Where Are You From: Home Location Profiling of Crowd Sensors from Noisy and Sparse Crowdsourcing Data

Chao Huang, Dong Wang, Shenglong Zhu
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
chuang7@nd.edu, dwang5@nd.edu, szhu3@nd.edu

***Abstract*—Crowdsourcing has emerged as an important data collection paradigm in participatory and human-centric sensing applications. While many crowdsourcing studies focus on sensing and recovering the status of the physical world, this paper investigates the problem of profiling the crowd sensors (i.e., humans). In particular, we study the problem of accurately inferring the home locations of people from the noisy and sparse crowdsourcing data they contribute. In this study, we propose a semi-supervised framework, *Where Are You From (WAYF)*, to accurately infer the home locations of people by explicitly exploring the localness of people and the dependency between people based on their check-in behaviors under a rigorous analytical framework. We perform extensive experiments to evaluate the performance of our scheme and compared it to the state-of-the-art techniques using three real world data traces collected from Foursquare. The results showed the effectiveness of our scheme in accurately profiling the home locations of people.**

***Index Terms*—Home Location Profiling, Crowdsourcing, Location Based Social Networks (LBSN)**

I. Introduction

With the rapid growth of ubiquitous Internet connectivity and Location-Based Social Network (LBSN) services (e.g., Foursquare, Gowalla, Google Places), crowdsourcing has become a key data collection paradigm in human-centric sensing applications. While many crowdsourcing studies focus on sensing and recovering the status of physical world, this paper investigates the problem of profiling the crowd sensors (i.e., humans). In particular, we study the problem of accurately estimating the home locations of people who are local residents in a city from noisy and sparse crowdsourcing data they contribute. User's home location is an important piece of information for many location based information services such as targeted ads of local business [4], urban planning [11], and location-aware recommendations [33], [17].

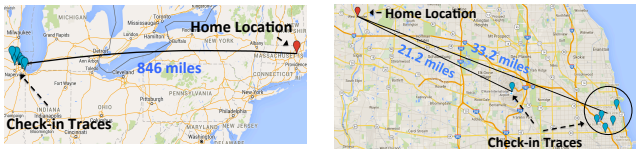
One simple approach for home location profiling task is to take the average of all venue locations the user visited as the estimated home location of the user. The assumption here is that check-in points of users come from the places which are close to their home locations. To check the accuracy of this simple method, we compute the average error distance between the estimated home location and the real home location of users over the real world datasets collected from three cities (i.e., Boston, Chicago and Washington D.C.) on Foursquare. The results show that the average estimation error of the simple average approach is *224 miles*, *150 miles* and *260 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.

Previous work has made significant progress to study the problem of geo-locating people in a city using online social network information [24], [19], [14], [23], [26]. However, most of current solutions either ignore the localness of users or the dependency between users based on their check-in behaviors. Such limitations has led to suboptimal estimation results by treating the non-local users as local ones or using the home locations of non-related users to estimate the location for each other [25]. To address these limitations, this work develops a new principled framework to investigate the problem of home location estimation of people by explicitly exploring both the localness of users and dependency between users from their check-in traces on LBSNs.

Two key challenges exist in order to solve the home location profiling problem:

- Noisy Data: both local and non-local people of a city can check-in at the venues in this city, which makes it a challenging task to separate local users from non-local ones solely based on their check-in traces (see Figure 1(a)). Furthermore, users can check-in at venues whose location are close or far away from his/her home locations (see Figure 1(b)).
- Data Sparity: the data (i.e., check-in points) from social media platforms are often incomplete and highly sparse: a person might not check in at every venue she/he visits in a city or she/he might intentionally choose not to check in due to some privacy concerns [6].

To address the above challenges, this paper develops a new semi-supervised framework, *Where Are You From (WAYF)*,



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

Figure 1. Check-in Points and Home Locations of Random Users on Foursquare

to explicitly explore both spatial and temporal information embedded in the crowdsourcing data publicly available on LBSNs. In particular, we propose a Maximum Likelihood Estimation (MLE) based algorithm to identify the local people in a city and a Bayesian inference approach to discover the dependent users based on their check-in behaviors. We then formulate the home location profiling problem as an optimization problem and derive an optimal solution for it. Finally, we evaluate the performance of the WAYF scheme and compare it with the state-of-the-art techniques through real-world datasets which are collected from the city of Boston, Chicago and Washington, D.C. on Foursquare. The evaluation results demonstrate the effectiveness achieved by our scheme.

Finally, a note on disclaimer. First, we did not discuss the privacy issue in this paper because the user identified in collected datasets from LBSNs are all anonymized [12]. Additionally, there exists a rich set of literature on the topic of protecting user's privacy in online social media applications [21], [13]. 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 profiling problem a trivial problem to solve.). Instead, we only used publicly available data from LBSNs with the goal to develop a new principled approach as an open-source resource for the research community.

