Towards Unsupervised Home Location Inference from Online Social Media

Chao Huang, Dong Wang, Shenglong Zhu, Daniel (Yue) Zhang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
chuang7@nd.edu, dwang5@nd.edu, szhu3@nd.edu, yzhang40@nd.edu

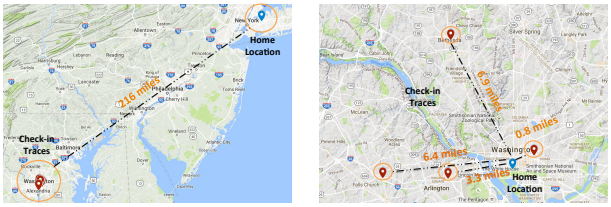
***Abstract*—Users' home location is important information for many advanced information services in big data applications (e.g., localized recommendation, target ads of local business and urban planning). In this paper, we study the problem of accurately inferring the home locations of people from the noisy and sparse data they voluntarily share on online social media. Previous studies have developed supervised learning approaches to predict a person's home location in a city. However, the accuracy of these techniques largely depends on a high quality training dataset, which is difficult and expensive to obtain in practice. In this study, we propose a new analytical framework, *Unsupervised Home Location Inference (UHLI)*, to accurately infer the home locations of people using a set of principle approaches. In particular, the UHLI scheme addresses the critical challenges of using sparse and noisy online social media data and derives an optimal solution to the home location inference problem. We evaluated the performance of our scheme and compared it to the state-of-the-art baselines using three real world data traces collected from Foursquare. The results showed that our scheme can accurately infer the home location of people and significantly outperform the state-of-the-art baselines.**

***Index Terms*—Home Location Inference,Unsupervised Learning,Social Media**

I. Introduction

Location-Based Social Network (LBSN) services have received increasing amount of attention due to the proliferation of portable devices with GPS sensors (e.g., smartphones) and the advent of online social media with location sharing functions (e.g., Foursquare, Gowalla, Google Places). Users can easily geotag their activities by sharing their "check-in" traces (a sequence of time-stamped GPS coordinates) on LBSN. In this paper, we study the problem of accurately inferring the home locations of people from the noisy and sparse data they contribute on online social media. User's home location is important information for many advanced information services in big data applications such as localized recommendations [31], targeted ads of local business [2], and urban planning [6].

There exist prior studies on geo-locating people in a city using online social network information [19], [3], [5], [8], [18], [26], [22]. Most of these previous studies used *supervised* learning approaches, which largely depend on high quality training datasets to predict a person's home location. However, such training datasets are difficult and expensive to obtain in practice since people usually are reluctant to publicize their real home locations [21]. For example, only 31.8% of users provide the home locations in the datasets we collected from Foursquare. Furthermore, since most of LBSNs have set up rate limits on their APIs for data collection and sharing [41], it is challenging to collect complete check-in traces of users at all venues they visited. To overcome the above challenges, this paper develops an *unsupervised* learning approach to accurately infer the home locations of users by exploiting the check-in points they voluntarily uploaded at their visited venues from LBSNs.

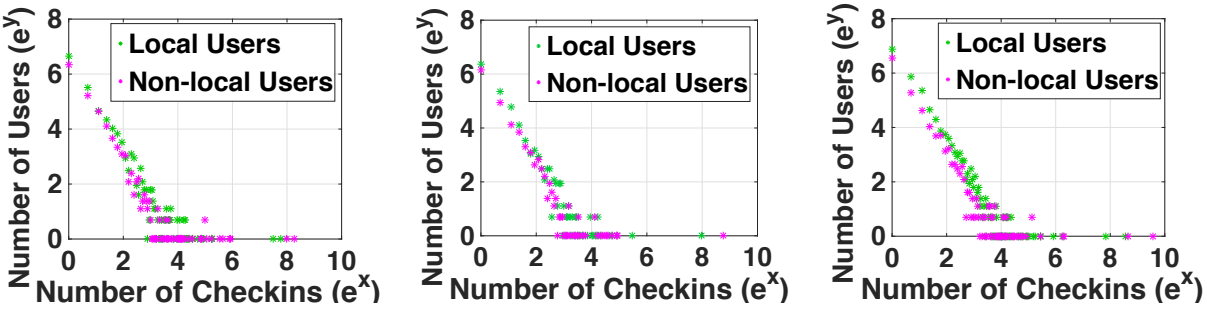


(a) User and Venues are in the Different Cities (b) User and Venues are in the Same City

Figure 1. Geographical Check-in Distribution of Random Users

A simple method of inferring a user's home location is to take the average of all venue locations the user visited as the estimated home location of the user. To check the accuracy of this simple method, we compute the average distance error between the estimated home location and the real home location of users over real world datasets collected from three cities (i.e., Washington D.C., Boston and Chicago) on Foursquare. The results show that the average estimation error of the simple average method is *260 miles*, *309 miles* and *150 miles* on the three datasets respectively. Such large estimation errors indicate the simple average method cannot accurately estimate the user's home location. The reasons are mainly twofold: (i) users might visit venues that are not in the same city as they live in (e.g., tourists); (ii) users might visit venues that are in the same city as they live in but are far away from their home locations. Figure 1 shows examples of the above two scenarios.

Another important aspect that might affect the home location inference is the influence scope of venues the user visited. We observe that a more popular venue is often less influential



(a) Washington, D.C. (b) Boston (c) Chicago

Figure 2. Distribution of Check-in Points Per User

(useful) to infer the home location of its visiting users due to the fact that many users from different places might visit it. For example, a 7-Eleven store is more useful than the White House for inferring the home locations of users and hence considered to have a large influence scope. In this paper, we develop an *unsupervised* user home location inference scheme that jointly explores both the localness of a user (i.e., whether the user is a local resident of a city or not) and the influence scope of a venue (i.e., how influential a venue is in terms of estimating the user's home location).

Two key challenges exist to solve the unsupervised home location inference problem: (i) *Sparse Data Challenge*: the spatial-temporal data (i.e., check-in points) are often incomplete and highly sparse: a person might not check in at every venue he/she visits in a city; (ii) *Noisy Data Challenge*: the collected data is "noisy" in the sense that a venue might have check-in points from both local and non-local people. We observed that local and non-local users have very similar distributions on all these dimensions, which makes it a challenging task to separate local users from non-local ones solely based on their check-in traces. In fact, the check-in points are just GPS coordinates with timestamps, which themselves do not provide much useful information to infer the home location of the user.

