique to find the maximum-likelihood estimates of parameters in a statistical model where data is incomplete [31]. We use the EM algorithm to solve the MLE problem formulated in the previous section because: (i) we do not know the interestingness of places a priori (i.e., incomplete data) and we use the latent variables in EM to model the unknown interestingness of those places; (ii) the likelihood function of our problem is intractable and we develop an EM based algorithm to derive the optimization solution.

To apply the EM algorithm to solve an MLE problem, we first need to define a likelihood function $L(\theta; X, Z) = p(X, Z|\theta)$, where θ denotes the parameter vector, X is the observed data, and Z represents the latent variables. The iterative computation of an EM algorithm mainly contains two steps: the expectation step (E-step) and the maximization step (M-step). In particular, the E-step estimates the conditional expectation of the latent variables Z and the M-step finds the parameters θ that maximize the expectation function in the E-step. Formally, they are given as:

$$\text{E-step: } Q(\theta|\theta^{(n)}) = E_{Z|x,\theta^{(n)}}[logL(\theta; x, Z)] \tag{5}$$

$$\text{M-step: } \theta^{(n+1)} = \arg\max_{\theta} Q(\theta|\theta^{(n)}) \tag{6}$$

In order to solve the MLE problem we formulated in the previous section, let us first define the likelihood function of our MLE problem. In the interesting place finding problem, the observed data is the *User-Place* Matrix H and the *User-Dependency* Matrix D. The estimation parameter $\theta = (T_1, ..., T_M; F_1, ..., F_M; T_{1,j}, ..., T_{M,j}; F_{1,j}, ..., F_{M,j}; d)$, where T_i, F_i, $T_{i,j}$, $F_{i,j}$ are defined in Equation (2) and d is defined in Equation (3). d reflects the prior probability a randomly chosen place is interesting. Moreover, 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 such that $z_k = 1$ if P_k is interesting and $z_k = 0$ otherwise. The variables calculated from the likelihood function include the estimation parameter θ and hidden variables. These variables can be used to estimate the interestingness of places and the travel experience of users.

Hence, the likelihood function of social-aware interesting place finding problem can be written as:

$$
\begin{aligned}
& L(\theta; X, Z) = \Pr(X, Z|\theta) \\
& = \prod_{k=1}^{N} \Bigg\{ \prod_{g \in C} \Bigg[\prod_{i \in g} (T_i^{H_{i,k}} (1 - T_i)^{(1-H_{i,k})})^{(|g|==1)} \\
& \prod_{j \in g} ((T_{i,j}^{H_{i,k} \ \&\& \ H_{j,k}} (1 - T_{i,j})^{(1-H_{i,k}) \ \&\& \ H_{j,k}})^{D_{i,j}})^{(|g|>1)} \Bigg] \times d \times z_k \\
& + \Bigg[\prod_{i \in g} (F_i^{H_{i,k}} (1 - F_i)^{(1-H_{i,k})})^{(|g|==1)} \\
& \prod_{j \in g} ((F_{i,j}^{H_{i,k} \ \&\& \ H_{j,k}} (1 - F_{i,j})^{(1-H_{i,k}) \ \&\& \ H_{j,k}})^{D_{i,j}})^{|g|>1} \Bigg] \times (1 - d) \times (1 - z_k) \Bigg\}
\end{aligned}
\tag{7}
$$

where $H_{i,k} = 1$ when user U_i visits place P_k and 0 otherwise. $D_{i,j} = 1$ when user U_i is a friend of U_j and 0 otherwise. $|g|$ denotes the size of the independent group g. The "&&" represents the logical "AND" for binary variables. The likelihood function represents the likelihood of the observed data (i.e., H and D) and the values of hidden variables (i.e., Z) given the estimation parameters (i.e., θ).