We summarized the contributions of this paper as follows:

- In this paper, we investigate the problem of home location profiling of crowd sensors from *noisy* and *sparse* crowdsourcing data which are collected from online social media platform.
- We develop a principled framework (i.e., WAYF scheme) that allows us to derive an *optimal* solution to accurately estimate the home location of users by explicitly exploring the localness of users and dependency between users based on their check-in behaviors.
- Our proposed scheme is evaluated through real-world datasets which are collected from online social media (i.e., Foursquare) and compared to the state-of-the-art techniques. Experimental results demonstrate that the effectiveness achieved by our proposed scheme

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

II. Problem Formulation

In this section, we introduce the problem of inferring the home locations of people in a crowdsourcing application. In particular, we consider a crowdsourcing application where a set of I venues (i.e., $M_1, M_2, ..., M_X$) have been visited by a group of J users (i.e., $N_1, N_2, ..., N_Y$). Here we define M_i to be the i-th venue and N_j to be the j-th user. $N_j = 1$ if the user is a local resident of the city and $N_j = 0$ if she/he is not. We further define the following inputs to our model.

- *Definition 1. Venue-User Matrix* MN. We define Venue-User Matrix $MN_{I\times J}$ to indicate *which venue is visited by which user*. In particular, $MN_i^j = 1$ indicates that user N_j has check-in points at venue M_i and $MN_i^j = 0$ otherwise.
- *Definition 2. Time Length Vector* TL. We define a Time Length Vector TL_J to represent the time length of user's check-in points (i.e., the time difference between the first and last check-in points of the user in the dataset). In particular, $h_j = \zeta$ denotes that user N_j's check-in points in a city lasts for ζ days.

One key challenge in crowdsourcing applications lies in the fact that both local and non-local users (e.g., tourist) can generate check-in points in a city. To address this challenge, we first define a few important terms. In particular, we denote the *local attractiveness* of a venue M_i as η_i, which is the probability that a user is local given that the user has check-in points at the venue M_i. Furthermore, considering a user may have different time length of her/his check-in points, we define $\eta_{i,\zeta}$ as the probability of a venue M_i to attract local users whose check-in points in a city last for ζ days. Formally, η_i and $\eta_{i,\zeta}$ can be given as:

$$\begin{aligned}\eta_i &= p(N_j = 1 | MN_i^j = 1)\\ \eta_{i,\zeta} &= p(N_j = 1 | MN_i^j = 1, h_j = \zeta)\end{aligned} \tag{1}$$

We denote the prior probability that venue M_i is visited by a user whose check-in points lasts for ζ days by $o_{i,\zeta}$. The relationship between η_i and $\eta_{i,\zeta}$ can be expressed as:

$$\eta_i = \sum_{\zeta=1}^{B} \eta_{i,\zeta} \cdot \frac{o_{i,\zeta}}{o_i} \quad \zeta = 1, ..., B \tag{2}$$

where $o_i = p(MN_i^j = 1)$. Formally, $o_{i,\zeta}$ and o_i are given as:

$$\begin{aligned}o_{i,\zeta} &= p(MN_i^j = 1, h_j = \zeta)\\ o_i &= \sum_{\zeta=1}^{B} o_{i,\zeta}\end{aligned} \tag{3}$$

We further denote $\varphi_{i,\zeta}$ as the probability of a *local* user (whose check-in points in a city lasts for ζ days) visits a venue M_i. Similarly, we denote $\phi_{i,\zeta}$ as the probability of a *non-local*

user (whose check-in points in a city lasts for ζ days) visits a venue N_j. $\varphi_{i,\zeta}$ and $\phi_{i,\zeta}$ are formally defined below:

$$\varphi_{i,\zeta} = p(MN_i^j = 1, h_j = \zeta | N_j = 1)$$
$$\phi_{i,\zeta} = p(MN_i^j = 1, h_j = \zeta | N_j = 0) \quad (4)$$

Based on Bayes' theorem, the conditional probabilities $\varphi_{i,\zeta}$ and $\phi_{i,\zeta}$ can be future derived as:

$$\varphi_{i,\zeta} = \eta_{i,\zeta} \times o_{i,\zeta}/c$$
$$\phi_{i,\zeta} = (1 - \eta_{i,\zeta}) \times o_{i,\zeta}/(1 - c) \quad (5)$$

where c represents the prior probability that a randomly chosen user is local.

In addition, we further define the following items to be used in the identification of dependent users based on their check-in behaviors.

- *Definition 3. Check-Category Matrix CC.* We define a Visit Category Matrix $CC_{J\times L}$ to represent the number of check-in points users have on each category of venues. Here L is the set of venue categories. In particular, each element CC_j^l is the number of check-in points of user N_j on venues whose category belongs to category l.
- *Definition 4. Check-Time Matrix CT.* We define a Check-Time Matrix $CT_{J\times 2}$ to represent the number of check-in points of users at different time of a day. In particular, each element CT_j^d, $(d \in [1, 2])$ is the number of check-in points of user N_j on daytime and nighttime respectively.

The above two features (i.e., CC and CT) are used to classify the users into different clusters where users in the same cluster are expected to have similar check-in behaviors. We denote such clusters of users as $H_1, H_2, ..., H_k, ..., H_Y$. For user N_j, we use ϱ_j to represent which cluster it belongs to (i.e., $\varrho_j = k$ if $N_j \in H_k$).

