A

An Unsupervised Approach to Inferring the Localness of People Using Incomplete Geo-Temporal Online Check-in Data

Chao Huang, University of Notre Dame
Dong Wang, University of Notre Dame
Jun Tao, University of Notre Dame

Inferring the localness of people is to classify people who are local residents in a city from people who visit the city by analyzing online check-in points that are contributed by online users. This information is critical for the urban planning, user profiling and localized recommendation systems. Supervised learning approaches have been developed to infer the location of people in a city by assuming the availability of high quality training datasets with complete geo-temporal information. In this article we develop an *unsupervised* model to accurately identify local people in a city by using the incomplete online check-in data that are publicly available. In particular, we develop an Incomplete-Geo-Temporal Expectation Maximization (IGT-EM) scheme, which incorporates a set of hidden variables to represent the localness of people and a set of estimation parameters to represent the likelihood of venues to attract local and non-local people respectively. Our solution can accurately classify local people from non-local nones without requiring any training data. We also implement a parallel IGT-EM algorithm by leveraging the computing power of a Graphic Processing Unit (GPU) that consists of 2496 cores. In the evaluation, we compared our new approach with the existing solutions through four real-world case studies using data from the city of New York, Chicago, Boston and Washington D.C. The results showed that our approach can identify the local people and significantly outperform the compared baselines in estimation accuracy and execution time.

Additional Key Words and Phrases: Localness of People, Unsupervised Learning, Online Social Networks, Crowdsourcing, Maximum Likelihood Estimation, GPU Implementation

1. INTRODUCTION

Two recent technical trends have changed people's daily lives fundamentally: the advent of online social media (Twitter, Facebook, Foursquare) and the proliferation of location based sensors (e.g., GPS sensors). The combination of them has resulted in Location-Based Social Network (LBSN) services. In those services, people check in at the venues they visited and share their check-in data through online social network services. Examples of such services include Foursquare, BrightKite, Citysense, GyPSii, MobiLuck. Millions of users have already adopted LBSN services and their check-in points become an important open data source for a city to develop *crowdsourcing-based smart city applications* (e.g., intelligent transportation, urban infrastructure monitoring, geotagging, crowdsensing, etc.) [Wang and Huang 2015; Cardone et al. 2013; Wang et al. 2014b].

Along with this trend, an important problem is to infer the *localness of people*, where the goal is to classify people who are local residents in a city from people who visit the city by analyzing online check-in points that are contributed by online users. Solving this problem will allow for (i) inferring different activity/mobility patterns of local and non-local people in a city; (ii) identifying the venues that are more likely to attract local or non-local people respectively. Such information is critical for many location based applications such as ads targeted for local business [Provost et al. 2009; Ahmed et al. 2011], urban planning [Ratti et al. 2006; Gonzalez et al. 2008], and localized news recommendations [White et al. 2009; Bennett et al. 2012].

A city may have a record of its local residents through census. However, online check-in data are distinct from census data in several ways. First, the census data is typically not publicly available or just includes some aggregated statistics, and thus cannot be used for the above-mentioned location based applications. Second, the census data does not include the venues that individuals visit in their daily lives and hence is not helpful in understanding the activity/mobility patterns or identifying hot spots visited

by local/non-local people in a city. Third, many people do not put their real names and home locations in their profiles on LBSN services, making it difficult (if possible) to map their LBSN accounts to the census information.

Supervised learning approaches have been developed to infer the location of people in a city by assuming the availability of high quality training datasets with complete geo-temporal information [Li et al. 2012b; Backstrom et al. 2010; Cheng et al. 2010; Cheng et al. 2011; Li et al. 2012a]. However, such assumption does not always hold in practice for a couple of reasons. First, many people do not feel comfortable to upload their localness or home location information to the online services for privacy concerns [Ma et al. 2013]. Second, many LBSNs do not share all their collected data with the general public [Zhang and Chow 2015]. Therefore, only a portion of the user's check-in points are available for analysis and we refer to such data as *geo-temporal incomplete data*. In particular, we define "incomplete geo-temporal check-in point data" in this article to be the check-in traces users share on LBSN.