4.2. Social-aware Interesting Place Finding Scheme

Given the above mathematical formulation, we derive E and M steps of the proposed SIPF+ scheme. First, we derive the Q function for the E-step, given by Equation (5), using the likelihood function derived in Equation (7). The

E-step is given as follows:

$$
\begin{aligned}
&Q(\theta|\theta^{(n)}) = E_{Z|X,\theta^{(n)}}[logL(\theta; X, Z)] \\
&= \sum_{k=1}^{N} \sum_{g \in C} \Bigg\{ Z(n,k) \times \Bigg[(|g| == 1) \sum_{i \in g} ((H_{i,k} logT_i + (1 - H_{i,k}) log(1 - T_i)) \\
&+ (|g| > 1)(\sum_{j \in g} D_{i,j}((H_{i,k} \;\&\&\; H_{j,k}) logT_{i,j} + (1 - H_{i,k}) \;\&\&\; H_{j,k}) log(1 - T_{i,j})) + logd) \Bigg] \\
&+ (1 - Z(n,k)) \times \Bigg[(|g| == 1) \cdot \sum_{i \in g} ((H_{i,k} logF_i + (1 - H_{i,k}) log(1 - F_i)) \\
&+ (|g| > 1)(\sum_{j \in g} D_{i,j}((H_{i,k} \;\&\&\; H_{j,k}) logF_{i,j} \\
&+ (1 - H_{i,k}) \;\&\&\; H_{j,k}) log(1 - F_{i,j})) + log(1 - d)) \Bigg] \Bigg\}
\end{aligned} \tag{8}
$$

where $Z(n,k) = \Pr(z_k = 1|X_k, \theta^{(n)})$. It is the conditional probability of the place P_k to be interesting given the observed data X_k and current estimate of θ, where X_k represents the k^{th} column of the User-Place Matrix H.

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 $\frac{\partial Q}{\partial T_i} = 0$, $\frac{\partial Q}{\partial F_i} = 0$, $\frac{\partial Q}{\partial T_{i,j}} = 0$, $\frac{\partial Q}{\partial F_{i,j}} = 0$ and $\frac{\partial Q}{\partial d} = 0$ in each iteration, we can get expressions of the optimal T_i^*, F_i^*, $T_{i,j}^*$, $F_{i,j}^*$ and d^*:

$$
\begin{aligned}
&T_i^{(n+1)} = T_i^* = \frac{\sum_{k \in H_i} Z(n,k)}{\sum_{k=1}^{N} Z(n,k)} \qquad F_i^{(n+1)} = F_i^* = \frac{\sum_{k \in H_i} (1 - Z(n,k))}{\sum_{k=1}^{N} (1 - Z(n,k))} \\
&T_{i,j}^{(n+1)} = T_{i,j}^* = \frac{\sum_{k \in H_{i,j}} Z(n,k)}{\sum_{k \in H_j} Z(n,k)} \qquad F_{i,j}^{(n+1)} = F_{i,j}^* = \frac{\sum_{k \in H_{i,j}} (1 - Z(n,k))}{\sum_{k \in H_j} (1 - Z(n,k)}) \\
&d^{(n+1)} = d^* = \frac{\sum_{k=1}^{N} Z(n,k)}{N}
\end{aligned} \tag{9}
$$

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

5. Evaluation

In this section, we evaluate the performance of the SIPF+ scheme using two real-world datasets collected from location-based social network services. We first describe the experiment settings and data pre-processing steps. Then, we introduce the state-of-the-art baselines and evaluation metrics used in our experiments. Finally, we present the evaluation results of SIPF+ scheme in comparison to all baselines and demonstrate the effectiveness of explicitly considering the user's travel experience and social dependency in solving the interesting place finding problem in social sensing.

5.1. Experiment Settings

5.1.1. Dataset Statistics

In this evaluation, we use two different datasets, which are collected from location-based social network services, namely, Brightkite[3] and Gowalla[4] [4]. In the location-based social network services, users check in and share their location information using the following format: (user ID, latitude, longitude, timestamp). The Brightkite dataset was collected from April 2008 until October 2010 and the Gowalla dataset was collected from February 2009 until October 2010. Other statistics of these two datasets are shown in Table 3.

Table 3: Dataset 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

Table 4 shows the statistics on the percentage of check-in points at interesting places and non-interesting places from the two datasets in the city of Chicago and San Francisco, respectively. In Table 4, we can observe that users

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

check in more frequently at interesting places (e.g., parks, museums, historic sites) than non-interesting places (e.g., work place, transit centers). This observation provides the basis for our proposed model to find interesting places by analyzing the collective check-in behaviors of online social media users. We also observe that users check in at non-interesting places, as well, which makes the interesting place finding problem using social sensing data an interesting, but non-trivial, problem to solve. We show that our SIPF+ scheme can find interesting places more accurately than the state-of-the-art baselines through real-world experiments in the rest of this section.

