2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Exploiting Spatial-Temporal-Social Constraints for Localness Inference Using Online Social Media

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

***Abstract*—The localness inference problem is to identify whether a person is a local resident in a city or not and the likelihood of a venue to attract local people. This information is critical for many applications such as targeted ads of local business, urban planning, localized news and travel recommendations. While there are prior work on geo-locating people in a city using supervised learning approaches, the accuracy of those techniques largely depends on a high quality training dataset, which is difficult and expensive to obtain in practice. In this study, we propose to exploit spatial-temporal-social constraints from noisy online social media data to solve the localness inference problem using an *unsupervised* approach. The spatial-temporal constraint represents the correlations between people and venues they visit and the social constraint represents social connections between people. In particular, we develop a Spatial-Temporal-Social-Aware (STSA) inference framework to jointly infer i) the localness of a person and ii) the local attractiveness of a venue without requiring any training data. We evaluate the performance of STSA scheme using three real-world datasets collected from Foursquare. Experimental results show that STSA scheme outperforms the state-of-the-art techniques by significantly improving the estimation accuracy.**

***Index Terms*—Spatial-Temporal-Social Constraints, Localness Inference, Localness of People, Local Attractiveness of Venues, Online Social Media**

I. Introduction

Understanding the *localness of users* (whether a user is a local resident in a city or not) and *local attractiveness of venues* (the likelihood of a venue to attract local users) is important to many applications such as targeted ads for local business [2], urban planning [10], and localized news and travel recommendations [21]. Recent years have witnessed an exponential growth of Location-Based Social Network (LBSN) services where people voluntarily share their location information through mobile applications. These services allow users to explicitly or automatically report the GPS coordinates (often called "check-in points") of their visited venues in a city. Examples of such services include Foursquare, Yelp, Gowalla, Instagram, and Google Places. In this paper, we develop a novel principled approach to accurately infer the localness of users and the local attractiveness of venues using publicly available LBSN data.

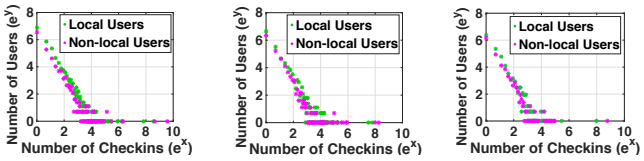
There exist prior works on geo-locating people in a city using online social network information [3], [12], [19], [22]. Most of these previous studies used *supervised* learning approaches, which largely depend on high quality training datasets to predict a person's home location. However, such training datasets are difficult and expensive to obtain in practice since people usually are reluctant to publicize their real home locations [20]. Furthermore, since most of LBSNs have set up rate limits on their APIs for data collection and sharing [43], it is very challenging to collect complete check-in traces of users at all venues they visited. A more practical scenario is that only partial check-in points of users in a city for a certain period of time can be obtained for analysis. In this paper, we prove the hypothesis that it is possible to use such sparse and incomplete check-in data trace to accurately estimate the localness of people and local attractiveness of venues using an *unsupervised* learning approach.

Several challenges exist in order to address the problem of inferring localness of users and local attractiveness of venues: (i) *Sparse Data Challenge*: the spatial-temporal data (i.e., check-in points) are often incomplete and sparse: a person might not check in at every venue he/she visits in a city or even turn off the check-in function sometime due to privacy concerns; (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 (e.g., tourists). The check-in points are just GPS coordinates with timestamps, which themselves do not provide much useful information to separate local people from the non-local ones.

A simple solution is to differentiate local users from non-local ones by analyzing the statistics of their check-in traces such as the number of check-in points, the length of check-in trace (the time difference between first and last check-in point of a user) and the activity range (the largest distance among all check-in points of a user). To investigate the feasibility of the simple solution, we plotted the distribution of i) the number of check-in points per user; ii) length of check-in trace; and iii) user's activity range of three real world datasets collected from Chicago, Washington D.C. and Boston on Foursquare in Figure 1, Figure 2 and Figure 3 respectively. We observed that local and non-local users have very similar distributions on all of three metrics. The above observation suggests that it is challenging to solve the local inference problem by simply analyzing the statistics of user's check-in trace.

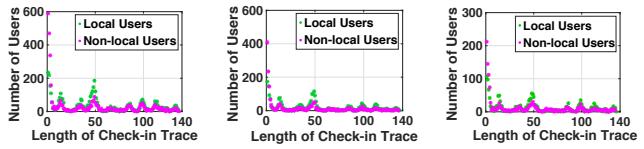
To address the above challenges, we develop a new unsupervised approach to infer the localness of users and local attractiveness of venues in a city by exploiting Spatial-Temporal-

IEEE/ACM ASONAM 2016,August 18-21, 2016, San Francisco, CA, USA 978-1-5090-2846-7/16/$31.00