To address the above challenges, this paper develops a new analytical framework, *Unsupervised Home Location Inference (UHLI)* scheme. The UHLI scheme explicitly explores the spatial (i.e., venues a user visited and her/his activity range), temporal (i.e., the length of a user's check-in trace) and social (i.e., the social connections of a user) dimensions of the problem. In particular, we develop two new Expectation Maximization (EM) based algorithms, Location-Aware EM (LA-EM) and Influence-Aware EM (INA-EM), to jointly estimate the localness of users and the influence scope of venues in a city. We further formulate the home location inference problem as an optimization problem that leverages the localness of users and influence scope of venues and derive an optimal solution. Finally, we evaluate the UHLI scheme using three real-world datasets collected from Foursquare. The results showed that our scheme can accurately infer the home location of people and significantly outperform the state-of-the-art baselines.

Finally, a note on disclaimer. First, we did not discuss the privacy issue in this paper because the user identities in collected datasets from LBSNs are all anonymized [7]. Additionally, there exists a rich set of literature on the topic of protecting user's privacy in online social media applications [39], [16]. These works can be used to address the privacy challenges if there is such a need. Second, we did not use any private data from a third party (e.g., Google Map search data, which could make the home location inference problem a trivial problem to solve). Instead, we only used publicly available data from LBSNs with the goal to develop a new unsupervised user home location inference scheme as an open-source resource for the research community.

In summary, our contributions are as follows:

- This paper addresses the problem of inferring the home location of people from sparse and noisy online social media data using an *unsupervised* approach. (Section III)
- We develop a principled framework (i.e., UHLI scheme) that allows us to derive an *optimal* solution to accurately infer the home location of users by jointly estimating the localness of people and the influence scope of venues. (Section IV)
- We perform extensive experiments to compare the performance of our UHLI scheme with the-sate-of-the-art baselines using three large scale real-world data sets collected from Foursquare. The results demonstrate that the proposed approach outperforms existing methods in terms of estimation accuracy. (Section V)

The rest of this paper is organized as follows: we review related work in Section II. In Section III, we present the problem formulation of inferring the home location of people. The proposed unsupervised home location inference scheme is discussed in Section IV. Experiment and evaluation results are presented in Section V. Finally, we conclude the paper in Section VI.

II. Related Work

Previous work has made significant progress on user profiling [27], [1], [15], [17]. For example, Mislove et al. proposed a community detection approach to infer the missing attributes of a user on Facebook from the attributes of his/her friends in the network [27]. Abel et al. developed a semantic approach

to construct the user's profile on Twitter by exploiting the links between the user's tweets and related news articles [1]. Lampe et al. investigated how different elements in the user's profile influence the formation of online connections [15]. Li et al. studied the problem of user profiling by capturing the correlation between attributes and social connections of the user's ego networks. However, none of these techniques can be directly used to infer the home location of people in a city because i) people may have social connections with friends living far away; ii) people may also report news/events that are not local to the city they live. In this paper, we solve the problem of inferring the home location of users by jointly estimating the localness of users and influence scope of venues they visited.

Our work is also related to user behavior understanding based on their home locations and digital traces [5], [30], [32], [38], [37]. For example, a content-based approach was proposed by Cheng et al. [5] to identify Twitter users' home cities and their movement patterns. Specifically, they extract a set of words which are related to a city (e.g., Washington D.C) and use those words as features to classify users to different cities. Home location was also used to model people's living conditions and lifestyles in [30]. Furthermore, user's home location has been considered as a key factor to compute the distance between social users in a pairwise fashion [32]. Our work is complementary to the above works in the sense that more accurate estimations of users' home location normally lead to a better understanding of user's behavior and movement patterns in a city.

Our work is closely related to the works that directly address the user's location inference problem [8], [18], [3], [19], [26], [22]. In particular, Backstrom et al. estimated a user's location by exploring both the geographic and social relationship between users [3]. Li et al. [19] developed a system to infer a user's location by integrating network and user-centric data via a unified influence model. They further extended their model to handle cases where users have multiple home locations [18]. McGee [26] proposed a network-based approach for location estimation by correlating the social tie strength with physical proximity. Hu et al. [8] designed a machine learning method to capture the inherent properties of users' homes by exploring their mobility features. Mahmud et al. [22] proposed a hierarchical ensemble algorithm to predict the home location of users by leveraging the domain knowledge and advanced classifications. However, the above solutions all used *supervised learning* approaches, which need sufficient training data with complete spatial-temporal information to accurately estimate an individual's home location. In contrast, this paper develops an *unsupervised approach* to address the problem of inferring the home location of people which does not require any training data.

Our work is also related to the work on social media analysis [20], [12], [9], [36], [11], [13], [10], [14] and information inference [40], [34], [33], [25], [24], [23]. In particular, Lu et al. analyzed the underlying connections between promotion campaigns in social media. Huang et al. studied the problem of discovering interesting places in a city from location-based social media [12], [9]. Different features (e.g., topic relevance, source dependency and time-sensitive information) have been analyzed to detect trustworthy information on social media [11], [36], [13]. A new framework has been proposed to infer the user's localness and venue's local attractiveness [10]. Emre [14] focused on identifying action-outcome relationships from social media data. This paper leveraged the insights of the above social media analysis work and addressed a new problem of inferring home location of users using online social media data.

III. Problem Formulation

In this section, we introduce the problem of inferring the home location of people from a LBSN application. In particular, we consider a LBSN application where a set of X venues (i.e., $V_1, V_2, ..., V_X$) have been visited by a group of Y users (i.e., $U_1, U_2, ..., U_Y$). Here we define V_x to be the x^{th} venue and U_y to be the y^{th} user. $U_y = 1$ if the user is a local resident of the city and $U_y = 0$ if she/he is not. We further define the following inputs to our model.

Definition 1. **Venue-User Matrix VU**. We define Venue-User Matrix $VU_{X\times Y}$ to indicate *which venue is visited by which user*. In particular, $VU_{x,y} = 1$ indicates that user U_y has check-in points at venue V_x and $VU_{x,y} = 0$ otherwise.

Definition 2. **Check-in Matrix CI**. We define Check-in Matrix $CI_{Y\times X}$ to be the transpose of $VU_{X\times Y}$ (i.e., $CI_{Y\times X} = \mathbf{VU}^T_{X\times Y}$).

Definition 3. **Temporal Vector T**. We define a Time Vector T_Y to represent the time length of user's check-in points (i.e., the time difference between the first and last check-in points in the dataset). In particular, $t_y = d$ denotes that user U_y's check-in points in a city lasts for d days.

Definition 4. **Spatial Vector S**. We define a Spatial Vector S_Y to represent the activity range of user's check-in points. In particular, $s_y = h$ denotes that the largest pairwise distance between all check-in points of the user U_y is h miles.

Definition 5. **Social Relationship Matrix SR**. We define a Social Relationship Matrix $SR_{Y\times Y}$ to represent the social connections between users. In particular, $SR_{y,y'} = 1$ if there exists social connection between two users $U_y, U_{y'}$ and $SR_{y,y'} = 0$ otherwise.