Table 4: Percentage of Check-in Points at Interesting vs Non-interesting Places

Data Trace	Interesting Place	Non-interesting Place
Chicago - Brightkite	60.5%	39.5%
Chicago - Gowalla	63.2%	36.8%
San Francisco - Brightkite	59.8%	40.2%
San Francisco - Gowalla	67.2%	32.8%

5.1.2. Data Pre-Processing

To evaluate our method in real-world settings, we conducted data pre-processing in two steps: (i) clustering all raw geographical check-in points into meaningful clusters that represent places in the physical world; (ii) identifying independent groups from all users based on their social connections. Using the meta-data generated by the above steps, we can create the User-Place Matrix H and User-Dependency Matrix D we discussed in Section 3. In our evaluation, we select two popular tourist destinations, San Francisco and Chicago, as our target cities from the two real-world datasets.

For the clustering step, we used the K-means clutering algorithm [32] to first cluster the raw check-in records into intermediate clusters without any geospatial-semantic meanings allocated to those clusters. In the data pre-processing step, our goal is to identify raw clusters that can be identified as specific places in our model. In our experiments, we found that most of the user's check-in points were naturally centered on some locations in the city.

Therefore, we use a K-means clustering algorithm by taking the places with dense check-in points as centroids and minimizing the mean squared distance from each data point to its nearest centroid [33]. We compared the clustering outputs of the K-means algorithm with other clustering algorithms (e.g., DB-SCAN) and found that the results from K-means make the most sense (i.e., identify the largest number of places in a city). Then, we re-organized the raw clusters into meaningful places by referring to the Point-of-Interest information from Google Map [5]. For the Brightkite dataset, we identified 83 places in San Francisco, with 36 interesting places and 47 not interesting places. In Chicago, we identified 70 places in total, with 36 interesting places and 34 not interesting places. For the Gowalla trace dataset, we identified 92 places in San Francisco, with 39 interesting places and 53 not interesting places. In Chicago, we identified 74 places in total, with 39 interesting places and 35 not interesting places. As a result, we created the User-Place Matrix H by associating each user with the places the user visited.

For the independent group identification step, we used a community detection algorithm called SLPA [34] to find independent groups of users. We first obtain the social connections between users from the friendship information within the dataset. In particular, we generated the user dependency graph as an undirected graph $G = (V, E)$ where V and E represents the set of users and their friendship, respectively: if u is a friend of v in the dataset, 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. Using the output of this step, we generated the User-Dependency Matrix D.

5.2. Baselines and Evaluation Metric

5.2.1. Baselines

In the evaluation, we compare the performance of the *SIPF+* scheme with the following state-of-the-art baselines. The first baseline is *Voting*, which com-

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

putes the interestingness of a place simply by counting the number of times the place is visited. The second baseline is *Sums-Hubs* [35], which explicitly considers the difference in users' travel experience when it computes 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 social sensing data [10]. Additionally, we also compare the performance of SIPF+ with four recent Point-of-Interest recommendation solutions. The first baseline is *iGSLR* [15], which explored the influence that geographical proximity has on users' check-in behaviors when computing the interestingness of a place. The second baseline is *GeoSoCa* [16], which explored geographical, social, and categorical information for Point-of-Interest recommendations. The third baseline is *STT* [18], which captured the spatial and temporal aspects of check-ins to recommend locations. The fourth baseline is *GTM* [17], which developed a geo-topic model to consider the user's activity area in recommending interesting places.

5.2.2. 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 identifying interesting places. These metrics include *precision*, *recall*, and *F1-measure* [36]. The second set of metrics are used to evaluate the ranking performance of different schemes. [6] These metrics include *mean average precision (MAP)* [37] and *normalized discounted cumulative gain (NDCG)* [38].

5.3. Evaluation Results

In this section, we conduct experiments on two real-world datasets to compare *SIPF+* scheme with the above state-of-the-art baselines in terms of *estimation accuracy* and *ranking performance*. Independent from the two datasets we

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

used in evaluation, we collected ground truth values (i.e. whether a place is interesting) from four widely used travel recommendation websites: TripAdvisor, Planet Aware, San Francisco Travel, and CityPass. We then decide whether a place is interesting using the following rubric:

- *Interesting places:* Places that have been recommended by at least two of the above travel recommendation websites or manually verified by the researchers using external sources: online social media that are different from those used in this paper (e.g., Foursquare and Twitter). For example, if a place has been frequently reported as an interesting place on social media by average people, we will consider this place as an interesting place (even though this place might not be recommended by traditional travel websites due to the coverage and freshness limitations that we discussed in the introduction of the paper).
- *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.