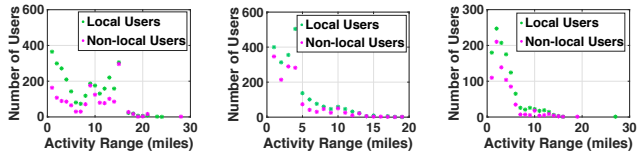
(a) Chicago (b) Washington D.C. (c) Boston

Figure 1: Distribution of Number of Check-in Points for Local and Non-Local Users



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

Figure 2: Distribution of Length of Check-in Trace for Local and Non-Local Users



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

Figure 3: Distribution of Activity Range for Local and Non-Local Users

Social constraints from social media. In particular, we develop a Spatial-Temporal-Social-Aware (STSA) framework to infer the localness of people by considering the venues they visited and their activity range (spatial), the time length of their check-in traces (temporal) and the social connections between people (social). Our framework can jointly estimate i) the localness of a person and ii) the attractiveness of a venue without requiring any training data. We evaluate our new approach using three real-world datasets collected from Foursquare. The results showed that STSA scheme outperforms the state-of-the-art techniques by significantly improving the estimation accuracy. The results of this paper are important because they provide accurate estimations on the localness of users and local attractiveness of venues, which are important elements in many recommendation and smart city applications [16], [31].

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 [11]. Additionally, there exists a rich set of literature on the topic of protecting user's privacy in online social media applications [40]. 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 localness inference problem a trivial problem to solve). Instead, we only used publicly available data from LBSNs with the goal to develop a new unsupervised localness inference scheme as an open-source resource for the research community.

The main contributions of this paper can be summarized as follows:

- We study the localness inference problem using an *unsupervised* approach by exploiting *spatial-temporal-social* constraints extracted from the sparse and noisy online social media data. (Section III)
- We develop a principled STSA framework that allows us to derive an *optimal* solution that is most consistent with users' check-in data traces and their social connections. (Section IV)
- We perform extensive experiments to compare the performance of our STSA framework and other the-sate-of-the-art baselines using three large scale real-world data sets collected from Foursquare. Experimental results demonstrate that the proposed approach outperforms existing methods by significantly improving estimation accuracy. (Section V)

II. Related Work

User Profiling. User profiling is an important problem in social media analysis. Previous works have made significant progress towards addressing this problem [1], [8], [17], [28]. 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 [1]. Dong et al. studied human interactions on demographics profiles by investigating mobile social network [8]. Li et al. studied the problem of user profiling by capturing the correlation between attributes and social connections of the user's ego networks [17]. However, none of these techniques can be directly applied to infer the localness of users and local attractiveness of venues 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 localness inference problem by using the publicly available check-in data trace from LBSN.

Localized Recommendation Systems. Our work is also related to localized recommendation system [5], [9], [21], [41]. In particular, Macedo et al. solved the local event recommendation problem using a learning-to-rank approach that leverages multiple context-aware recommendation models as features [21]. Chen et al. proposed a greedy algorithm that leverages the information coverage to encode the location categories in its recommendations [5]. Yin et al. developed a LCA-LDA probabilistic model to infer both the item content and the local preference in its recommendation [41]. Gao et al. studied the content information on LBSNs for POI recommendation [9]. Our work is complementary to the above recommendation systems in the sense that the correctly estimated localness of users and local attractiveness of venues are critical for more accurate and effective localized recommendations.

Truth Discovery in Social Sensing. Our work is also related to the work on truth discovery in social sensing applications [13]–[15], [32]–[36]. In particular, Wang et. al developed an estimation theoretical framework to solve the truth discovery problem (i.e., joint estimation of the source reliability

and claim correctness without prior knowledge on either of them) in social sensing applications [32]–[35]. Chao et. al extended the truth discovery framework to consider additional features of the problem such as time, location, confidence and topic relevance [13]–[15], [36]. Marshall et al. further explored the semantic dimension of the truth discovery problem and considered features such as emotion, claim hardness and mood sensitivity in the truth discovery solutions [24]–[26], [39]. This paper leveraged the insights of the above truth discovery solutions and addressed a new problem of inferring localness of users and local attractiveness of venues using online social media data. The proposed framework incorporates spatial, temporal and social constraints in the solution.

Geo-locating People. Finally, our work is closely related to the works that address the problem of geo-locating people in a city [3], [6], [19], [22]. For example, Cheng et al. proposed a probabilistic framework to estimate a Twitter user's location at the city level purely based on the content of the user's tweets [6]. 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. T 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 used supervised learning approaches, which require sufficient training data with complete spatial-temporal information to accurately estimate an individual's home location. In contrast, this paper developed an unsupervised learning approach to address the problem of inferring the localness of user and the local attractiveness of venues that does not require any training data.

III. Problem Formulation

In this section, we formulate the problem of inferring the localness of users and local attractiveness of venues in a city as a maximum likelihood estimation problem. In particular, we consider a set of X venues in a city, namely, V_1, V_2,...,V_X, which have check-in points from a set of Y users, namely, U_1, U_2,..,U_Y. Let V_x represent the x^{th} venue and U_y represent the y^{th} user. We define $U_y = 1$ if the user is a local resident of the city and $U_y = 0$ if he/she is not. We further define the following inputs to our model.