Finally, we formulate our user home location inference problem as follows: *given the Venue-User Matrix MN, Time Length Vector TL_J, Check-Category Matrix CC and Check-Time Matrix CT*, our goal is to accurately estimate the home locations of all local users in a city who have check-in trace on LBSNs. We define ω_j and δ_j to represent the latitude and longitude of user N_j's home location. Formally, our goal can be given as follows:

$$(\omega_j, \delta_j | MN, TL, CC, CT) \quad \forall j, 1 \le j \le J \quad (6)$$

We summarize the defined notations in Table I.

Table I
SUMMARY OF NOTATIONS

Symbol	Interpretation
M	set of venues
N	set of users
MN	venue-user matrix
TL	time length vector
CC	check-category matrix
CT	check-time matrix

III. THE WHERE ARE YOU FROM (WAYF) FRAMEWORK

In this section, we present our solution *Where Are You From (WAYF)* scheme to infer people's home locations by exploring the localness of users and dependency between users based on their check-in behavior. The WAYF scheme consists of two major components: *Local People Identification* and *Home Location Inference*. We will explain these two components in detail in the following subsections.

A. Local People Identification

In this subsection, we present the user localness identification scheme: Local People Identification (LPI). The objective of the LPI scheme is to identify local users from non-local ones by using the MN matrix, TL vector, CC matrix and CT matrix.

Based on the terms and variables we defined in Section II, the likelihood function $F = (\Omega; \Gamma, \Phi)$ for LPI is as follows:

$$\begin{aligned} &L(\Omega; \Gamma, \Phi) \\ &= p(\Gamma, \Phi | \Omega) \\ &= \prod_{j=1}^{J} \Big[\prod_{i=1}^{I} \prod_{\zeta=1}^{B} \Upsilon_{i,j,\zeta} \cdot p(\Phi_j) \Big] \cdot p(\Phi_j | \Gamma_j, \Omega) \end{aligned} \quad (7)$$

where Ω can be given as:

$$\Omega = (\varphi_{i,\zeta}, \varphi_{2,\zeta}, ..., \varphi_{I,\zeta}; \phi_{1,\zeta}, \phi_{2,\zeta}, ..., \phi_{I,\zeta}; c) \quad (8)$$

Γ is the observed data (i.e., Matrix MN and Vector TL). Φ represents a set of latent variables that indicate whether a user is local or not. More specially, we define a corresponding variable Φ_j for each user N_j such that $\Phi_j = 1$ if N_j is local and $\Phi_j = 0$ otherwise. Additionally, $\Upsilon_{i,j,\zeta}$ and $p(\Phi_j)$ are defined in Equation (9) and (10).

$$\Upsilon_{i,j,\zeta} = \begin{cases} \varphi_{i,\zeta} & MN_i^j = 1, h_j = \zeta, \Phi_j = 1 \\ 1 - \sum_{\zeta=1}^{B} \varphi_{i,\zeta} & MN_i^j = 0, h_j = \zeta, \Phi_j = 1 \\ \phi_{1,\zeta} & MN_i^j = 1, h_j = \zeta, \Phi_j = 0 \\ 1 - \sum_{\zeta=1}^{B} \phi_{1,\zeta} & MN_i^j = 0, h_j = \zeta, \Phi_j = 0 \end{cases} \quad (9)$$

$$\Phi(n, j) = \begin{cases} c & MN_i^j = 1, h_j = \zeta, \Phi_j = 1 \\ c & MN_i^j = 0, h_j = \zeta, \Phi_j = 1 \\ 1 - c & MN_i^j = 1, h_j = \zeta, \Phi_j = 0 \\ 1 - c & MN_i^j = 0, h_j = \zeta, \Phi_j = 0 \end{cases} \quad (10)$$

Given the above formulated likelihood function, we can derive E and M steps of the proposed LPI scheme. First, the E-step is derived as follows:

$$\begin{aligned} &Q(\Omega | \Omega^{(n)}) \\ &= E_{\Phi || \Gamma, \Omega^{(n)}} [log L(\Omega; \Gamma, \Phi)] \\ &= \sum_{j=1}^{J} \Phi(n, j) \times \sum_{i=1}^{I} (log \Upsilon_{i,j,\zeta} + log p(\Phi_j)) \end{aligned} \quad (11)$$

Notation	Solution
$\varphi^*_{i,\zeta}$	$(\Sigma_{y\in SW_{x,d}} p(\Phi_j = 1\|\Gamma_j, \Omega^{(n)}))/(\Sigma^Y_{y=1} p(\Phi_j = 1\|\Gamma_j, \Omega^{(n)}))$
$\phi^*_{i,\zeta}$	$(\Sigma_{y\in SW_{x,d}} (1 - p(\Phi_j = 1\|\Gamma_j, \Omega^{(n)})))/(\Sigma^Y_{y=1} (1 - p(\Phi_j = 1\|\Gamma_j, \Omega^{(n)})))$
c^*	$(\Sigma^Y_{y=1} p(\Phi_j = 1\|\Gamma_j, \Omega^{(n)}))/(J)$