A few technical challenges exist to solve the problem of localness inference. *First*, the online check-in trace is incomplete and sparse: we cannot expect a person to check in at every place in the city (e.g., he/she either forgets to check in or intentionally chooses not to do so due to privacy concerns). *Second*, the collected data is "noisy": both local people and non-local visitors can check in at the same venue. There is no clear difference between the check-in points provided by different group of people.

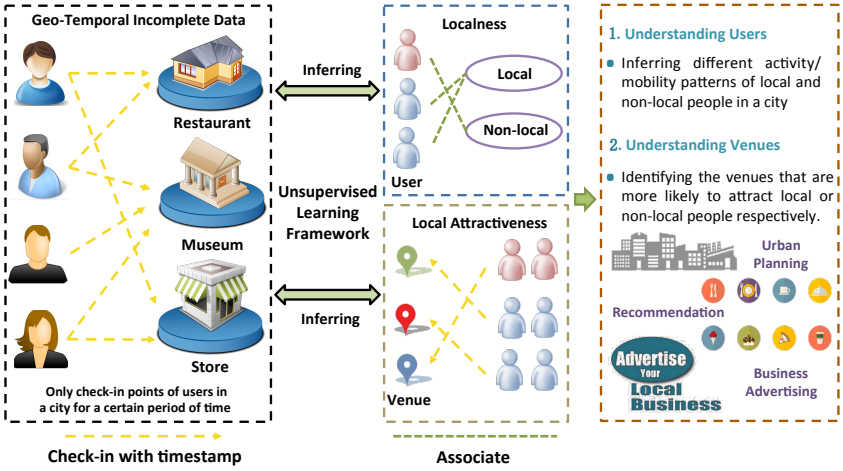


Fig. 1. The Overview of Incomplete-Geo-Temporal Expectation Maximization (IGT-EM) Framework

In this article, we develop an *unsupervised* learning approach to find local people in a city (Figure 1). In particular, we develop an Incomplete-Geo-Temporal Expectation Maximization (IGT-EM) scheme, which incorporates a set of hidden variables to represent the localness of people and a set of estimation parameters to represent the likelihood of venues to attract local and non-local people respectively. Our solution can identify the localness of people without prior knowledge on the people and the venues they visited. Additionally, we implement a parallel IGT-EM algorithm by leveraging the computing power of a Graphic Processing Unit (GPU) that consists 2496 cores. In the evaluation, we compared our new approach with the existing solutions through four real-world case studies using data from the city of New York, Chicago, Boston and Washington D.C. The results showed that our approach can identify the local people and significantly outperform the compared baselines in estimation accuracy and execution time.

We summarize the main contributions of this article as follows:

— In this article, we address the problem of inferring the localness of people using an *unsupervised* approach based on *incomplete geo-temporal* data.
— We develop a new approach to identify the local people without the prior knowledge on the people and the venues they visited. (Section 4)
— We implement a parallel IGT-EM scheme that can run on GPU processors and demonstrate that the parallel algorithm runs *a few orders of magnitude* faster than its sequential counterpart. (Section 5)
— We evaluate the estimation accuracy and execution time of IGT-EM scheme through four real world case studies using data from Foursquare. The IGT-EM scheme is shown to significantly outperform the compared baselines in identifying the localness of people. (Section 6)

A preliminary version of this work has been published in [Huang and Wang 2016a]. This journal article significantly expands our previous work and makes new contributions from the following aspects. First, we developed a scalable framework of IGT-EM to implement our proposed scheme on a parallel platform (i.e., GPU), which can efficiently handle big data and is more suitable for large-scale online social network applications (Section 5). Second, we performed a set of experiments on a new dataset collected from the largest city in U.S. (i.e., New York) and further evaluated the robustness and efficiency of our scheme in more real world scenarios (Section 6). Third, we studied the performance of the IGT-EM scheme with respect to a set of key parameters in our model: the threshold for time length of check-in trace, number of check-in points, and the density of the dataset. The new results verified the robustness of our model over different dimensions of the problem (Section 6). Finally, we also investigated the convergence of the IGT-EM scheme and presented the results of the convergence analysis of IGT-EM over different datasets (Section 6).