5.3.1. 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 San Francisco using the Brightkite dataset are shown in Figure 1. We observe that the *SIPF+* outperforms all the compared baselines in terms of precision, recall, and F1-measure. The largest performance gain achieved by SIPF+ on precision over the best performed baseline (i.e., SIPF) is 6%. From the above results, we can observe that our extended model (i.e., SIPF+) outperforms the previous model (i.e., SIPF), as well as other state-of-the-art baselines. The performance gain is achieved by explicitly considering more complex social dependency (i.e., arbitrary user dependency graphs that include cycles) between users in the new SIPF+ scheme. The results on Gowalla dataset are shown in Figure 2. Again, we observe that SIPF+ continues to outperform other baselines and the largest performance gain achieved by *SIPF+* on recall, compared to the best performed baseline, is 3%. We repeated the same experiments for the city of Chicago. The results for Brightkite and Gowalla are shown in Figure 3 and 4, respectively. We observe that SIPF+ continue to outperform all compared baselines.

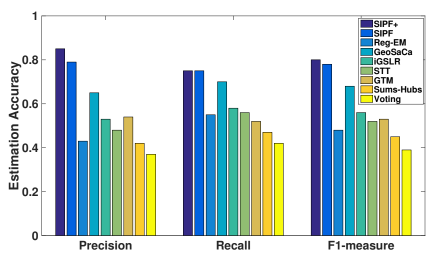

Figure 1: Estimation Accuracy (San Francisco, Brightkite Dataset)

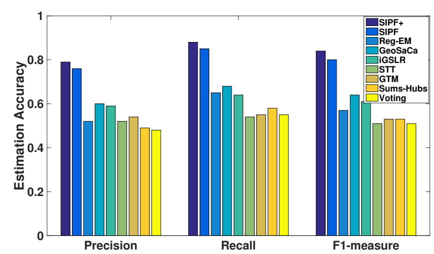

Figure 2: Estimation Accuracy (San Francisco, Gowalla Dataset)

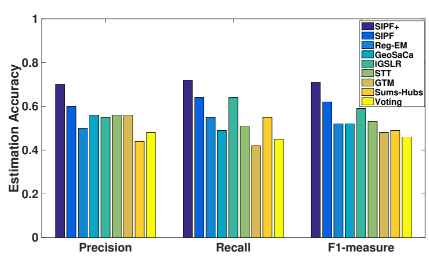

Figure 3: Estimation Accuracy (Chicago, Brightkite Dataset)

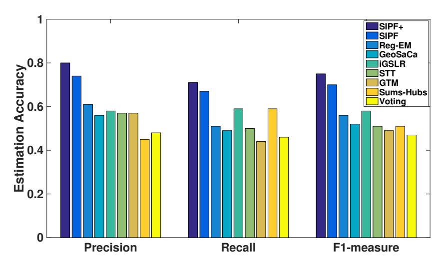

Figure 4: Estimation Accuracy (Chicago, Gowalla Dataset)

5.3.2. Ranking Performance

We also evaluate the ranking performance of all schemes and use *MAP* [37], *NDCG@10, NDCG@15, and NDCG@20* [38] as the evaluation metrics. In Figure 5 and Figure 6, we compare the performance of SIPF+ to all baselines, in terms of MAP, on the Brightkite dataset and the Gowalla dataset, respectively.

We observe that *SIPF+* achieves the highest MAP score compared to all baselines. In Figure 7 and Figure 8, we compare the performance of SIPF+ to all baselines, in terms of NDCG@n, on two datasets, respectively. We observe that *SIPF+* continues to outperform all baselines at different values of n. These results demonstrate that SIPF+ achieves the best ranking performance among all compared schemes. Similarly, the results for Chicago, in terms of MAP, are shown in Figure 9 and Figure 10. The results for Chicago, in terms of NGCDn, are shown in Figure 11 and Figure 12. We consistently observe that SIPF+ achieves the best performance compared to all other baselines.