One key challenge of using LBSN data lies in the fact that both local and non-local people (e.g., tourist) can have check-in points in a city, and people may have check-in points in the places both near and far away from their home locations. To address such challenge, we develop two new models (i.e., *User Localness Model (ULM)* and *Venue Influence Model (VIM)*) based on the Maximum Likelihood Estimation (MLE) principle. In particular, the ULM estimates the localness of

people (i.e., whether a user is a local resident of a city or not) from massive check-in points contributed by the crowd. The VIM identifies the influence scope of each venue (i.e., how influential a venue is in terms of estimating a user's home location). The outputs of the two estimation models (i.e., the localness of users and venues' influence scope) are further taken as inputs into the Home Location Estimation (HLE) scheme to accurately infer the home location of users.

A. User Localness Model

We denote the *local attractiveness* of a venue V_x as ε_x, which is the probability that a user is local given that the user has check-in points at the venue V_x. Furthermore, considering a user may have different time length of her/his check-in points, we define $la_{x,d}$ as the probability of a venue V_x to attract local users whose check-in points in a city last for d days. Formally, ε_x and $\varepsilon_{x,d}$ can be given as:

$$\begin{aligned}\varepsilon_x &= \Pr(U_y = 1 | VU_{x,y} = 1) \\ \varepsilon_{x,d,h} &= \Pr(U_y = 1 | VU_{x,y} = 1, t_y = d, s_y = h)\end{aligned} \quad (1)$$

We denote the prior probability that venue V_x is visited by a user whose check-in points lasts for d days and activity range is h miles by $r_{x,d}$. The relationship between ε_x and $\varepsilon_{x,d,h}$ can be expressed as:

$$\varepsilon_x = \sum_{d=1}^{D}\sum_{h=1}^{H} \varepsilon_{x,d,h} \times \frac{r_{x,d,h}}{r_x} \quad d \in [1, D]; h \in [1, H] \quad (2)$$

where $r_x = \Pr(VU_{x,y} = 1)$ and $r_{x,d,h} = \Pr(VU_{x,y} = 1, t_y = d, s_y = h)$. Note that $r_x = \sum_{d=1}^{D}\sum_{h=1}^{H} r_{x,d,h}$.

We further denote $M_{x,d,h}$ as the probability of a *local* user (whose check-in points in a city lasts for d days and activity range is h miles) visits a venue V_x. Similarly, we denote $N_{x,d,h}$ as the probability of a *non-local* user (whose check-in points in a city lasts for d days and activity range is h miles) visits a venue V_x. $M_{x,d,h}$ and $N_{x,d,h}$ are formally defined below:

$$\begin{aligned}M_{x,d,h} &= \Pr(VU_{x,y} = 1, t_y = d, s_y = h | U_y = 1) \\ N_{x,d,h} &= \Pr(VU_{x,y} = 1, t_y = d, s_y = h | U_y = 0)\end{aligned} \quad (3)$$

We define the prior probability that a randomly chosen user is local as q, Using Bayes' theorem, we have:

$$\begin{aligned}M_{x,d,h} &= \frac{\varepsilon_{x,d,h} \times r_{x,d,h}}{q} \\ N_{x,d,h} &= \frac{(1 - \varepsilon_{x,d,h}) \times r_{x,d,h}}{1 - q}\end{aligned} \quad (4)$$

B. Venue Influence Model

We observe that venues have different influence scopes that actually affect the inference of a user's home location. In particular, we define Inf_x and Pop_x to be the influence scope and popularity of a venue V_x respectively such that $Inf_x + Pop_x = 1$.

To estimate the popularity of a venue, we define the following terms: te_y is defined as the *travel experience* of user U_y, which is the probability that a venue V_x is popular given that the user visits V_x. Furthermore, $V_x = P$ denotes that venue V_x is popular and $V_x = \bar{P}$ denotes that venue V_x is not popular. Formally, te_y is defined as follows: $te_y = \Pr(V_x = P | CI_{y,x} = 1)$.

We further define a few relevant conditional probabilities: I_y and J_y are defined as the probability that U_y visits a venue V_x given that V_x is popular (or not) respectively. Formally, I_y and J_y are defined as:

$$\begin{aligned}I_y &= \Pr(CI_{y,x} = 1 | V_x = P) \\ J_y &= \Pr(CI_{y,x} = 1 | V_x = \bar{P})\end{aligned} \quad (5)$$

Observing that users may visit different numbers of places, we denote the probability that user U_y visits a place by p_y (i.e., $p_y = \Pr(CI_{y,x} = 1)$) where V_x is a randomly chosen venue. We further denote b as the prior probability that a randomly chosen venue is popular (i.e., $b = \Pr(V_x = P)$). Using the Bayes' theorem, we can obtain the relationship between the items defined above:

$$\begin{aligned}I_y &= \frac{te_y \times p_y}{b} \\ J_y &= \frac{(1 - te_y) \times p_y}{(1 - b)}\end{aligned} \quad (6)$$

Using the above definitions, the problem of inferring home location of users is formulated as follows: given the Venue-User Matrix $\mathbf{VU}_{X\times Y}$, Check-in Matrix $\mathbf{CI}_{Y\times X}$, Time Vector $\mathbf{T}_Y$, Spatial Vector $\mathbf{S}_Y$ and Social Relationship Matrix $\mathbf{SR}_{Y\times Y}$, the goal is to accurately infer the home location of users. Here, α_u and β_u are defined as the latitude and longitude of user u's home location. Formally, we compute:

$$(\alpha_u, \beta_u | VU, CI, T, S, SR) \quad \forall u, 1 \le u \le Y \quad (7)$$

IV. Solution

In this section, we present our solution *Unsupervised Home Location Inference (UHLI)* scheme to infer people's home location by exploring the localness of users and influence scope of venues as we discussed in the previous section. The UHLI scheme consists of three major components: *User Localness Identification*, *Venue Influence Inference* and *Home Location Estimation*, which are shown in Figure 3. We will explain these three components in detail in the following subsections.

A. User Localness Identification

In this subsection, we present the user localness identification scheme: Localness-Awareness Expectation Maximization (LA-EM). The objective of the LA-EM scheme is to identify local users from non-local ones by using the VU matrix, T vector and S vector.