2. RELATED WORK

With the proliferation of the online social media and networks, some seminal work explores social network for user profiling [Lampe et al. 2007; Pfeil et al. 2009; Golbeck 2009]. For example, Lampe et al. explored the relationship between profile structure and the number of friends to decide which profile elements are most likely to predict the friendship links on Facebook [Lampe et al. 2007]. Pfei et al. investigated the age differences and similarities in the use of the MySapce to explore potential differences in social capital among older people and teenagers [Pfeil et al. 2009]. Golbeck investigated features of profile similarity and their relationship to the user's trust by isolating profile features beyond the overall similarity and examined their solution on FilmTrust [Golbeck 2009]. However, none of the above works studied the problem of inferring the localness of people in a city. In contrast, this paper focuses on the problem of identifying local people by using the incomplete geo-temporal online social media data from a city.

Prior works exist on the topic of mining the location of people in a city from their online check-in data [Li et al. 2012b; Backstrom et al. 2010; Cheng et al. 2010; Cheng et al. 2011; Li et al. 2012a; Jurgens 2013]. For example, Cheng et al. developed a statistical model to estimate the location of Twitter users by analyzing the key phrases used in tweets [Cheng et al. 2010; Cheng et al. 2011]. They used a set of location related phrases to train their model and associate those phrases with locations. Backstrom et al. estimated the location of a user by exploring the relationship between geography and friendship on online social networks [Backstrom et al. 2010]. Jurgens et al developed a framework to infer locations of users by studying their spatial correlations and using a small number of ground truth locations as the initiation of its algorithm [Jur-

gens 2013]. Li et al developed a novel analytical model to predict user's home location from leveraging the information from both social network and user generated data [Li et al. 2012b]. They further extended their model to handle cases where users have multiple home locations [Li et al. 2012a]. However, the above solutions all need sufficient training data to build up their models and generate accurate estimation results, hence are all supervised learning approaches. On the contrary, this paper takes an unsupervised learning approach to infer the localness of people without prior knowledge on the people and the venues they visited.

A category of location-based recommendation systems also bear a resemblance to our work [Yin et al. 2013; Chen et al. 2015; Park et al. 2007; Ramaswamy et al. 2009; Kodama et al. 2009]. For example, Yin et al. built a recommendation system that offers a particular user a set of venues or events by considering the user's personal interest and local preference [Yin et al. 2013]. Furthermore, Chen et al. developed a Point-of-Interests (POI) recommendation system to find the top-K location category based POI recommendation by considering the information coverage in the recommendation process [Chen et al. 2015]. Part et al. developed recommendation system for venues by exploring the visitor's profile information (e.g., location, age or cuisine preferences) [Park et al. 2007]. Ramaswamy et al. proposed a location recommendation system that leverages user's social affinity [Ramaswamy et al. 2009]. Kodama et al. designed an approach to recommend items to a user by taking into account his current location and preferences [Kodama et al. 2009]. In contrast, we focus on solving the problem of inferring the localness of people using LBSN data, which could be used in many recommendation systems like the ones discussed above.

3. MODEL

In this section, we describe the model we used to infer the localness of people in a city using incomplete geo-temporal data. Let us consider a LBSN application where a set of users U_1, U_2,..,U_Y generate check-in points in a set of venues V_1, V_2,...,V_X, in a city. Let V_i represent the i^{th} venue and U_j represent the j^{th} user. We introduce a hidden variable Z for each user to indicate the localness of that user. For example, $Z_j = 1$ when the user U_j is local and $Z_j = 0$ if he/she is not. We also define a ***Check-in Matrix*** VU: $VU_{i,j} = 1$ indicates that the user U_j checks in at venue V_i and $VU_{i,j} = 0$ otherwise.