Table II
SOLUTIONS OF LPI SCHEME

where $\Phi(n, j)$ is defined in Equation (12) and n is the iteration index.

$$p(\Phi_j) = \begin{cases} p(N_j = 1|\Gamma_j, \Omega^{(n)}) & MN^j_i = 1, h_j = \zeta, \Phi_j = 1 \\ p(N_j = 1|\Gamma_j, \Omega^{(n)}) & MN^j_i = 0, h_j = \zeta, \Phi_j = 1 \\ p(N_j = 0|\Gamma_j, \Omega^{(n)}) & MN^j_i = 1, h_j = \zeta, \Phi_j = 0 \\ p(N_j = 0|\Gamma_j, \Omega^{(n)}) & MN^j_i = 0, h_j = \zeta, \Phi_j = 0 \end{cases} \tag{12}$$

For the M-step, in order to get the optimal Ω^* that maximizes the Q function, we set partial derivatives of $Q(\Omega|\Omega^{(n)})$ with respect to Ω to 0. We can get the optimal estimation of the parameters for the next iteration (i.e., $(\varphi_{i,\zeta})^{(n+1)}$, $(\phi_{i,\zeta})^{(n+1)}$ and $(c)^{(n+1)}$) in Table II. $D_{i,\zeta}$ is the set of users who visit the venue M_i and their check-in points in a city last for ζ days

We can further optimize the user localness identification process by leveraging both dependency between users and the Cramer-Rao lower bounds (CRLB) of estimation results. In particular, we identify the users with inaccurate estimation results by computing the CRLBs and improve inaccurate estimations by leveraging the home locations of their dependent users based on their check-in behaviors.

We first cluster dependent users based on the features in Section II (i.e., CC matrix and CT matrix). In particular, we define a vector s_j where the l-th entry of s_j represents the probability of user N_j has check-in points at venues with category l. We draw s_j from a Dirichlet(ϑ) where ϑ is the Dirichlet parameter for cluster ϱ_j. Similarity, we represent the user N_j's temporal check-in behavior as a vector t_j drawn from a Dirichlet ι_j. Formally, they can be represented as:

$$s_j|\varrho = Dirichlet(\vartheta_k); \quad t_j|\varrho = Dirichlet(\iota_k) \tag{13}$$

We define CC_k and CT_k to represent the sub-matrices of Check-Category Matrix CC and Check-Time matrix CT which include the users in the k-th cluster. We further define the set of users in the k-th cluster as U_k. Based on the defined terms and variables, we define the likelihood function as:

$$\begin{aligned} &p(CC_k, CT_k|\vartheta_k, \iota_k) \\ &= \prod_{j\in U_k} p(CC_k|\vartheta_k) p(CT_k|\iota_k) \end{aligned} \tag{14}$$

Using the formulated likelihood functions, we can derive the cluster assignment for each user. We then update hyperparameters of each cluster (i.e., ϑ and ι) iteratively and the cluster assignment for each user (i.e., ϱ_j) using the Bayesian inference model until the values of the hyperparameters converge. To update cluster hyperparameters, we maximize the likelihood function $p(CC_k, CT_k|\vartheta_k, \iota_k)$ with respect to ϑ and ι respectively. Particularly, we compute the maximum likelihood updates for ϑ_k given the CC_k of users in the k-th cluster. Similarly, we compute the maximum likelihood updates for ι_k given the CT_k of users in the k-th cluster. The maximum can be computed via the fixed-point iteration algorithm [27]. The process of updating cluster hyperparameters is given as follows:

$$\begin{aligned} \vartheta^*_{k,l} &= \vartheta_{k,l} \frac{\sum_{j\in U_k} \frac{CC_{j,l}}{CC_{j,l} - 1 + \vartheta_{k,l}}}{\sum_{j\in U_k} \frac{CC_j}{CC_j - 1 + \sum^L_{l'=1} \vartheta_{k,l'}}} \\ \iota^*_{k,d} &= \eta_{k,d} \frac{\sum_{j\in U_k} \frac{CT_{j,d}}{CT_{j,d} - 1 + \iota_{k,d}}}{\sum_{j\in U_k} \frac{CT_j}{CT_j - 1 + \sum^2_{d'=1} \iota_{k,d'}}} \end{aligned} \tag{15}$$

Based on the updates of cluster hyperparamters, we adjust the cluster assignment ϱ_j by maximizing the likelihood function. In particular, maximizing the likelihood function $p(CC_k, CT_k|\varrho_j = k)$ is given as follows:

$$\varrho^*_j = \text{argmax}_k p(CC_j|\varrho_j = k) p(CT_j|\varrho_j = k) \tag{16}$$

The CRLB is defined as the inverse of Fisher information: $CRLB = E^{-1}$, where E is the Fisher information of the estimation parameter. Using the likelihood function from Equation (7) and the results of estimation parameters from Table II, we can compute CRLB to quantify the accuracy of our solution using a similar method we developed in [39]. Using the computed CRLB, we can compute the confidence interval a_i on the local attractiveness estimation of each venue. We further define Q_j to represent the estimation accuracy of a user's localness. Given the Venue-User matrix MN, Q_j can be computed as: $Q_j = 1 - \frac{\sum_{i\in MN_j} a_i}{|MN_j|}$ where MN_j is the set of venues user N_j has check-in points and a_i is the derived confidence interval on the local attractiveness estimation of venue M_i.