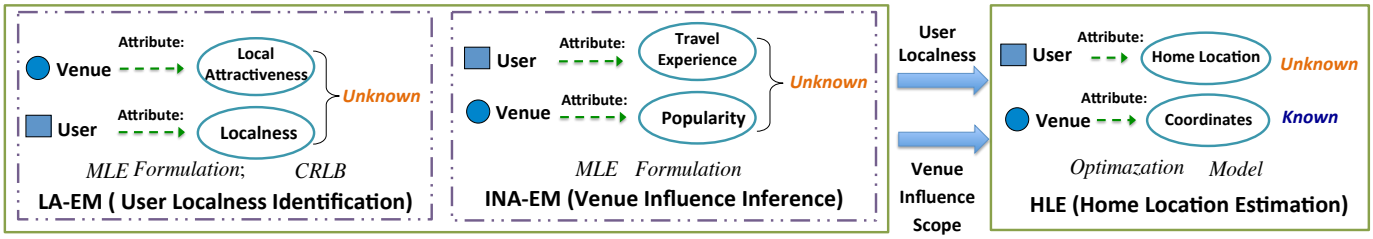


Figure 3. The UHLI Framework

1) The Likelihood Function: Given the terms and variables we defined in Section III, the likelihood function $L = (\Theta_{la}; O, \Lambda)$ for LA-EM is as follows:

$$
\begin{aligned}
&L(\Theta_{la}; O, \Lambda) = \Pr(O, \Lambda | \Theta_{la}) \\
&= \prod_{y=1}^{Y} \Big(\Lambda(n,y) \times \\
&\Big[\prod_{x=1}^{X} \prod_{d=1}^{D} \prod_{h=1}^{H} (M_{x,d,h})^{VU_{x,y} \;\&\&\; (t_y=d) \;\&\&\; (s_y=h)} \times \Pr(\lambda_y) \Big] \\
&+ \Big(\Lambda(n,y) \times \Big[\prod_{x=1}^{X} \prod_{d=1}^{D} \prod_{h=1}^{H} (1 - \sum_{d=1}^{D} \sum_{h=H} M_{x,d,h}) \times \Pr(\lambda_y) \Big] \\
&+ \Big(\Lambda(n,y) \times \\
&\Big[\prod_{x=1}^{X} \prod_{d=1}^{D} \prod_{h=1}^{H} (N_{x,d,h})^{VU_{x,y} \;\&\&\; (t_y=d) \;\&\&\; (s_y=h)} \times \Pr(\lambda_y) \Big] \\
&+ \Big(\Lambda(n,y) \times \Big[\prod_{x=1}^{X} \prod_{d=1}^{D} \prod_{h=1}^{H} (1 - \sum_{d=1}^{D} \sum_{h=H} N_{x,d,h}) \times \Pr(\lambda_y) \Big] \Big)
\end{aligned}
\tag{8}
$$

where $\Theta_{la} = (M_{1,d,h}, ..., M_{X,d,h}; N_{1,d,h}, ..., N_{X,d,h}; q)$. O is the observed data (i.e., Matrix VU, Vector T and S). Λ is a set of latent variables that indicate whether a user is local or not. More specially, we define a corresponding variable λ_y for each user U_y such that $\lambda_y = 1$ if U_y is local and $\lambda_y = 0$ otherwise. $\Lambda(n,y)$ and $\Pr(\lambda_y)$ are defined as follows:

$$
\Lambda(n,y), \Pr(\lambda_y) = \begin{cases} \Pr(U_y = 1 | O_y, \Theta_{la}^{(n)}), & q & \lambda_y = 1 \\ \Pr(U_y = 0 | O_y, \Theta_{la}^{(n)}), & (1-q) & \lambda_y = 0 \end{cases} \tag{9}
$$

2) The LA-EM Scheme: Given the above likelihood function, we can derive E and M steps of the proposed LA-EM scheme. First, the E-step is derived as follows:

$$
\begin{aligned}
&Q(\Theta_{la} | \Theta_{la}^{(n)}) = E_{\Lambda | O, \Theta_{la}^{(n)}} [log L(\Theta_{la}; O, \Lambda)] \\
&= \sum_{y=1}^{Y} \Lambda(n,y) \times \sum_{x=1}^{X} (log \Omega_{x,y,d,h} + log \Pr(\lambda_y))
\end{aligned}
\tag{10}
$$

where $\Omega_{x,y,d,h}$ is defined as:

$$
\begin{aligned}
&\Omega_{x,y,d,h} \\
&= \begin{cases} M_{x,d,h} & VU_{x,y} = 1, t_y = d, s_y = h, \lambda_y = 1 \\ 1 - \sum_{d=1}^{D} \sum_{h=H} M_{x,d,h} & VU_{x,y} = 0, t_y = d, s_y = h, \lambda_y = 1 \\ N_{x,d,h} & VU_{x,y} = 1, t_y = d, s_y = h, \lambda_y = 0 \\ 1 - \sum_{d=1}^{D} \sum_{h=H} N_{x,d,h} & VU_{x,y} = 0, t_y = d, s_y = h, \lambda_y = 0 \end{cases}
\end{aligned}
\tag{11}
$$

For the M-step, in order to get the optimal Θ^* that maximizes the Q function, we set partial derivatives of $Q(\Theta_{la} | \Theta_{la}^{(n)})$ with respect to Θ_{la} to 0. We can get the optimal estimation of the parameters for the next iteration (i.e., $(M_{x,d,h})^{(n+1)}$, $(N_{x,d,h})^{(n+1)}$ and $(q)^{(n+1)}$) as follows:

$$
\begin{aligned}
M_{x,d,h}^* &= \frac{\sum_{y \in SW_{x,d,h}} \Pr(\lambda_y = 1 | O_y, \Theta^{(n)})}{\sum_{y=1}^{Y} \Pr(\lambda_y = 1 | O_y, \Theta_{la}^{(n)})} \\
N_{x,d,h}^* &= \frac{\sum_{y \in SW_{x,d,h}} (1 - \Pr(\lambda_y = 1 | O_y, \Theta_{la}^{(n)}))}{\sum_{y=1}^{Y} (1 - \Pr(\lambda_y = 1 | O_y, \Theta_{la}^{(n)}))} \\
q^* &= \frac{\sum_{y=1}^{Y} \Pr(\lambda_y = 1 | O_y, \Theta_{la}^{(n)})}{Y}
\end{aligned}
\tag{12}
$$

where $SW_{x,d,h}$ is the set of users who visit the venue V_x and their check-in points in a city last for d days and their activity range is h miles.

We further optimize the inference process by leveraging both Cramer-Rao lower bounds (CRLB) of estimation results obtained in the previous subsection and the social connections between users.

The CRLB is defined as the inverse of Fisher information: $CRLB = J^{-1}$, where J is the Fisher information of the estimation parameter. The CRLB can be used to obtain approximate confidence bounds of the maximum likelihood estimation [28]. Using the likelihood function from Equation (8) and the results of estimation parameters from Equation (12), we can compute CRLB to quantify the accuracy of our solution using a similar method we developed in [35].

In particular, we can assess the estimation accuracy of the estimation on λ_x by computing its confidence bounds.

Formally, the confidence bounds of λ_x are given as:

$$(\hat{\lambda}_x^{MLE} - c_p\sqrt{var(\hat{\lambda_x}^{MLE})}, \hat{\lambda_x}^{MLE} + c_p\sqrt{var(\hat{\lambda_x}^{MLE})}) \tag{13}$$

where cp is the standard score of confidence level p. $var(\hat{\lambda_x}^{MLE})$ is the estimation variance on λ_x, which can be computed from CRLB based on Equation (4).