We define the time length of a user's check-in trace as the time difference between the earliest and latest check-in points contributed by the user in the dataset. It varies from one person to another. We further define a ***Time Vector*** T and the element t_j is the time length of the check-in trace from user U_j. In this paper, we explicitly consider *both venue and time information* and formulate an optimization problem to estimate the localness of people.

We then define a_i as the ***local attractiveness*** of a venue V_i. Formally a_i is given by:

$$a_i = P(U_j = 1 | VU_{i,j} = 1) \tag{1}$$

We define a_i^k as the local attractiveness of V_i to attract local people whose time length of check-in trace lasts for k days. Formally, a_i^k is given by:

$$a_i^k = P(U_j = 1 | VU_{i,j} = 1, t_j = k) \tag{2}$$

Therefore,

$$a_i = \sum_{k=1}^{K} a_i^k \times \frac{r_i^k}{r_i} \quad k = 1, ..., K \tag{3}$$

where $r_i^k = P(VU_{i,j} = 1, t_j = k)$. For each venue, r_i^k can be computed by counting and dividing the frequencies of users check-in points at V_i using the Check-in Matrix VU and the Time Vector T. Note that $r_i = \Sigma_{k=1}^{K} r_i^k$.

We further define $b_{i,k}$ and c_i^k as follows:

$$b_i^k = P(VU_{i,j} = 1, t_j = k | U_j = 1)$$
$$c_i^k = P(VU_{i,j} = 1, t_j = k | U_j = 0) \tag{4}$$

Follow the Bayes theorem, we can obtain the relation between b_i^k, c_i^k and a_i^k, r_i^k as follows:

$$b_i^k = \frac{a_i^k \times r_i^k}{l} \quad c_i^k = \frac{(1 - a_i^k) \times r_i^k}{1 - l} \tag{5}$$

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

Therefore, we can formulate the problem of identify local people using incomplete geo-temporal data studied as a constraint maximum likelihood estimation (MLE) problem: given only the Check-in Matrx VU and Time Vector T, our goal is to compute:

$$\forall j, 1 \leq j \leq Y : P(U_j = 1 | VU, T)$$
$$\forall i, 1 \leq i \leq X : P(U_j = 1 | VU_{i,j} = 1) \tag{6}$$

4. SOLUTION

In this section, we solve the problem of local people identification formulated in Section 3 by developing an Incomplete-Geo-Temporal Expectation Maximization (IGT-EM) algorithm.

4.1. Background

Expectation Maximization (EM) is a commonly used optimization technique for the MLE problem where the model contains hidden variables that cannot be directly observed from the data. Specifically, it contains two steps:

$$\text{E-step: } Q(\theta | \theta^{(n)}) = E_{Z|x,\theta^{(n)}}[logL(\theta; x, Z)] \tag{7}$$
$$\text{M-step: } \theta^{(n+1)} = \arg\max_{\theta} Q(\theta | \theta^{(n)}) \tag{8}$$

We solve the localness inference problem by developing an Incomplete-Geo-Temporal EM scheme. Specifically, the observed data X in our problem is the Check-in Matrix VU and the Time Period Vector T. The estimation parameter vector is defined as $\theta = (b_1^k, b_2^k, ..., b_X^k; c_1^k, c_2^k, ..., c_X^k; l)$, where $k = 1, 2, .., K$. b_i^k and c_i^k are defined in Equation (4). We further introduce a vector of hidden variables Z to indicate whether a user is local or not. In particular, the variable z_j for user U_j such that $z_j = 1$ if U_j is local and $z_j = 0$ otherwise. Finally, we define a set of indication variables t_j^k such that $t_j^k = 1$ if $t_j = k$ in Time Period Vector T and $t_j^k = 0$ otherwise.