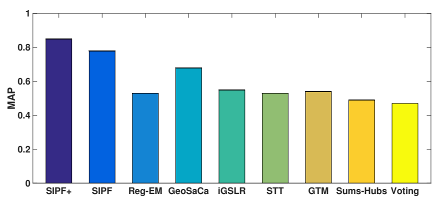

Figure 5: MAP Evaluation (San Francisco, Brightkite Dataset)

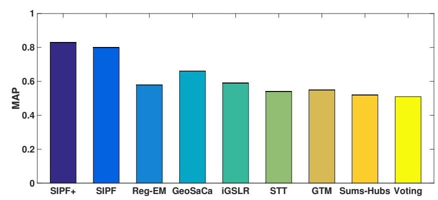

Figure 6: MAP Evaluation (San Francisco, Gowalla Dataset)

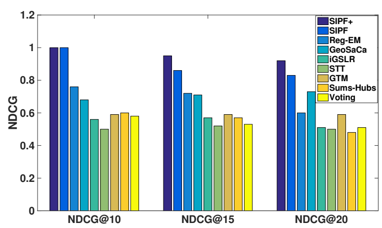

Figure 7: NDCG@n Evaluation (San Francisco, Brightkite Dataset)

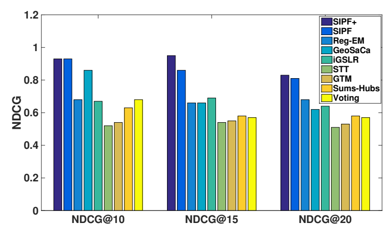

Figure 8: NDCG@n Evaluation on San Francisco in Gowalla Dataset

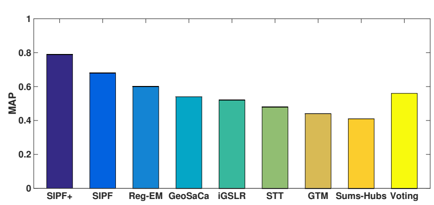

Figure 9: MAP Evaluation (Chicago, Brightkite Dataset)

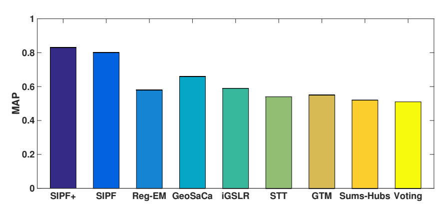

Figure 10: MAP Evaluation (Chicago, Gowalla Dataset)

In addition to the above quantitative analysis, we also present the *top 10* interesting places identified by all schemes from the Brightkite dataset in San

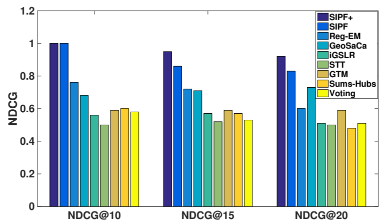

Figure 11: NDCG@n Evaluation (Chicago, Brightkite Dataset)

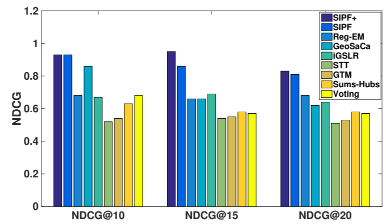

Figure 12: NDCG@n Evaluation (Chicago, Gowalla Dataset)

Francisco in Table 5. Note that the interesting places (i.e., true positives) are *highlighted* in the table. We observe all the top ten places found by *SIPF+* in the Brightkite dataset are, indeed, interesting. However, quite a few top places found by the best performing baseline are not really interesting. The performance improvement achieved by SIPF+ can help users schedule their visits more efficiently and save their time and money from making unnecessary trips to places that are not interesting. Results from the Gowalla dataset are similar and we do not report them due to limited space.