Using the computed CRLB, we can compute the confidence bound cb_x on the local attractiveness estimation of each venue. We further define AC_y to represent the estimation accuracy of a user's localness. Given the Venue-User matrix VU, AC_y can be computed as:

$$AC_y = \frac{\sum_{x \in VU_y}(cb_x)}{|VU_y|} \tag{14}$$

where CV_y is the set of venues user U_y has check-in points.

We then optimize the inference of a user's localness as follows: if a user U_y's localness estimation accuracy AC_y is less than a certain threshold (we use 0.5 in our experiment) and have social connections with others, we compute an optimized localness of U_y by leveraging its social constraints (i.e., SR Matrix). In particular, we define the objective function of our problem as follows:

$$f = \sum_{y \in SU} \sum_{y' \in SR_y} |\Lambda_y^* - \Lambda_y'| \cdot w(y, y') \tag{15}$$

where SU is the set of users who have social connections, SR_y is the set of users who have social connections with user U_y. $w(y, y')$ is the strength of social connection between user U_y and $U_{y'}$, which is reflected by the number of same venues the two users visited together. Additionally, Λ_y^* is the optimized inference of localness estimation of user U_y and $\Lambda_{y'}$ is the localness estimation of user $U_{y'}$ from the Spatial-Temporal Modeling component. The goal is to find the Λ_y^* for every user in SU that minimizes the defined objective function. This optimization problem can be solved in linear time using weighted median algorithm [4].

B. Venue Influence Inference

In this subsection, we present the venue influence scope inference scheme: Influence-Awareness Expectation Maximization (INA-EM). The objective of the INA-EM scheme is to infer the influence scope of each venue.

1) The Likelihood Function: Similarly as we derive LA-EM algorithm, we first setup the likelihood function to infer the venue influence scope. The estimation parameter $\Theta_{ina} = (I_1, ..., I_Y; J_1, ..., J_Y; b)$, where I_y, J_y and b are defined in Equation (5). Furthermore, we define a vector of latent variables Z to indicate the popularity of venues (which is directly related to the influence scope of a venue). Specifically, we have a corresponding variable z_x for each venue V_x (i.e., z_x=1 if $V_x = P$ and z_x=0 otherwise). S is the observed data (i.e., CI Matrix). Hence, the likelihood function $L(\Theta_{ina}; S, Z)$ for the INA-EM scheme can be written as follows:

$$\begin{aligned} L(\Theta_{ina}; S, Z) &= \Pr(S, Z|\Theta_{ina}) \\ &= \prod_{x \in V} \Pr(z_x|S_x, \Theta_{ina}) \times \prod_{y \in U} \delta_{y,x} \times \Pr(z_x) \end{aligned} \tag{16}$$

where $\delta_{y,x}$ is define in Table I.

Table I
NOTATIONS FOR INA-EM

$\delta_{y,x}$	$\Pr(z_x)$	$Z(n,x)$	Constrains
I_y	b	$\Pr(V_x = P\|S_x, \Theta_{ina}^{(n)})$	$CI_{x,y} = 1, z_x = 1$
$1 - I_y$	b	$\Pr(V_x = P\|S_x, \Theta_{ina}^{(n)})$	$CI_{x,y} = 0, z_x = 1$
J_y	$1 - b$	$\Pr(V_x = \bar{P}\|S_x, \Theta_{ina}^{(n)})$	$CI_{x,y} = 1, z_x = 0$
$1 - J_y$	$1 - b$	$\Pr(V_x = \bar{P}\|S_x, \Theta_{ina}^{(n)})$	$CI_{x,y} = 0, z_x = 0$

2) The INA-EM Scheme: Given the above likelihood function, we can derive E and M steps of the proposed INA-EM scheme. First, the E-step is given as follows:

$$\begin{aligned} Q(\Theta_{ina}|\Theta_{ina}^{(n)}) &= E_{Z|S,\Theta_{ina}^{(n)}}[logL(\Theta_{ina}; S, Z)] \\ &= \sum_{x=1}^{X} Z(n,x) \times \sum_{y=1}^{Y}(log\delta_{y,x} + log\Pr(z_x)) \end{aligned} \tag{17}$$

where $Z(n,x)$ is defined in Table I and n is the iteration index.

In the M-step, as before, we choose Θ_{ina}^* that maximizes the $Q(\Theta_{ina}|\Theta_{ina}^{(n)})$ function in each iteration to be the $\Theta_{ina}^{(n+1)}$ for the next iteration. The optimal solutions of the parameters for the next iteration (i.e., $(I_y)^{(n+1)}$, $(J_y)^{(n+1)}$ and $(h)^{(n+1)}$) are as follows:

$$\begin{aligned} I_y^* &= \frac{\sum_{y \in SI_y} \Pr(z_x = 1|S_x, \Theta_{ina}^{(n)})}{\sum_{x=1}^{X} \Pr(z_x = 1|S_x, \Theta^{(n)})} \\ J_y^* &= \frac{\sum_{y \in SI_y}(1 - \Pr(z_x = 1|S_x, \Theta_{ina}^{(n)}))}{\sum_{x=1}^{X}(1 - \Pr(z_x = 1|S_x, \Theta_{ina}^{(n)}))} \\ h^* &= \frac{\sum_{x=1}^{X} \Pr(z_x = 1|S_x, \Theta_{ina}^{(n)})}{X} \end{aligned} \tag{18}$$

where SI_y is the set of venues user U_y visits. It turns out our proposed INA-EM scheme could also provide a quantitative metric to evaluate exactly how popular a venue would be (i.e., $\Pr(z_x = 1|S_x, \Theta_{ina}^{(n)})$). Based on the definition of influence scope of a venue (Inf_x) in the previous section, we can compute it for each venue as $Inf_x = 1 - \Pr(z_x = 1|S_x, \Theta_{ina}^{(n)})$.

C. Home Location Estimation

In this section, we formulate the problem of estimating the location of users as an optimization problem by incorporating the outputs of the LA-EM and INA-EM from the previous subsections. We present a new algorithm called Home Location Estimation (HLE) to solve the home location estimation problem. The objective of the HLE algorithm is to accurately estimate the home locations of users.