We optimize the user's localness identification as follows: if a user N_j's localness estimation accuracy Q_j is less than a certain threshold and have dependency with others, we compute an optimized localness of N_j by leveraging the clustering results derived earlier. In particular, based on the above definitions, we define users within the same cluster as dependent users. For the user N_j, we use De_j to represent the set of her/his dependent users in the same cluster. Finally, we define the objective function as: $f^* = \sum_{j\in N} \sum_{j'\in De_j} |\Phi^*_j - \Phi_{j'}| \cdot w(N_j, N_{j'})$, where $w(N_j, N_{j'})$ is the dependency strength between user N_j and $N_{j'}$, which is computed as the number

of common venues the two users visited in our model. This problem can be solved using weighted median algorithm [7].

B. User Home Location Estimation

Finally, we formulate the problem of user's home location estimation as an optimization problem by incorporating the users' localness and dependency obtained from the previous subsection. In particular, we can estimate the home location of each identified local user by leveraging the home locations of his/her dependent users in the same dependent group (i.e., cluster). We define a distance function $dis(N_j, N_{j'})$ to represent the geographical distance between the home location of user N_j and user $N_{j'}$. We define the objective function of our estimation problem as follows: $f_{hl} = \sum_{j \in N} \sum_{j' \in De_j} dis(N_j, N_{j'}) \cdot w(N_j, N_{j'})$. The goal is to find the home location (ω_j, δ_j) for each user in M that minimizes the defined objective function. To simplify the notations used in the distance function $dis(N_j, N_{j'})$ based on cartesian coordinate system [8], we define a few additional notations:

$$\begin{aligned} x &= sin(\frac{\pi}{2} - \omega)cos(\delta) = cos(\omega)cos(\delta) \\ y &= sin(\frac{\pi}{2} - \omega)sin(\delta) = cos(\omega)cos(\delta) \\ z &= cos(\frac{\pi}{2} - \omega) = sin(\omega) \end{aligned} \tag{17}$$

Using the defined terms, we can convert our problem to the following convex optimization one:

$$\begin{aligned} &\min\ f_{hl}(x, y, z) = \\ &\sum_{j' \in De_j} [(x_{j'} - x)^2 + (y_{j'} - y)^2 + (z_{j'} - z)^2] \\ &s.t.\ x^2 + y^2 + z^2 = 1 \end{aligned} \tag{18}$$

The convex optimization in Equation (17) can be rewritten as:

$$\begin{aligned} \min\ f_{hl}(x, y, z) \quad &= 2 \sum_{j' \in De_j} w_{j'} - 2 \sum_{j' \in De_j} w_{j'} x - \\ &2 \sum_{j' \in De_j} w_{j'} y - 2 \sum_{j' \in De_j} w_{j'} z \end{aligned} \tag{19}$$

We note that $f_{hl}(x, y, z)$ is a linear function and the minimum value can be obtained at the extreme point on the sphere $x^2 + y^2 + z^2 = 1$. Therefore, we can obtain the corresponding (ω_j, δ_j) that minimizes f_{hl} based on the derivations from Equation (17).

IV. Evaluation

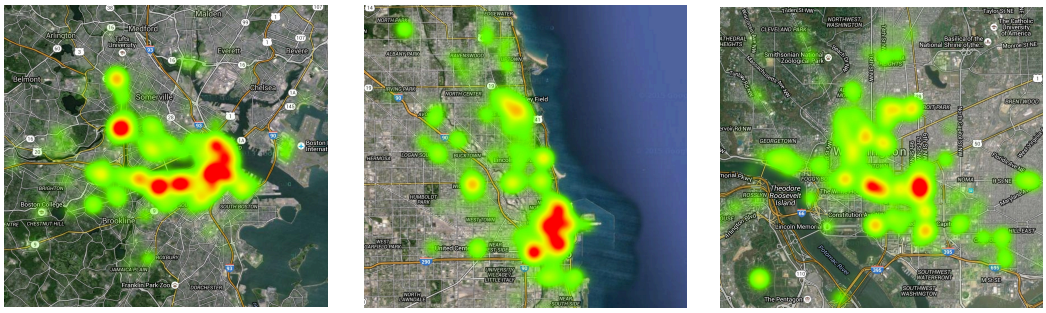
In this section, we evaluate the performance of the *WAYF (Where Are You From)* scheme using three real-world data traces collected from a social media platform (i.e., Foursquare). We demonstrate the effectiveness of our proposed scheme on these data traces and compare the performance of our scheme to the state-of-the-art baselines. In the rest of this section, we first present the experiment settings and data pre-processing steps that were used to prepare the data for performance validation. Then we introduce the state-of-the-art techniques and evaluation metrics we used in our experiments. Finally, we present the evaluation results that demonstrate the *WAYF* scheme can infer the home location of users more accurately than the compared baselines.