- *Definition 1*. **Check-in Matrix CI**. We define Check-in Matrix $\mathbf{CI}_{X\times Y}$ to indicate *who visit where*. In particular, $CI_{x,y} = 1$ indicates that user U_y has check-in points at venue V_x and $CI_{x,y} = 0$ otherwise.
- *Definition 2*. **Temporal Vector T**. We define a Temporal Vector $\mathbf{T}_Y$ to represent the time length of user's check-in points. In particular, $t_y = k$ denotes that user U_y's check-in points in a city lasts for k days.
- *Definition 3*. **Spatial Vector S**. We define a Spatial Vector $\mathbf{S}_Y$ to represent the activity range of user's check-in points. In particular, $s_y = h$ denotes that the the largest distance among all check-in points of the user U_y is h miles.
- *Definition 4*. **Social Relationship Matrix SR**. We define a Social Relationship Matrix $\mathbf{SR}_{Y\times Y}$ 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.

First, let us define a few important terms that will be used in the problem formulation. We denote the *local attractiveness* of a venue V_x as la_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 and activity range of his/her check-in points, we define $la_{x,k,h}$ as the probability of a venue V_x to attract local users whose check-in points in a city last for k days and the activity range is h miles. Formally, la_x and $la_{x,k,h}$ can be given as:

$$\begin{aligned} la_x &= \Pr(U_y = 1 | CI_{x,y} = 1) \\ la_{x,k,h} &= \Pr(U_y = 1 | CI_{x,y} = 1, t_y = k, s_y = h) \end{aligned} \quad (1)$$

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

$$la_x = \Sigma_{k=1}^{K}\Sigma_{h=1}^{H} la_{x,k,h} \times \frac{r_{x,k,h}}{r_x} \quad k \in [1, K]; h \in [1, H] \quad (2)$$

where $r_x = \Pr(CI_{x,y} = 1)$ and $r_{x,k,h} = \Pr(CI_{x,y} = 1, t_y = k, s_y = h)$. Note that $r_x = \Sigma_{k=1}^{K}\Sigma_{h=1}^{H} r_{x,k,h}$. Let us further define $E_{x,k,h}$ to denote the probability of a *local* user whose check-in points in a city lasts for k days and activity range is h miles visits a venue V_x. Similarly, let $F_{x,k,h}$ denote the probability of a *non-local* user whose check-in points in a city lasts for k days and activity range is h miles visits a venue V_x. $E_{x,k,h}$ and $F_{x,k,h}$ are formally defined as:

$$\begin{aligned} E_{x,k,h} &= \Pr(CI_{x,y} = 1, t_y = k, s_y = h | U_y = 1) \\ F_{x,k,h} &= \Pr(CI_{x,y} = 1, t_y = k, s_y = h | U_y = 0) \end{aligned} \quad (3)$$

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

$$E_{x,k,h} = \frac{la_{x,k,h} \times r_{x,k,h}}{q}, \quad F_{x,k,h} = \frac{(1 - la_{x,k,h}) \times r_{x,k,h}}{1 - q} \quad (4)$$

Inferring the Localness of Users and Local Attractiveness of Venues. The problem of inferring the localness of users and local attractiveness of venues is formulated as follows: given the Check-in Matrix $\mathbf{CI}_{X\times Y}$, Temporal Vector $\mathbf{T}_Y$, Spatial Vector $\mathbf{S}_Y$ and Social Relationship Matrix $\mathbf{SR}_{Y\times Y}$, the goal is to jointly estimate both the localness of each user and the probability of each venue in a city to attract local people. Formally, we compute:

$$\begin{aligned} &\forall y, 1 \le y \le Y : P(U_y = 1 | CI, T, S, SR) \\ &\forall x, 1 \le x \le X : P(U_y = 1 | CI_{x,y} = 1) \end{aligned} \quad (5)$$

IV. The Spatial-Temporal-Social-Aware (STSA) Framework

In this section, we present the Spatial-Temporal-Social-Aware (STSA) framework to solve the localness inference problem by exploring the spatial-temporal-social constraints

embedded in the online social media data. The framework consists of two major components: *Spatial-Temporal Modeling* and *Social-Aware Localness Inference*. The overview of the framework is shown in Figure 4.