The likelihood function for IGT-EM is as follows:

$$
\begin{aligned}
&L(\theta; X, Z) = \Pr(X, Z|\theta) \\
&= \prod_{j=1}^{Y} \Bigg\{ \prod_{i=1}^{X} \prod_{k=1}^{K} (\beta_{i,j}^{k})^{V_i U_j \times t_j^k} \times (\beta_{i,j}^{k})^{(1-V_i U_j)} \times l \times z_j \\
&+ \prod_{i=1}^{X} \prod_{k=1}^{K} (\beta_{i,j}^{k})^{VU_{i,j} \times t_j^k} \times (\beta_{i,j}^{k})^{(1-VU_{i,j})} \times (1-l) \times (1-z_j) \Bigg\}
\end{aligned}
\tag{9}
$$

where $VU_{i,j} = 1$ when user U_j visits venue V_i and 0 otherwise. The β_i^k are defined as follows:

$$
\beta_{i,j}^{k} = \begin{cases}
b_i^k & if\ VU_{i,j} = 1, t_j^k = 1, z_j = 1 \\
(1 - \sum_{k=1}^{K} b_i^k) & if\ VU_{i,j} = 0, t_j^k = 1, z_j = 1 \\
c_i^k & if\ VU_{i,j} = 1, t_j^k = 1, z_j = 0 \\
(1 - \sum_{k=1}^{K} c_i^k) & if\ VU_{i,j} = 0, t_j^k = 1, z_j = 0
\end{cases}
\tag{10}
$$

4.2. Incomplete-Geo-Temporal Expectation Maximization

We can derive the E-step as follows:

$$
\begin{aligned}
&Q(\theta|\theta^{(n)}) = E_{Z|X,\theta^{(n)}}[logL(\theta; X, Z)] \\
&= \sum_{j=1}^{Y} \Bigg\{ \Pr(z_j = 1|X_j, \theta^{(n)}) \times \Bigg[\sum_{i=1}^{X} \sum_{k=1}^{K} (VU_{i,j} \times t_j^k) \times log\beta_{i,j}^k + (1 - VU_{i,j}) \times log\beta_{i,j}^k + logl \Bigg] \\
&+ \Pr(z_j = 0|X_j, \theta^{(n)}) \times \Bigg[\sum_{i=1}^{X} \sum_{k=1}^{K} (VU_{i,j} \times t_j^k) \times log\beta_{i,j}^k + (1 - VU_{i,j}) \times log\beta_{i,j}^k + log(1-l) \Bigg] \Bigg\}
\end{aligned}
\tag{11}
$$

Note that in the Q function, the estimation parameters are represented by $\beta_{i,j}^k$ which is defined in Equation (10). $\beta_{i,j}^k$ represents different parameters under different conditions.

To drive the M-step, we set partial derivatives of $Q(\theta|\theta^{(n)})$ given by Equation (11) with respect to θ to 0 to get the optimal θ^*. Specifically, we get the solutions of $\frac{\partial Q}{\partial b_i^k} = 0$, $\frac{\partial Q}{\partial c_i^k} = 0$ and $\frac{\partial Q}{\partial l} = 0$ in each iteration. The optimal estimation of the parameters for the next iteration (i.e., $(b_i^k)^{(n+1)}$, $(c_i^k)^{(n+1)}$ and $(l)^{(n+1)}$) are as follows:

$$(b_i^k)^{(n+1)} = \frac{\Sigma_{j \in SW_i^k} \mathbf{Pr}(z_j = 1 | X_j, \theta^{(n)})}{\Sigma_{j \in U} \mathbf{Pr}(z_j = 1 | X_j, \theta^{(n)})}$$
$$(c_i^k)^{(n+1)} = \frac{\Sigma_{j \in SW_i^k} (1 - \mathbf{Pr}(z_j = 1 | X_j, \theta^{(n)}))}{\Sigma_{j \in U} (1 - \mathbf{Pr}(z_j = 1 | X_j, \theta^{(n)}))}$$
$$(l)^{(n+1)} = \frac{\Sigma_{j \in U} \mathbf{Pr}(z_j = 1 | X_j, \theta^{(n)})}{|U|} \quad (12)$$

where SW_i^k is the group of users who check in at venue U_i and the time length of their check-in points is k days in a city. U is the set of all users. The E and M steps of the IGT-EM scheme are shown in Figure 2.