#	Interesting Places found by SIPF+	Interesting Places found by the Best Baseline
1	**San Francisco Museum of Modern Art**	**AT&T Park**
2	**Chinatown in San Francisco**	**Contemporary Jewish Museum**
3	**Union Square in San Francisco**	**Ghirardelli Square**
4	**AT&T Park**	**Coit Tower**
5	**San Francisco Fisherman's Wharf**	Food Court Restaurants near Cabrillo Street
6	**The Cable Car Museum**	**Aquarium of the Bay**
7	**Aquarium of the Bay**	Marina Middle School
8	**Coit Tower**	**Japanese Tea Garden**
9	**Japanese Tea Garden**	Kaiser Permanente Medical Center
10	**San Francisco City Hall**	**San Francisco Ferry Building**

Table 5: Top 10 Interesting Places Found by SIPF+ and Best Performed Baselines in the Brightkite Dataset in San Francisco (Highlighted Places are Truth Positives)

Finally, we evaluate the execution time and memory cost of all compared algorithms. We run all algorithms on a regular laboratory computer (4 cores and

2 GHZ for each core, 8GB memory). Table 6 and Table 7 show the execution time and memory cost of all algorithms on the Brightkite and the Gowalla dataset, respectively. From those results, we observe that the execution time of our proposed *SIPF+* is comparable to most of the non-trivial baselines and the memory overhead is small and affordable on a regular laboratory computer.

Table 6: Execution Time and Memory Cost Evaluation on Brightkite Datasets

	Chicago		San Francisco	
Algorithm	Execution Time (s)	Memory Cost (KB)	Execution Time (s)	Memory Cost (KB)
SIPF+	5.43	6583	5.51	6941
SIPF	5.02	6560	5.36	6852
Reg-EM	1.36	6468	1.54	6786
GeoSaCa	7.54	6908	7.73	8876
iGSLR	5.56	6528	6.30	6984
STT	5.52	6644	6.40	7512
GTM	2.43	6592	2.57	6652
Sums-Hubs	1.14	6320	1.23	6732
Voting	0.54	5748	0.65	6348

Table 7: Execution Time and Memory Cost Evaluation on Gowalla Datasets

	Chicago		San Francisco	
Algorithm	Execution Time (s)	Memory Cost (KB)	Execution Time (s)	Memory Cost (KB)
SIPF+	5.72	7713	6.01	7801
SIPF	5.35	7662	5.84	7702
Reg-EM	1.48	7600	1.74	7656
GeoSaCa	9,82	9656	10.03	9056
iGSLR	5.82	7844	6.96	7863
STT	5.61	7500	6.52	7749
GTM	2.51	7968	2.81	7541
Sums-Hubs	1.25	7332	1.31	7456
Voting	0.64	6860	0.71	6986

6. Discussions and Future Work

An important assumption made in our framework is that all places are assumed to be independent. However, some places may have underlying dependency imposed by the physical world. One example is the location dependency. For example, many national museums in Washington D.C. are located in the same area. Users who visit one museum are also likely to visit others. However, they may not provide check-in points at each museum they visit. Ignorance of such physical dependency between places is likely to generate suboptimal results in finding all interesting places. Hence, the following problem becomes interesting to investigate: how can we incorporate the dependency between places appropriately into our framework so that the estimation accuracy in finding interesting places can be further improved? We have successfully applied the MLE framework to handle non-independent variables in cyber-physical system (CPS) applications [39]. We believe that similar insights could be leveraged to effectively address the aforementioned place dependency problem.

The only input to the SIPF+ scheme is the User-Place Matrix H and User-Dependency Matrix D. This simple requirement on the input makes the proposed method very robust and generally applicable to different application scenarios. However, we might still be able to improve the performance of SIPF+ if additional information about users (e.g., user's travel experience, home city, working place, etc.) and places (e.g., construction time, historic background, etc.) is known to the application. For example, by knowing some of the users' travel experience a priori, we can initialize the SIPF+ scheme with a better start point (compared to a random start point). This will greatly expedite the convergence process of the EM algorithm and improve the response time of SIPF+ in large scale social sensing applications. The key challenge here is how to incorporate the additional information into the proposed model without sacrificing the rigidity of the analytical framework. We are actively working on the above extensions.

7. Conclusion

This paper develops a new social-aware maximum likelihood estimation framework to accurately identify interesting places in a city through the social sensing application paradigm. The proposed SIPF+ scheme explicitly incorporates both the user's *travel experience* and *social relationship* into a rigorous analytical framework. The proposed approach jointly estimates both the user's travel experience and the interestingness of a place using an Expectation Maximization algorithm. Compared to the traditional travel recommendation websites, the SIPF+ can find more interesting places that are visited by common citizens. We evaluated the SIPF+ scheme on two real-world datasets collected from location-based social network services. The results showed that the SIPF+ scheme achieved non-trivial performance gains in identifying more interesting places, while simultaneously lowering the number of not interesting places misidentified as interesting, when compared to the state-of-the-art baselines. The results of this paper are important because they lay out an analytical foundation to improve the accuracy in interesting place finding by using a principled approach.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. CBET-1637251, CNS-1566465 and IIS-1447795 and Army Research Office under Grant W911NF-16-1-0388. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] D. Wang, T. Abdelzaher, L. Kaplan, Social Sensing: Building Reliable Systems on Unreliable Data, Morgan Kaufmann, 2015.