A. Experimental Setups and Evaluation Metrics

1) Data Trace Statistics: In this paper, we evaluate WAYF 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). 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 also note that *such ground truth home location information is not globally available for all users in all cities* [33], which is the main motivation to develop *WAYF* scheme to infer the home location of people from their check-in points. In the evaluation, we selected the data traces from three cities in U.S where the ground truth information is available: Boston, Chicago and Washington D.C.. Figure 2 shows the heat map of venues in three cities. The statistics of these traces are summarized in Table III.

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 WAYF scheme: (i) Venue-User Matrix (MN Matrix) Generation; (iii) Time Lenght Vector (TL Vector) Generation; Check-Category Matrix (CC Matrix) Generation and Check-Time Matrix (CT Matrix) Generation. They are summarized as follows:

- *Venue-User Matrix Generation*: We generate the MN 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 N_j visits venue M_i in the data trace, we set the element MN_i^j in MN to 1 and 0 otherwise.
- *Time Length Vector Generation*: For simplicity, we generate a binary time vector T based on the time length of user's check-in points. In particular, if the time length of user's check-in points (i.e., the time difference between the first and last check-in points) is larger than a certain threshold, we set the corresponding element h_j in vector TL as 1. Otherwise, we set the h_j as 0.
- *Check-Category Matrix Generation*: We generate the CC matrix by counting the number of check-in points of users on different categories of venues in the city. In particular, we set the element CC_j^l in matrix CC as the number of check-in points of user N_j on venues with the category of l.
- *Check-Time Matrix Generation*: We generate the CT matrix by counting the number of check-in points of users at different time of a day. In particular, we set the element CT_j^d in matrix CT as the number of check-in points of user N_j during daytime (i.e., $d = 1$) or nighttime (i.e., $d = 2$).



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

Figure 2. Heat Maps of Venues in Three Cities

Table III
DATA TRACES STATISTICS

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

After the above pre-processing steps, we generate all the inputs (i.e., MN Matrix, TL Vector, CC Matrix and CT Matrix) for the *WAYF* scheme.

3) Evaluation Metric: In our evaluation, we define two metrics to evaluate the performance of the *WAYF* scheme. The first metric is *Average Error Distance for Top-k% Users (AED@Top-k%)*. In particular, we denote g_j and $\hat{g_j}$ as the user N_j's real and estimated home location respectively. $dis(g_j, \hat{g_j})$ is defined as the distance between g_j and $\hat{g_j}$. The Top-k% users are the top k% users who are ranked by $dis(g_j, \hat{g_j})$. 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 [9]. Particularly, $ACC@M$ is used to measure the fraction of users who can be accurately geo-located within 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 the above metrics (i.e., AED-Top-k% and $ACC@M$) are given in Table IV.

Table IV
METRIC DEFINITIONS

Metric	Definition
AED@Top-k%	$\frac{\sum_{j \in N} dis(g_j, \hat{g_l}) \vert Rank(j) < k\%}{\vert N \vert}$
$ACC@M$	$\frac{\{j \vert j \in N \wedge dis(g_j, \hat{g_j}) < M\}}{\vert N \vert}$

B. Performance Validation

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

- *HLI*: it proposes a machine learning approach that locate people's home location by integrating the spatial and temporal features of people's trajectories [14].
- *MLP*: it proposes a generative probabilistic approach that infers a user's locations by leveraging the home locations of the user's online friends [23].
- *UHLI*: it develops an unsupervised approach to solve the problem of inferring the home locations of people by exploring the localness of users and influence scope of venues [19].
- *FM*: it infers a user's location by utilizing the home locations of the people that visit similar places as the user [5].
- *UDI*: it proposes a unified framework for profiling users' home locations by exploring both social network between users and influence probabilities of different locations [24].
- *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 [25].
- *Aver*: it simply computes the home location of a user by taking the average of the coordinates of all places the user visited.

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 Boston data trace are shown in Table V. We observe that the proposed *WAYF* scheme

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

	Top-k% Accurate Geo-locating Users				
Alg	20%	40%	60%	80%	100%
WAYF	**0.43**	**0.68**	**1.39**	**2.81**	**20.09**
HLI	0.63	1.25	2.03	7.66	79.99
MLP	0.54	1.05	1.92	10.22	86.89
UHLI	0.50	1.02	1.52	2.93	25.49
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 VI
ESTIMATION ACCURACY ON BOSTON DATA TRACE IN TERMS OF ACC@Y MILES

	M (mile)				
Alg	1	3	5	7	9
WAYF	**0.240**	**0.564**	0.663	**0.724**	**0.754**
HLI	0.168	0.462	0.583	0.615	0.643
MLP	0.202	0.474	0.569	0.602	0.631
UHLI	0.221	0.557	**0.673**	0.713	0.743
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

outperforms the compared baselines over different values of $k\%$. Specifically, it has the smallest average error distance on the estimation of users' home location. 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 224 miles).