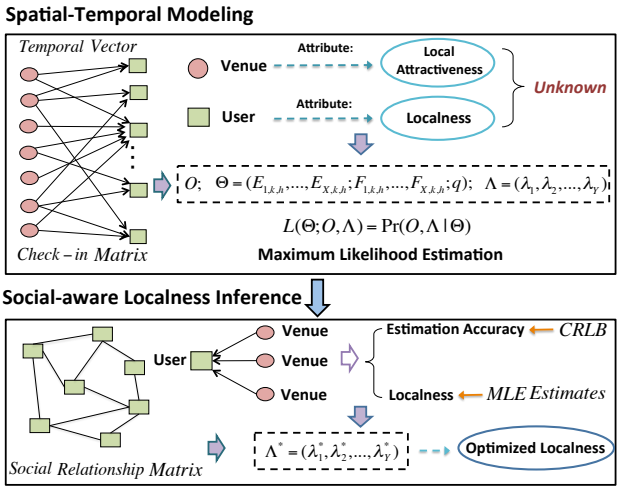


Figure 4: The STSA Framework

A. Spatial-Temporal Modeling

We first present the Spatial-Temporal Modeling component of the STSA framework. EM is an optimization scheme that is commonly used to solve the MLE problem where unobserved latent variables exist in the model [7]. Specifically, it iterates between two key steps: expectation step (E-Step) and maximization step (M-step). In E-step, it computes the expectation of the log likelihood function based on the current estimates of the model parameters. In M-step, it computes the new estimates of the model parameters that maximize the expected log-likelihood function in E-step.

Given the terms we defined in the previous section, the likelihood function that describes the user's check-in behavior together with the spatial-temporal constraints is given follows:

$$L(\Theta;O,\Lambda)=\Pr(O,\Lambda|\Theta)$$
$$=\prod_{y=1}^{Y}\Pr(\lambda_y|O_y,\Theta^{(n)})\times\prod_{x=1}^{X}\prod_{k=1}^{K}\prod_{h=1}^{H}\Psi_{x,y,k,h}\times\Pr(\lambda_y) \quad (6)$$

where $\Theta=(E_{1,k,h},...,E_{X,k,h};F_{1,k,h},...,F_{X,k,h};q)$. O is the observed data (i.e., Matrix CI, Vector T and S). Λ is a set of latent variables that indicate whether a user is local or not. More specially, we have 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. Additional variables are defined in Table I.

Table I: Notations for Spatial-Temporal Modeling

$\Psi_{x,y,k,h}$	$\Pr(\lambda_y)$	$\Lambda(n,y)$	Spatial-Temporal Constrains
$E_{x,k,h}$	q	$\Pr(U_y=1\|O_y,\Theta^{(n)})$	$CI_{x,y}=1,t_y=k,s_y=h,\lambda_y=1$
$1-\sum_{k=1}^{K}\sum_{h=1}^{H}E_{x,k,h}$	q	$\Pr(U_y=1\|O_y,\Theta^{(n)})$	$CI_{x,y}=0,t_y=k,s_y=h,\lambda_y=1$
$F_{x,k,h}$	$1-q$	$\Pr(U_y=0\|O_y,\Theta^{(n)})$	$CI_{x,y}=1,t_y=k,s_y=h,\lambda_y=0$
$1-\sum_{k=1}^{K}\sum_{h=1}^{H}F_{x,k,h}$	$1-q$	$\Pr(U_y=0\|O_y,\Theta^{(n)})$	$CI_{x,y}=0,t_y=k,s_y=h,\lambda_y=0$

Given the above mathematical formulation, we develops an Expectation and Maximization (EM) scheme to solve the problem. The E-step is derived as follows:

$$Q(\Theta|\Theta^{(n)})=E_{\Lambda|O,\Theta^{(n)}}[logL(\Theta;O,\Lambda)]$$
$$=\sum_{y=1}^{Y}\Lambda(n,y)\times\sum_{x=1}^{X}(log\Psi_{x,y,k,h}+log\Pr(\lambda_y)) \quad (7)$$

where $\Lambda(n,y)$ is defined in Table I and n is the iteration index.

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

$$E^*_{x,k,h}=\frac{\Sigma_{y\in CV_{x,k,h}}\Pr(\lambda_y=1|O_y,\Theta^{(n)})}{\Sigma_{y=1}^{Y}\Pr(\lambda_y=1|O_y,\Theta^{(n)})}$$
$$F^*_{x,k,h}=\frac{\Sigma_{y\in CV_{x,k,h}}(1-\Pr(\lambda_y=1|O_y,\Theta^{(n)}))}{\Sigma_{y=1}^{Y}(1-\Pr(\lambda_y=1|O_y,\Theta^{(n)}))}$$
$$q^*=\frac{\Sigma_{y=1}^{Y}\Pr(\lambda_y=1|O_y,\Theta^{(n)})}{Y} \quad (8)$$

where $CV_{x,k,h}$ is the set of users who visit the venue V_x and the check-in points of these users last for k days and the activity range of those users is h miles.

B. Social-Aware Localness Inference

In this section, we demonstrate how we can 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 [29]. Using the likelihood function from Equation (6) and the results of estimation parameters from Equation (8), we can compute CRLB to quantify the accuracy of our solution using a similar method we developed in [37].