[2] Y. Ge, Q. Liu, H. Xiong, A. Tuzhilin, J. Chen, Cost-aware travel tour recommendation, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011, pp. 983–991.

[3] M. Harding, J. Finney, N. Davies, M. Rouncefield, J. Hannon, Experiences with a social travel information system, in: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, ACM, 2013, pp. 173–182.

[4] E. Cho, S. A. Myers, 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, ACM, 2011, pp. 1082–1090.

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

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

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

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

[9] Y. Zheng, X. Xie, W.-Y. Ma, Geolife: A collaborative social networking service among user, location and trajectory., IEEE Data Eng. Bull. 33 (2) (2010) 32–39.

[10] D. Wang, L. Kaplan, H. Le, 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), 2012.

[11] C. Huang, 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.

[12] S. Khetarpaul, R. Chauhan, S. Gupta, L. V. Subramaniam, 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, ACM, 2011, p. 8.

[13] R. R. Vatsavai, A. Ganguly, V. Chandola, A. Stefanidis, S. Klasky, 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, ACM, 2012, pp. 1–10.

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

[15] J.-D. Zhang, 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, ACM, 2013, pp. 334–343.

[16] J.-D. Zhang, 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, ACM, 2015, pp. 443–452.

[17] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya, 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, ACM, 2013, pp. 375–384.

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

[19] E. Cambria, B. White, T. S. Durrani, N. Howard, Computational intelligence for natural language processing (2014).

[20] A. Gangemi, V. Presutti, D. Reforgiato Recupero, Frame-based detection of opinion holders and topics: a model and a tool, Computational Intelligence Magazine, IEEE 9 (1) (2014) 20–30.

[21] R. Y. Lau, Y. Xia, Y. Ye, A probabilistic generative model for mining cybercriminal networks from online social media, Computational Intelligence Magazine, IEEE 9 (1) (2014) 31–43.

[22] T. Denoeux, Maximum likelihood estimation from uncertain data in the belief function framework, Knowledge and Data Engineering, IEEE Transactions on 25 (1) (2013) 119–130.

[23] M. Bańbura, M. Modugno, Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data, Journal of Applied Econometrics 29 (1) (2014) 133–160.

[24] T. Denoeux, Likelihood-based belief function: justification and some extensions to low-quality data, International Journal of Approximate Reasoning 55 (7) (2014) 1535–1547.

[25] E. Tsakonas, J. Jalden, N. Sidiropoulos, B. Ottersten, Sparse conjoint analysis through maximum likelihood estimation, Signal Processing, IEEE Transactions on 61 (22) (2013) 5704–5715.

[26] K. Li, W. Lu, S. Bhagat, L. V. Lakshmanan, C. Yu, On social event organization, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2014, pp. 1206–1215.

[27] J. She, Y. Tong, L. Chen, Utility-aware social event-participant planning, in: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, 2015, pp. 1629–1643.

[28] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, J. Han, Event-based social networks: linking the online and offline social worlds, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012, pp. 1032–1040.

[29] J. She, Y. Tong, L. Chen, C. Cao, Conflict-aware event-participant arrangement and its variant for online setting, Knowledge and Data Engineering, IEEE Transactions on.

[30] Y. Tong, J. She, R. Meng, Bottleneck-aware arrangement over event-based social networks: the max-min approach, World Wide Web (2015) 1–27.

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

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

[33] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu, An efficient k-means clustering algorithm: Analysis and implementation, IEEE transactions on pattern analysis and machine intelligence 24 (7) (2002) 881–892.

[34] J. Xie, B. K. Szymanski, 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, IEEE, 2011, pp. 344–349.

[35] J. M. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM (JACM) 46 (5) (1999) 604–632.

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

[37] R. Baeza-Yates, B. Ribeiro-Neto, et al., Modern information retrieval, Vol. 463, ACM press New York, 1999.

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

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