Furthermore, we also evaluate the estimation performance of all schemes in terms of $ACC@M$. The evaluation results on Boston data trace are shown in Table VI. We observe that the proposed *WAYF* scheme also outperforms the compared baselines over most values of M. In particular, 24% and 56% of users can be geo-located within 1 and 3 miles of their real home locations respectively.

We also evaluate the estimation performance of all schemes in terms of AED-Top-$k\%$ and $ACC@M$ on Chicago and Washington D.C data trace. The results on Chicago data trace in terms of AED-Top-$k\%$ and $ACC@M$ are shown in Table VII and Table VIII respectively. The results on Washington D.C data trace in terms of AED-Top-$k\%$ and $ACC@M$ are shown in Table IX and Table X respectively. Similar results are observed in those tables. The performance improvements of *WAYF* are achieved by explicitly incorporating the localness of users and dependency between users based on their check-in behaviors from crowdsourcing data into an optimal home location estimation solution.

The above evaluation results from real world data traces demonstrate that the proposed *WAYF* scheme can effectively infer the home locations of users compared to the state-of-the-art techniques.

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

	Top-k% Accurate Geo-locating Users				
Algorithm	20%	40%	60%	80%	100%
WAYF	**0.46**	**0.83**	**1.39**	**2.77**	**9.81**
HLI	0.77	1.68	2.98	5.22	54.32
MLP	1.10	2.11	3.30	4.79	64.79
UHLI	0.67	1.11	1.79	2.77	29.59
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 VIII
ESTIMATION ACCURACY ON CHICAGO DATA TRACE IN TERMS OF ACC@M MILES

	M (mile)				
Algorithm	1	3	5	7	9
WAYF	**0.232**	**0.492**	**0.688**	0.760	0.789
HLI	0.145	0.332	0.494	0.565	0.681
MLP	0.083	0.295	0.447	0.606	0.714
UHLI	0.190	0.482	0.681	**0.766**	**0.808**
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

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

	Top-k% Accurate Geo-locating Users				
Algorithm	20%	40%	60%	80%	100%
WAYF	**0.46**	**0.72**	**1.08**	2.49	**16.79**
HLI	0.80	1.37	2.29	9.33	82.97
MLP	0.75	1.22	2.28	15.19	93.83
UHLI	0.66	0.97	1.42	**2.43**	21.71
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

V. RELATED WORK

Crowdsourcing has emerged as a new application paradigm of collecting data measurements about the physical world from a crowd of humans or devices on their behalf [3], [1]. This emerging paradigm is now widely used in many real-world applications and systems [41], [20], [18], [37], [36]. Recent research work starts to address new challenges in crowdsourcing applications such as incentive mechanism design [43] and privacy protection [30] and data reliability [38], [15]. An emerging problem of user profile inference arises in crowdsourcing applications due to the proliferation of mobile

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

Algorithm	M (mile)				
	1	3	5	7	9
WAYF	**0.238**	**0.563**	0.676	**0.774**	**0.795**
HLI	0.142	0.455	0.549	0.621	0.655
MLP	0.150	0.426	0.539	0.599	0.627
UHLI	0.206	0.560	**0.696**	0.765	0.794
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

sensing devices (e.g., smartphones) and the rapid growth of location-based social network services (e.g., Foursquare, Google Places, Gowalla) [10]. These services empower common individuals to easily share their location and visiting information at scale. To address this emerging problem, this paper develops a novel scheme to accurately infer the home locations of people by using sparse and noisy check-in points contributed by the crowd.

Previous work has made significant progress on user profiling [28], [2], [22]. 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 [28]. 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 [2]. Li et al. studied the problem of user profiling by capturing the correlation between attributes and social connections of the user's ego networks [22]. 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 without the requirement on the knowledge of the user's social connections and content (e.g., tweets, blogs) they generate.

In addition, our work is also related to user behavior understanding based on their home locations. For example, a content-based approach was proposed by Cheng et al. [9] 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., New York) 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 [32]. Furthermore, user's home location has been considered as a key factor to compute the distance between social users in a pairwise fashion [35], [34]. 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 [14], [23], [5], [24], [26], [25], [19]. In particular, Backstrom et al. estimated a user's location by exploring both the geographic and social relationship between users [5]. Li et al. [24] 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 [23]. McGee [26] proposed a network-based approach for location estimation by correlating the social tie strength with physical proximity. Hu et al. [14] designed a machine learning method to capture the inherent properties of users' homes by exploring their mobility features. Mahmud et al. [25] proposed a hierarchical ensemble algorithm to predict the home location of users by leveraging the domain knowledge and advanced classifications. Huang et al. studied the problem of user's location inference using an unsupervised approach that considers the localness of users and influence scope of venues [19]. In this work, we develop an semi-supervised framework to address the problem of profiling the home locations of local users on LBSNs by leveraging both the localness of users and the dependency between users based on their check-in behaviors.