In particular, we can assess the estimation accuracy of the estimation on la_x by computing its confidence bounds. Formally, the confidence bounds of la_x are given as:

$$(\hat{la}_x^{MLE}-c_p\sqrt{var(\hat{la}_x^{MLE})},\hat{la}_x^{MLE}+c_p\sqrt{var(\hat{la}_x^{MLE})}) \quad (9)$$

where cp is the standard score of confidence level p. $var(\hat{la}_x^{MLE})$ is the estimation variance on la_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 EA_y to represent the estimation accuracy of a user's localness. Given the Check-in matrix CI, EA_y can be computed as:

$$EA_y=\frac{\sum_{x\in CV_y}(cb_x)}{|CV_y|} \quad (10)$$

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 EA_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:

$$f = \sum_{y \in SU} \sum_{y' \in SR_y} |\Lambda_y^* - \Lambda_y'| \cdot w(y, y') \tag{11}$$

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]. We summarize the STSA scheme in Algorithm 1.

Algorithm 1 STSA Algorithm

Input: Check-in Matrix CI, Temporal Vector T, Spatial Vector S and Social Relationship Matrix SR
Output: Estimations of Venue's Local Attractiveness and User's Localness

Initialize Θ ($E_{x,k} = r_{x,k}, F_{x,k} = 0.5 \times r_{x,k}$, $q \in (0, 1)$)
$n = 0$
repeat
  $n = n + 1$
  for Each $y \in U$ **do**
    compute $\Pr(\lambda_y = 1 | O_y, \Theta^{(n)})$
  end for
  for Each $x \in V$ **do**
    compute $(E_{x,k})^{(n)}, (F_{x,k})^{(n)}, (q)^{(n)}$
  end for
until $\Theta^{(n)}$ and $\Theta^{(n-1)}$ converge
Let $(\Lambda_y)^c$ = converged value of $\Pr(\lambda_y = 1 | O_y, \Theta^{(n)})$
for Each $x \in V$ **do**
  compute ab_x based on Equation (9)
end for
for Each $y \in SU$ **do**
  compute EA_y based on Equation (10)
  if $EA_y \geq 0.5$ **then**
    compute Λ_y^* in Equation (11)
    $\Lambda_y \leftarrow \Lambda_y^*$
  end if
end for
for Each $y \in U$ **do**
  if $(\Lambda_y)^c \geq threshold\ value$ **then**
    user U_y is local
  else
    user U_y is non-local
  end if
end for
for Each $x \in V$ **do**
  calculate $(la_{x,k,h})^*$ from converge values of $(E_{x,k,h})$, $(F_{x,k,h})$ and (q) based on Equation (2)
end for

V. Evaluation

In this section, we conduct experiments to evaluate the performance of the *Spatial-Temporal-Social-Aware (STSA)* scheme on three real-world data traces collected from a location-based social network service: Foursquare. We demonstrate the effectiveness of our proposed framework on these data traces and compare the performance of our scheme to the state-of-the-art baselines. In the rest of this section: (i) we present the experiment settings and data pre-processing steps that were used to prepare the data for evaluation. (ii) We introduce the state-of-the-art baselines and evaluation metrics we used in our experiments. (iii) We present the evaluation results that demonstrate the *STSA* scheme can estimate the localness of users and local attractiveness of venues in a city more accurately than the compared baselines.

A. Experiment Setups and Evaluation Metrics

1) Data Trace Statistics: In this paper, we evaluate our proposed 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 contain home location information of users, which serves as the ground truth to decide the localness of users in our evaluation. One should note that such home location information is not available for all users in all cities [31], which is the main motivation to develop *STSA* scheme to infer the localness of users and local attractiveness of venues in a city 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 [1]: Chicago, Washington D.C. and Boston. The statistics of these traces are summarized in Table II.

Table II: Data Traces Statistics

Data Trace	Chicago	Washington D.C.	Boston
Number of Users	31,615	17,070	12,804
Number of Venues	2,529	1,932	1,478
Number of Check-ins	48,605	25,722	18,296

2) Data Pre-Processing: To evaluate our methods in real world settings, we went through the following data pre-processing steps: (i) Check-in Matrix (CI Matrix) Generation; (ii) Temporal Vector (T Vector) Generation; (iii) Social Relationship Matrix SR Generation.

- *Check-in Matrix Generation*: We generate the CI Matrix by associating each venue with the users who visited this venue. In particular, if user U_y visited venue V_x in the data trace, we set the element $CI_{x,y}$ in CI to 1 and 0 otherwise.
- *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, $t_y = k$ if the difference between the first and last check point of user U_j in a city is k days.
- *Activity 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 $SR_{x,y}$ in SR to 1 and 0 otherwise.

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

3) Evaluation Metric: In the experiments, we use two category of evaluation metrics to evaluate the performance of *STSA* scheme. The first category of metrics are used to evaluate the estimation accuracy of different techniques in terms of inferring the localness of users. They include *accuracy*, *precision*, *recall*, *F1-score* [23]. The second category of metric is used to evaluate the estimation accuracy of the *local attractiveness of venues* defined in Section III. We use the term *Root Mean Squared Error (RMSE)* to characterize the difference between the estimation value and ground truth value of a venue's local attractiveness. The mathematical definitions of the above metrics are given in Table III.