1) Optimization Model Formulation: We formulate the home location estimation problem as an optimization problem. In particular, we can generate a weighted bipartite graph $G = (U, V; E, W)$ based on the Check-in Matrix CI and estimated influence scope score inf_x of venues. The U and V represent the set of users and venues respectively (i.e., nodes in G). E and W represent the set of edges in the G and the influence scope scores of all venues respectively. For simplicity, we use u and v to represent user U_y and venue V_x in this subsection. If the element $CI_{x,y}$ in CI matrix is 1, we have a link between node u and v. All edges that end at venue node v will be assigned the weight inf_v. We further define a distance function $dist(u, v)$ to represent the distance (i.e., cosine-haversine formula [29]) between the home location of user u and venue v. Given the generated bipartite graph G and the GPS coordinates (i.e., latitude α_v and longitude β_v) of all venues, we define the objective function f of our problem as:

$$f = \sum_{u \in U} \sum_{(u,v) \in E} w(u, v) \cdot dist(u, v) \tag{19}$$

where $w(u, v)$ is the weight of the edge (u, v). The goal is to find the (α_u, β_u) for every user in U that minimizes the defined objective function. In another word, the estimated user's home location is the point that has least weighted distance to all venues she/he visited.

Note that for each $u \in U$, if we can find (α_u, β_u) that minimizes $\sum_{(u,v) \in E} w(u, v) \cdot dist(u, v)$, then the objective function f is minimized. So we only need to focus on finding (α, β) of user u that minimizes the objective function $f(\alpha, \beta) = \sum_{i=1}^{m} w_i \cdot dist(\alpha, \beta, \alpha_i, \beta_i)$, where i is the index of neighbors of user u and m is the number of visited venues of user u.

2) The HLE Scheme: We present the HLE scheme to solve the above optimization problem. We first define the following terms:

$$\begin{aligned} A &= \frac{1}{2} \sum_{v=1}^{m} w_v \\ B &= \frac{1}{2} \sum_{v=1}^{m} w_v \sin(\alpha_v) \\ C &= \frac{1}{2} \sum_{v=1}^{m} w_v \cos(\alpha_v) \cos(\beta_v) \\ D &= \frac{1}{2} \sum_{v=1}^{m} w_v cos(\alpha_v) \sin(\beta_v) \end{aligned} \tag{20}$$

To find the (α_u, β_u) that minimizes $f(\alpha_u, \beta_u)$, we first consider the boundary cases. Let $\alpha_u = \pm\frac{\pi}{2}$, we have $f(\pm\frac{\pi}{2}, \beta_u) = A \mp B$. Let $\beta_u = \pm\pi$, we have $f(\alpha_u, \pm\pi) = A - B\sin(\alpha_u) + C\cos(\alpha_u)$. For the regular case, we set derivatives of $f(\alpha_u, \beta_u)$ with respect to α_u and β_u to 0 and we can find the desired (α_u, β_u) that minimizes $f(\alpha_u, \beta_u)$.

V. Evaluation

In this section, we conduct experiments to evaluate the performance of the *UHLI (Unsupervised Home Location Inference)* scheme on three real-world data traces collected from a location-based social network service: Foursquare. We demonstrate the effectiveness of our proposed scheme on these data traces and compare the performance of our scheme to the state-of-the-art baselines.

A. Experimental Setups and Evaluation Metrics

1) Data Trace Statistics: In this paper, we evaluate the UHLI scheme on three real-world data traces collected from Foursquare. In Foursquare, users can easily share their location information (i.e., check-in points) at different venues they visit in a city. Each check-in point is formatted as: (user ID, venue ID, timestamp). In the evaluation, we selected the data traces from three cities in U.S [1]: Washington D.C., Boston and Chicago. The statistics of these traces are summarized in Table II. The data traces we collected also contains home location information of users, which serves as the ground truth to decide the home location of users in our evaluation. One should note that such home location information is not globally available for all users in all cities [31], which is the main motivation to develop *UHLI* scheme to infer the home location of people from their check-in points.

2) Data Pre-Processing: To evaluate our methods in real world settings, we went through the following data pre-processing steps to generate the inputs for the UHLI scheme: (i) Venue-User Matrix (VU Matrix) Generation; (ii) Check-in Matrix (CI Matrix) Generation; (iii) Temporal Vector (T Vector) Generation; (iv) Spatial Vector (S Vector) Generation; (v) Social Relationship Matrix SR Generation. They are summarized as follows:

- *Venue-User Matrix Generation*: We generate the VU Matrix by associating each venue with the users who visited this venue (i.e., the users who had check-in points at the venue). In particular, if user U_y visits venue V_x in the data trace, we set the element $VU_{x,y}$ in VU to 1 and 0 otherwise.
- *Check-in Matrix Generation*: By definition, CI Matrix is simply the transpose of VU Matrix.
- *Temporal Vector Generation*: we generate the T vector by setting the corresponding element as the time length of the user's check-in trace in a city. In particular, $t_y = d$ if the difference between the first and last check point of user U_y in a city is d days.
- *Spatial Vector Generation*: we generate the S vector by setting the corresponding element as the user's activity range in a city. In particular, $s_y = h$ if the activity range of user U_y in a city is h miles.
- *Social Relationship Matrix Generation*: We generate the SR Matrix as follows: if user U_y and user $U_{y'}$ have a social connection, we set the element $SR_{y,y'}$ in SR to 1 and 0 otherwise.

[1] https://archive.org/details/201309_foursquare_dataset_umn

Table II
DATA TRACES STATISTICS

Data Trace	Washington D.C	Boston	Chicago
Number of Users	17,231	12,804	31,965
Number of Venues	1,932	1,478	2,529
Number of Check-ins	25,722	18,296	48,605

3) Evaluation Metric: In our evaluation, we define two metrics to evaluate the performance of the *UHLI* scheme. The first metric is *Average Error Distance for Top-k% Users (AED@Top-k%)*. In particular, we denote l_u and $\hat{l_u}$ as the user u's real and estimated home location respectively. $dis(l_u, \hat{l_u})$ is defined as the distance between l_u and $\hat{l_u}$. The Top-k% users are the top k% users who are ranked by $dis(l_u, \hat{l_u})$. In our experiments, we evaluate the performance of all schemes by varying the value of k%. A low AED-Top-k% value means that the approach can geo-locate users close to their real home location on average for the Top-k% users. The second metric is *Accuracy within M miles (ACC@M)* which we borrowed from [5]. Particularly, $ACC@M$ is used to measure the fraction of users who can be accurately geo-located within a certain M miles from her/his real home location. In our experiment, we evaluate the performance of different techniques by varying the values of M. A high $ACC@M$ value means that the approach can geo-locate a larger fraction of users within a given error bound. The mathematical definitions of AED-Top-k% and $ACC@M$ are given in Table III.