Maximum likelihood estimation (MLE) framework has been widely used in the domain of sensor networks [40], [29], [31], [16], [42]. For example, Pereira et al. proposed a diffusion-based MLE algorithm for distributed estimation in WSN in the presence of noisy measurements and data faults [31]. Eric et al. deisgned a MLE based approach to aggregate the signals from noisy measurements at remote sensor nodes to a fusion center without any inter-sensor collaborations [29]. Wang et. al [40] developed an estimation theoretical framework to solve the truth discovery problem in social sensing applications. In contrast, this paper studied a new problem of inferring the user's home location using sparse and noisy crowdsourcing data.

VI. CONCLUSION

This paper proposes a semi-supervised approach to infer the home locations of crowd sensors by exploring the sparse and noisy crowdsourcing data from location based social networks (LBSN). In particular, we develop the *Where Are You From (WAYF)* scheme to accurately estimate users' home locations by explicitly exploring the localness of users and the dependency between users based on their check-in behaviors. We perform extensive experiments to evaluate the performance of our new scheme using the real-world datasets collected from Foursquare. The evaluation results demonstrated the effectiveness of our new scheme in profiling the home location of users. The results of our paper are important because they can directly contribute to *crowd data source profiling* in participatory sensing and other crowdsourcing applications.

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

[2] 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*. Springer, 2011.

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

[4] 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 Minging (KDD)*, pages 114–122. ACM, 2011.

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

[6] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, pages 199–208. ACM, 2012.

[7] D. Brownrigg. The weighted median filter. *Communications of the ACM*, pages 807–818, 1984.

[8] K.-t. Chang. *Introduction to geographic information systems*. McGraw-Hill Higher Education Boston, 2006.

[9] 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.

[10] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *International Conference on Knowledge Discovery and Data Minging (KDD)*, pages 1082–1090. ACM, 2011.

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

[12] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Mobisys*, pages 31–42. ACM, 2003.

[13] J. Hamm, A. C. Champion, G. Chen, M. Belkin, and D. Xuan. Crowd-ml: A privacy-preserving learning framework for a crowd of smart devices. In *International Conference on Distributed Computing Systems (ICDCS)*, pages 11–20. IEEE, 2015.

[14] T.-r. Hu, J.-b. Luo, H. Kautz, and A. Sadilek. Home location inference from sparse and noisy data: models and applications. In *International Conference on Data Mining (ICDM)*, pages 1382–1387. IEEE, 2015.

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

[16] 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. ACM/IEEE, 2016.

[17] 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.

[18] 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.

[19] C. Huang, D. Wang, S. Zhu, and Y. Zhang. Towards unsupervised home location inference from online social media. In *International Conference on Big Data (Big Data)*. IEEE, 2016.

[20] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. Understanding human mobility from twitter. *PloS one*, (7), 2015.

[21] L. Kong, L. He, X.-Y. Liu, Y. Gu, M.-Y. Wu, and X. Liu. Privacy-preserving compressive sensing for crowdsensing based trajectory recovery. In *International Conference on Distributed Computing Systems (ICDCS)*, pages 31–40. IEEE, 2015.

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

[23] R. Li, S. Wang, and K. C.-C. Chang. Multiple location profiling for users and relationships from social network and content. In *Very Large Data Bases (VLDB)*, pages 1603–1614. ACM, 2012.

[24] 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 Minging (KDD)*, pages 1023–1031. ACM, 2012.

[25] J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *Transactions on Intelligent Systems and Technology (TIST)*, (3):47, 2014.

[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] T. Minka. Estimating a dirichlet distribution, 2000.

[28] 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.

[29] E. J. Msechu and G. B. Giannakis. Sensor-centric data reduction for estimation with wsns via censoring and quantization. *Transactions on Signal Processing (TSP)*, pages 400–414, 2012.

[30] B. Niu, Q. Li, X. Zhu, G. Cao, and H. Li. Achieving k-anonymity in privacy-aware location-based services. In *International Conference on Computer Communications (Infocom)*, pages 754–762. IEEE, 2014.

[31] S. S. Pereira, R. Lopez-Valcarce, et al. A diffusion-based em algorithm for distributed estimation in unreliable sensor networks. *Signal Processing Letters, IEEE*, pages 595–598, 2013.

[32] 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.

[33] 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)*, 26(6):1384–1399, 2014.

[34] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. In *ICWSM*, pages 329–336. AAAI, 2011.

[35] 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 Minging (KDD)*, pages 1046–1054. ACM, 2011.

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

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

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

[39] 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)*, pages 1026–1037, 2013.

[40] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *International Conference on Information Processing in Sensor Networks (IPSN)*, April 2012.

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

[42] J. Wang, D. Wang, Y. Zhao, and T. Korhonen. Fast anti-collision algorithms in rfid systems. In *Mobile Ubiquitous Computing, Systems, Services and Technologies, 2007. UBICOMM'07. International Conference on*, pages 75–80. IEEE, 2007.

[43] X. Zhang, G. Xue, R. Yu, D. Yang, and J. Tang. Truthful incentive mechanisms for crowdsourcing. In *International Conference on Computer Communications (Infocom)*, pages 2830–2838. IEEE, 2015.