Table III: Metric Definitions

Metric	Definition
$Accuracy$	$\frac{TP+TN}{TP+TN+FP+FN}$
$Precison$	$\frac{TP}{TP+FP}$
$Recall$	$\frac{TP}{TP+FN}$
$F1\text{-}score$	$\frac{2\times Precison\times Recall}{Precison+Recall}$
$RMSE$	$\sqrt{\frac{\sum_{x=1}^{X}(la_x^{grouthtruth}-la_x)^2}{X}}$

In Table III, *TP*, *TN*, *FP* and *FN* represents True Positives, True Negatives, False Positives and False Negatives respectively. In our experiment, the True Positives and True Negatives are the users that are correctly classified by a particular scheme as local or non-local respectively. The False Positives and False Negatives are the non-local and local users that are misclassified to each other respectively.

B. Evaluation of Our Scheme

In this subsection, we evaluate the performance of the proposed *ULI* scheme and compare it to the state-of-the-art techniques as follows:

- *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 people that visit similar places as the user [3].
- *FL*: it proposes a network-based approach that leverages the evidence of social tie strength between users [27].
- *HLI*: it proposes a machine learning approach that locate people's home location by integrating the spatial and temporal features of people's trajectories [12].
- *Reg-EM*: it solves the localness inference problem using a similar EM approach but does not consider temporal and social information [38].
- *Average_Log*: it infers the localness of a user by considering both the location and the number of venues the user visits [30].
- *TruthFinder*: it estimates the localness of a user using a heuristic based pesudo-probabilistic model [42].
- *LC-based*: it assumes the localness of a user is reflected by the length of his/her check-in trace: the longer the length of check-in trace, the more likely the user is local.
- *AR-based*: it assumes the localness of a user is reflected by his/her activity range: the smaller the activity range, the more likely the user is local.
- *Freq-based*: it assumes the localness of a user is reflected by the number of venues he/she visited in a city: the more venues a user visits, the more likely the user is local.

STSA scheme differs from the above schemes in that it is an unsupervised approach which does not require any information on i) the home locations of the targeting users or their friends; and ii) the news or content (e.g., tweets, blogs) generated by users. Instead, it judiciously uses the venue locations, user visiting behavior and the social connections between users to solve the localness inference problem.

1) Evaluation Results: In our evaluation, we evaluated the above schemes using the ground truth information (i.e., home locations of users). In particular, a user is decided as a *local user* in a city if the user's home location is within X miles from the center of the city. To evaluate the robustness of all compared schemes, we evaluated the performance of them over different values of X (e.g., 15 miles, 30 miles, 50 miles and 100 miles), ranging from the core part of the city to the suburbs to the satellite towns.

The evaluation results of Chicago data trace are shown in Table IV. We observe that *STSA* outperforms the compared baselines in all evaluation metrics: it finds the most number of local users while keeping the falsely reported one the least. It also has the smallest error on the estimation of the local attractiveness of venues. The largest performance gain achieved by *STSA* on accuracy and F1-measure over the best performed baseline on 15 miles threshold value are 13% and 10% respectively. The results are also consistent over different X (distance threshold) values.

We repeated the above experiments on Washington D.C and Boston data trace. Considering the space limit, we only present the evaluation results for X=15 miles. The results on Washington D.C and Boston data traces are shown in Table V and Table VI respectively. In those tables, we observe that *STSA* continuously outperforms all compared baselines with nontrivial performance gains. The performance improvements of *STSA* are achieved by i) explicitly considering spatial-temporal-social constraints from social media data; ii) carefully handling the nonlinear relationship between the users localness and the venue's local attractiveness.

Table V: Estimation Results on Washington D.C. Trace (X=15 miles)

Algorithm	Accuracy	Precision	Recall	F1	RMSE
STSA	**0.691**	**0.710**	**0.825**	**0.761**	**0.387**
MLP	0.501	0.594	0.509	0.548	0.509
FM	0.493	0.601	0.504	0.548	0.522
FL	0.504	0.636	0.507	0.565	0.491
HLI	0.557	0.606	0.732	0.663	0.502
Reg-EM	0.523	0.586	0.674	0.627	0.581
Average_Log	0.547	0.643	0.538	0.586	0.600
TruthFinder	0.445	0.541	0.451	0.492	0.660
LC-based	0.627	0.722	0.607	0.659	0.426
AR-based	0.479	0.574	0.481	0.523	0.550
Freq-based	0.522	0.617	0.516	0.562	0.575

Furthermore, we validate the derived CRLBs and the performance bounds of the proposed *STSA* scheme discussed in Section IV. In particular, we randomly sampled 40 venues