Table III
METRIC DEFINITIONS

Metric	Definition
AED@Top-k%	$\frac{\sum_{u \in U} dis(l_u, \hat{l_u}) \mid Rank(u) < k\%}{\mid U \mid}$
$ACC@M$	$\frac{\{u \mid u \in U \wedge dis(l_u, \hat{l_u}) < M\}}{\mid U \mid}$

B. Evaluation of Our Scheme

Table IV
ESTIMATION ACCURACY ON WASHINGTON D.C. DATA TRACE IN TERMS OF AED@TOP-K%

	Top-k% Accurate Users				
Alg	20%	40%	60%	80%	100%
UHLI	**0.66**	**0.97**	**1.42**	**2.43**	**21.71**
HLI	0.80	1.37	2.29	9.33	82.97
MLP	0.75	1.22	2.28	15.19	93.83
FM	0.75	1.15	2.04	9.69	84.43
UDI	0.73	1.14	2.05	10.54	88.93
FL	0.72	1.16	2.13	14.64	93.34
OAlgo	0.79	1.32	2.24	9.29	82.93
Aver	10.78	1.22	2.27	20.01	259.86

In this subsection, we evaluate the performance of the proposed *UHLI* scheme and compare it to the state-of-the-art user geo-locating techniques that include:

Table V
ESTIMATION ACCURACY ON WASHINGTON D.C. DATA TRACE IN TERMS OF ACC@M MILES

	M (mile)				
Alg	1	3	5	7	9
UHLI	**0.206**	**0.560**	**0.696**	**0.765**	**0.794**
HLI	0.142	0.455	0.549	0.621	0.655
MLP	0.150	0.426	0.539	0.599	0.627
FM	0.160	0.454	0.569	0.632	0.661
UDI	0.161	0.451	0.566	0.629	0.658
FL	0.159	0.441	0.555	0.614	0.643
OAlgo	0.147	0.456	0.551	0.623	0.657
Aver	0.149	0.430	0.538	0.598	0.625

- *HLI*: it proposes a machine learning approach that locate people's home location by integrating the spatial and temporal features of people's trajectories [8].
- *MLP*: it proposes a generative probabilistic approach that infers a user's locations by leveraging the home locations of the user's online friends [18].
- *FM*: it infers a user's location by utilizing the home locations of the people that visit similar places as the user [3].
- *UDI*: it proposes a unified framework for profiling users' home locations by exploring both social network between users and influence probabilities of different locations [19].
- *FL*: it proposes a network-based approach that leverages the evidence of social tie strength between users [26].
- *OAlgo*: it presents a hierarchical ensemble algorithm for inferring the home location of users by exploring the tweeting behavior of users [22].
- *Aver*: it simply computes the home location of a user by taking the average of the coordinates of all places the user visited.

UHLI scheme differs from the above schemes in the sense that it is an unsupervised approach which does not require any training data on the home locations of the targeting users or their social connections. Instead, it judiciously leverage the venue locations visited by the users to infer their home locations using a principled approach.

1) Evaluation Results: In our evaluation, we evaluate the performance of above schemes using AED-Top-k% and $ACC@M$ metrics we introduced. The results of AED-Top-k% on Washington D.C. data trace are shown in Table IV. We observe that the proposed *UHLI* scheme outperforms all compared baselines. Specifically, it has the smallest average error distance on the estimation of users' home location.

In particular, the average estimation error of all users is 4 to 12 times better than the compared baselines. The *Aver* heuristic that takes the average coordinates of all venues that a user visited to estimate the user location failed to provide an accurate estimation (i.e., average error of users is 259.86 miles).

Furthermore, we also evaluate the estimation performance of all schemes in terms of $ACC@M$. The evaluation results on Washington D.C. data trace are shown in Table V. We observe that the proposed *UHLI* scheme also outperforms all compared baselines over different values of M. In particular, 21% and 56% of users can be geo-located within 1 and 3 miles of their real home locations respectively, which is 5% and 11% better than the best performed baselines.

We repeated the above experiments on Boston and Chicago data trace. The results on Boston data trace in terms of AED-Top-$k\%$ and $ACC@M$ are shown in Table VI and Table VII respectively. In those tables, we observe that *UHLI* continuously outperforms all compared baselines with nontrivial performance gains. The results on Chicago data trace in terms of AED-Top-$k\%$ and $ACC@M$ are shown in Table VIII and Table IX respectively. The above evaluation results from real world data traces demonstrate that the proposed *UHLI* scheme can effectively infer the home location of users and achieved significant performance improvements without using any training data compared to the state-of-the-art techniques.

Table VI
ESTIMATION ACCURACY ON BOSTON DATA TRACE IN TERMS OF AED@TOP-K%

	Top-k% Accurate Geo-locating Users				
Alg	20%	40%	60%	80%	100%
UHLI	**0.50**	**1.02**	**1.52**	**2.93**	**25.49**
HLI	0.63	1.25	2.03	7.66	79.99
MLP	0.54	1.05	1.92	10.22	86.89
FM	0.62	1.22	2.89	38.09	131.51
UDI	0.67	1.30	2.11	8.22	80.77
FL	0.61	1.19	2.67	9.13	80.81
OAlgo	0.67	1.30	2.09	7.64	79.52
Aver	0.56	1.05	2.00	15.83	224.66

Table VII
ESTIMATION ACCURACY ON BOSTON DATA TRACE IN TERMS OF ACC@Y MILES

	M (mile)				
Alg	1	3	5	7	9
UHLI	0.221	0.557	0.673	0.713	0.743
HLI	0.168	0.462	0.583	0.615	0.643
MLP	0.202	0.474	0.569	0.602	0.631
FM	0.177	0.425	0.501	0.532	0.558
UDI	0.161	0.424	0.579	0.614	0.644
FL	0.181	0.433	0.508	0.540	0.566
OAlgo	0.155	0.459	0.579	0.612	0.640
Aver	0.203	0.477	0.559	0.592	0.619

Table VIII
ESTIMATION ACCURACY ON CHICAGO DATA TRACE IN TERMS OF AED@TOP-K%

	Top-k% Accurate Users				
Alg	20%	40%	60%	80%	100%
UHLI	**0.67**	**1.11**	**1.79**	**2.77**	**29.59**
HLI	0.77	1.68	2.98	5.22	54.32
MLP	1.10	2.11	3.30	4.79	64.79
FM	1.45	2.39	3.91	8.96	72.14
UDI	1.28	2.20	3.42	8.74	98.98
FL	1.42	2.32	3.41	4.67	53.77
OAlgo	0.75	1.61	2.97	5.30	54.42
Aver	1.45	2.38	3.50	4.92	151.16

Table IX
ESTIMATION ACCURACY ON CHICAGO DATA TRACE IN TERMS OF ACC@Y MILES

	M (mile)				
Alg	1	3	5	7	9
UHLI	**0.190**	**0.482**	**0.681**	**0.766**	**0.808**
HLI	0.145	0.332	0.494	0.565	0.681
MLP	0.083	0.295	0.447	0.606	0.714
FM	0.051	0.276	0.424	0.512	0.577
UDI	0.064	0.294	0.451	0.569	0.658
FL	0.052	0.284	0.451	0.624	0.737
OAlgo	0.151	0.336	0.495	0.554	0.674
Aver	0.051	0.275	0.437	0.605	0.715

VI. CONCLUSION

This paper proposes an unsupervised approach to infer the home location of people by using the sparse and noisy data they share on LBSN. We develop the *Unsupervised Home Location Inference (UHLI)* scheme that can accurately infer users' home locations by jointly estimating the localness of people and the influence scope of venues under a rigorous analytical framework. We evaluate our new approach on three real-world datasets collected from Foursquare. The results showed that our approach can accurately infer the home locations of people in a city and significantly outperform other state-of-the-art baselines in terms of estimation accuracy. The results of our paper are important because they can directly contribute to localized recommendation, targeted ads and urban planning applications where training data is difficult or expensive to obtain.