Table IV: Estimation Results on Chicago Trace

	X=15 miles					X=30 miles				
Algorithm	Accuracy	Precision	Recall	F1	RMSE	Accuracy	Precision	Recall	F1	RMSE
STSA	**0.731**	**0.760**	**0.907**	**0.827**	**0.435**	**0.765**	**0.804**	**0.908**	**0.853**	**0.421**
MLP	0.449	0.648	0.433	0.519	0.649	0.476	0.681	0.447	0.539	0.662
FM	0.452	0.724	0.382	0.501	0.652	0.477	0.749	0.407	0.528	0.669
FL	0.514	0.722	0.523	0.607	0.534	0.514	0.768	0.517	0.618	0.607
HLI	0.597	0.709	0.732	0.720	0.598	0.612	0.749	0.726	0.738	0.614
Reg-EM	0.600	0.692	0.785	0.736	0.547	0.620	0.730	0.783	0.755	0.549
Average_Log	0.525	0.735	0.517	0.607	0.567	0.521	0.772	0.513	0.617	0.578
TruthFinder	0.473	0.684	0.478	0.563	0.613	0.474	0.726	0.480	0.578	0.628
LC-based	0.539	0.748	0.528	0.619	0.566	0.551	0.801	0.534	0.641	0.569
AR-based	0.528	0.739	0.518	0.609	0.554	0.522	0.773	0.512	0.616	0.565
Freq-based	0.452	0.731	0.359	0.482	0.573	0.434	0.764	0.355	0.484	0.590
	X=50 miles					X=100 miles				
Algorithm	Accuracy	Precision	Recall	F1	RMSE	Accuracy	Precision	Recall	F1	RMSE
STSA	**0.774**	**0.814**	**0.908**	**0.859**	**0.415**	**0.778**	**0.825**	**0.904**	**0.863**	**0.409**
MLP	0.483	0.690	0.448	0.544	0.663	0.456	0.660	0.430	0.520	0.667
FM	0.461	0.718	0.412	0.523	0.675	0.443	0.711	0.377	0.493	0.682
FL	0.517	0.775	0.525	0.626	0.533	0.501	0.782	0.504	0.613	0.540
HLI	0.602	0.750	0.713	0.731	0.616	0.643	0.779	0.750	0.764	0.623
Reg-EM	0.623	0.738	0.782	0.759	0.552	0.633	0.752	0.783	0.767	0.547
Average_Log	0.518	0.778	0.511	0.617	0.583	0.520	0.793	0.512	0.622	0.585
TruthFinder	0.470	0.732	0.478	0.578	0.635	0.468	0.742	0.477	0.581	0.639
LC-based	0.550	0.809	0.533	0.643	0.574	0.548	0.820	0.531	0.645	0.579
AR-based	0.521	0.782	0.512	0.619	0.565	0.524	0.798	0.514	0.625	0.561
Freq-based	0.507	0.768	0.501	0.607	0.593	0.425	0.784	0.354	0.487	0.599

Table VI: Estimation Results on Boston Trace (X=15 miles)

Algorithm	Accuracy	Precision	Recall	F1	RMSE
STSA	**0.626**	**0.652**	**0.820**	**0.726**	**0.424**
MLP	0.488	0.580	0.503	0.539	0.485
FM	0.512	0.607	0.627	0.617	0.479
FL	0.484	0.629	0.502	0.558	0.526
HLI	0.562	0.611	0.761	0.678	0.491
Reg-EM	0.540	0.596	0.743	0.661	0.527
Average_Log	0.529	0.635	0.522	0.573	0.596
TruthFinder	0.476	0.581	0.477	0.524	0.580
LC-based	0.569	0.674	0.557	0.610	0.483
AR-based	0.479	0.585	0.480	0.527	0.579
Freq-based	0.514	0.620	0.510	0.560	0.531

from each city data trace and computed the confidence bounds of the local attractiveness values of these sampled venues. The results are shown in Figure 5. We can observe that there are only 4, 3 and 3 out of 40 venues whose ground-truth local attractiveness values fall out of the 90% confidence bounds on the Chicago, Washington D.C and Boston data trace respectively. These results validate that our derived confidence bounds correctly characterize the estimation errors of venue's local attractiveness at a given confidence level.

The above evaluation results from real world data traces demonstrate that the proposed *STSA* scheme can accurately infer the localness of users and local attractiveness of venues and achieved significant performance improvements over the baselines without using any training data.

VI. Conclusion

This paper proposes an unsupervised approach to jointly infer the localness of users and local attractiveness of venues by jexploiting spatial-temporal-social constraints from the social media data. We evaluate our framework using three real-world datasets collected from Foursquare. The results showed that the STSA framework outperforms the state-of-the-art baselines by significantly improving the estimation accuracy.

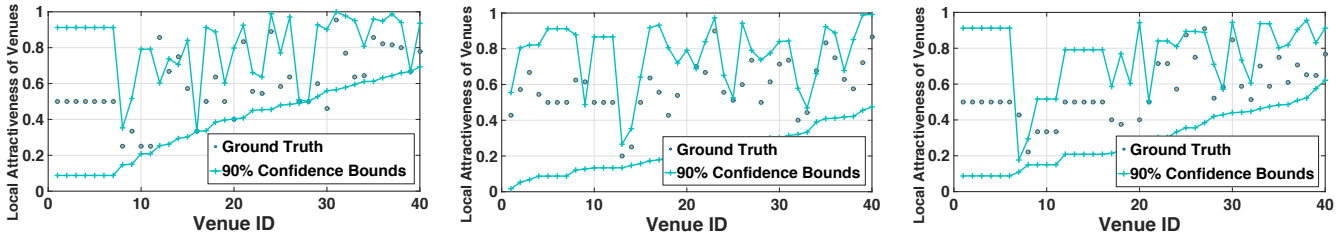
Acknowledgment

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

References

[1] 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 *KDD*, pages 114–122. ACM, 2011.