ACKNOWLEDGMENTS

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] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *The Semanic Web: Research and Applications*, pages 375–389. Springer, 2011.

[2] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 114–122. ACM, 2011.

[3] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Internation World Wide Web Conference (WWW)*, pages 61–70. ACM, 2010.

[4] D. Brownrigg. The weighted median filter. *Communications of the ACM*, 27(8):807–818, 1984.

[5] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *International Conference on Information and Knowledge Management (CIKM)*, pages 759–768. ACM, 2010.

[6] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[7] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *International Conference on Mobile Systems, Applications, and Services (Mobisys)*, pages 31–42. ACM, 2003.

[8] T. Hu, J. Luo, H. Kautz, and A. Sadilek. Home location inference from sparse and noisy data: Models and applications. In *International Conference on Data Mining Workshop (ICDMW)*, pages 1382–1387. IEEE, 2015.

[9] C. Huang and D. Wang. On interesting place finding in social sensing: An emerging smart city application paradigm. In *International Conference on Smart City*, pages 13–20. IEEE, 2015.

[10] C. Huang and D. Wang. Exploiting spatial-temporal-social constraints for localness inference using online social media. In *International Conference on Advances in Social Network Analysis and Mining (ASONAM)*. ACM/IEEE, 2016.

[11] C. Huang and D. Wang. Topic-aware social sensing with arbitrary source dependency graphs. In *International Conference on Information Processing in Sensor Networks (IPSN)*, pages 1–12. IEEE, 2016.

[12] C. Huang and D. Wang. Unsupervised interesting places discovery in location-based social sensing. In *International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 67–74. IEEE, 2016.

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

[14] E. Kıcıman and M. Richardson. Towards decision support and goal achievement: Identifying action-outcome relationships from social media. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 547–556. ACM, 2015.

[15] C. A. Lampe, N. Ellison, and C. Steinfield. A familiar face (book): profile elements as signals in an online social network. In *Conference on Human Factors in Computing Systems (CHI)*, pages 435–444. ACM, 2007.

[16] K. Lewis, J. Kaufman, and N. Christakis. The taste for privacy: An analysis of college student privacy settings in an online social network. *Journal of Computer-Mediated Communication*, 14(1):79–100, 2008.

[17] R. Li, C. Wang, and K. C.-C. Chang. User profiling in an ego network: co-profiling attributes and relationships. In *Internation World Wide Web Conference (WWW)*, pages 819–830. ACM, 2014.

[18] R. Li, S. Wang, and K. C.-C. Chang. Multiple location profiling for users and relationships from social network and content. *International Database on Very Large Data Bases*, 5(11):1603–1614, 2012.

[19] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 1023–1031. ACM, 2012.

[20] C.-T. Lu, S. Xie, X. Kong, and P. S. Yu. Inferring the impacts of social media on crowdfunding. In *International Conference on Web Search and Data Mining (WSDM)*, pages 573–582. ACM, 2014.

[21] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao. Privacy vulnerability of published anonymous mobility traces. *Transactions on Networking (TON)*, 21(3):720–733, 2013.

[22] J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *Transactions on Intelligent Systems and Technology (TIST)*, page 47, 2014.

[23] J. Marshall, M. Syed, and D. Wang. Hardness-aware truth discovery in social sensing applications. In *Distributed Computing in Sensor Systems (DCOSS), 2016 International Conference on*, pages 143–152. IEEE, 2016.

[24] J. Marshall and D. Wang. Mood-sensitive truth discovery for reliable recommendation systems in social sensing. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 167–174. ACM, 2016.

[25] J. Marshall and D. Wang. Towards emotional-aware truth discovery in social sensing applications. In *Smart Computing (SMARTCOMP), 2016 IEEE International Conference on*, pages 1–8. IEEE, 2016.

[26] J. McGee, J. Caverlee, and Z. Cheng. Location prediction in social media based on tie strength. In *International Conference on Information and Knowledge Management (CIKM)*, pages 459–468. ACM, 2013.

[27] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *International Conference on Web Search and Data Mining (WSDM)*, pages 251–260. ACM, 2010.

[28] V. Papathanasiou. Some characteristic properties of the fisher information matrix via cacoullos-type inequalities. *Journal of Multivariate analysis*, 44(2):256–265, 1993.

[29] C. Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, pages 38–40, 1957.

[30] A. Sadilek and H. Kautz. Modeling the impact of lifestyle on health at scale. In *International Conference on Web Search and Data Mining (WSDM)*, pages 637–646. ACM, 2013.

[31] M. Sarwat, J. J. Levandoski, A. Eldawy, and M. F. Mokbel. Lars: An efficient and scalable location-aware recommender system. *Transactions on Knowledge and Data Engineering (TKDE)*, pages 1384–1399, 2014.

[32] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 1046–1054. ACM, 2011.

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

[34] 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.

[35] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal. On credibility estimation tradeoffs in assured social sensing. *Journal on Selected Areas in Communications (JSAC)*, 31(6):1026–1037, 2013.

[36] D. Wang, J. Marshall, and C. Huang. Theme-relevant truth discovery on twitter: An estimation theoretic approach. In *International Conference on Web and Social Media (ICWSM)*, 2016.

[37] J. Wang, Y. Zhao, and D. Wang. A novel fast anti-collision algorithm for rfid systems. In *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, pages 2044–2047. IEEE, 2007.

[38] J.-w. WANG, D. WANG, K. TIMO, and Y.-p. ZHAO. A novel anti-collision protocol in multiple readers rfid sensor networks [j]. *Chinese Journal of Sensors and Actuators*, 8:026, 2008.

[39] P. Wisniewski, H. Jia, H. Xu, M. B. Rosson, and J. M. Carroll. Preventative vs. reactive: How parental mediation influences teens' social media privacy behaviors. In *International Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, pages 302–316. ACM, 2015.

[40] D. Zhang, R. Han, D. Wang, and C. Huang. On robust truth discovery in sparse social media sensing. In *2016 International Conference on Big Data (IEEE BigData 2016)*. IEEE, 2016.

[41] J.-D. Zhang and C.-Y. Chow. Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *special interest group on information retrieval (SIGIR)*, pages 443–452. ACM, 2015.