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

Figure 5: Confidence Bounds of STSA at 90% Confidence Level

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

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

[5] X. Chen, Y. Zeng, G. Cong, S. Qin, Y. Xiang, and Y. Dai. On information coverage for location category based point-of-interest recommendation. In *AAAI*. AAAI Press, 2015.

[6] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM*, pages 759–768. ACM, 2010.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Rpyal Statistical Society*, pages 1–38, 1977.

[8] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *KDD*, pages 15–24. ACM, 2014.

[9] H. Gao, J. Tang, X. Hu, and H. Liu. Content-aware point of interest recommendation on location-based social networks. In *AAAI*, pages 1721–1727, 2015.

[10] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, pages 779–782, 2008.

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

[12] T.-r. Hu, J.-b. Luo, H. Kautz, and A. Sadilek. Home location inference from sparse and noisy data: models and applications. In *ICDM*, pages 1382–1387. IEEE, 2015.

[13] C. Huang and D. Wang. Topic-aware social sensing with arbitrary source dependency graphs. In *IPSN*, pages 1–12. IEEE/ACM, 2016.

[14] C. Huang and D. Wang. Towards time-sensitive truth discovery in social sensing applications. In *MASS*, pages 154–161. IEEE, 2016.

[15] C. Huang and D. Wang. Unsupervised interesting places discovery in location-based social sensing. In *12th IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS 16)*. IEEE, 2016.

[16] M. Kaminskas, F. Ricci, and M. Schedl. Location-aware music recommendation using auto-tagging and hybrid matching. In *Recsys*, pages 17–24. ACM, 2013.

[17] R. Li, C. Wang, and K. C.-C. Chang. User profiling in an ego network: co-profiling attributes and relationships. In *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. In *PVLDB*, pages 1603–1614. ACM, 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 *KDD*, pages 1023–1031. ACM, 2012.

[20] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao. Privacy vulnerability of published anonymous mobility traces. *TON*, pages 720–733, 2013.

[21] A. Q. Macedo, L. B. Marinho, and R. L. Santos. Context-aware event recommendation in event-based social networks. In *Recsys*, pages 123–130. ACM, 2015.

[22] J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *TIST*, page 47, 2014.

[23] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge University Press, 2008.

[24] J. Marshall, M. Syed, and D. Wang. Hardness-aware truth discovery in social sensing applications. In *12th IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS 16)*. IEEE, 2016.

[25] J. Marshall and D. Wang. Mood-sensitive truth discovery for reliable recommendation systems in social sensing. In *10th ACM Conference on Recommender Systems (Recsys 2016)*. ACM, 2016.

[26] J. Marshall and D. Wang. Towards emotional-aware truth discovery in social sensing applications. In *The 2nd IEEE International Conference on Smart Computing (SMARTCOMP 2016)*. IEEE, 2016.

[27] J. McGee, J. Caverlee, and Z. Cheng. Location prediction in social media based on tie strength. In *CIKM*, pages 459–468. ACM, 2013.

[28] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *WSDM*, pages 251–260. ACM, 2010.

[29] V. Papathanasiou. Some characteristic properties of the fisher information matrix via cacoullos-type inequalities. *Journal of Multivariate analysis*, pages 256–265, 1993.

[30] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *COLING*. IEEE, 2010.

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

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

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

[34] 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.

[35] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, et al. Using humans as sensors: an estimation-theoretic perspective. In *Proceedings of the 13th international symposium on Information processing in sensor networks*, pages 35–46. IEEE Press, 2014.

[36] D. Wang and C. Huang. Confidence-aware truth estimation in social sensing applications. In *SECON*, pages 336–344. IEEE, 2015.

[37] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal. On credibility estimation tradeoffs in assured social sensing. *JSAC*, pages 1026–1037, 2013.

[38] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In *IPSN*, April 2012.

[39] D. Wang, J. Marshall, and C. Huang. Theme-relevant truth discovery on twitter: An estimation theoretical approach. In *The 10th International AAAI Conference on Web and Social Media (ICWSM 16)*. IEEE, 2016.

[40] 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 *CSCW*, pages 302–316. ACM, 2015.

[41] H. Yin, Y. Sun, B. Cui, Z. Hu, and L. Chen. Lcars: a location-content-aware recommender system. In *KDD*, pages 221–229. ACM, 2013.

[42] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *TKDE*, pages 796–808, 2008.

[43] J.-D. Zhang and C.-Y. Chow. Geosoca: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *SIGIR*, pages 443–452. ACM, 2015.