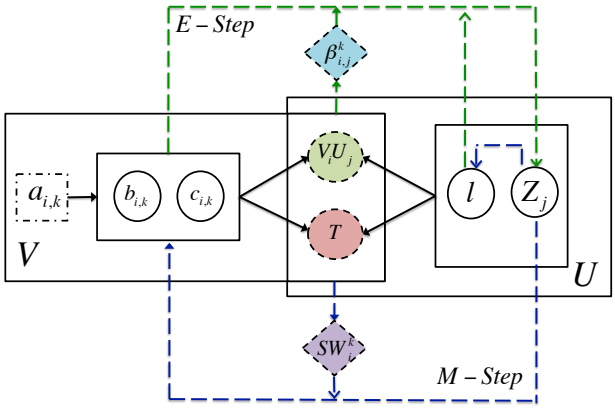


Fig. 2. IGT-EM Scheme

5. PARALLEL IGT-EM

To further improve the computing performance, we design a parallel implementation of the IGT-EM scheme on a Graphic Processing Unit (GPU) that uses Compute Unified Device Architecture (CUDA) programming model [Nvidia 2008]. GPU has emerged as a new computing platform for many computational intensive applications. CUDA is a parallel programming model invented by NVIDIA. In CUDA, a *kernel* is defined as a grid of thread blocks and a thread of execution is the smallest unit in the parallelization. In the parallelization process, each node (called a *thread node*) will take care of a part of the whole computation task and users need to specify a set of *kernels* to parallelize the computation task.

In this work, we implement a parallel version of the IGT-EM scheme on a Graphic Processing Unit (GPU) platform to improve its computation efficiency for large-scale datasets. While there exist prior studies on parallel implementation of EM algorithm, several challenges exist in order to implement parallel IGT-EM: (i) the memory of Graphics Card is limited, so we need to design efficient strategies to handle the large-scale datasets from LBSN on GPU; (ii) we need to design a mechanism to distribute the computation task of various estimation parameters and hidden variables of IGT-EM to different threads in an efficient way. To address these challenges, we designed the parallel IGT-EM based on the MLE model developed in this paper and optimized our implementation using the following techniques: (i) we set the variables used in each thread as local variables instead of global variables given the fact that it costs more

time to access global memory than local memory; (ii) we replaced the original conditional branch in the IGT-EM algorithm with directly index in corresponding arrays, which allows us to save threads waiting time during the branch execution. The above optimization leads to significant execution time improvement achieved by IGT-EM as shown in the next section. We summarize the parallel IGT-EM scheme in Algorithm 1.

ALGORITHM 1: Parallel Incomplete-Geo-Temporal EM Algorithm

Input: Check-in Matrix VU, Time Period Vector T
Output: Localness of Users and Local Attractiveness of Venues

Initialize θ ($b_i^k = r_i^k, c_i^k = 0.5 \times r_i^k, l$ =Random number in $(0, 1)$)
$n = 0$
repeat
  $n = n + 1$
  CUDA Kernel of E-Step:
  for Each $j \in U$ **do**
    computation of $j \rightarrow$ one thread
    compute $\Pr(z_j = 1 | X_j, \theta^{(n)})$
  end for
  CUDA Kernel of M-Step:
  for Each $i \in V$ **do**
    computation of $i \rightarrow$ one thread
    compute $(b_i^k)^{(n)}, (c_i^k)^{(n)}, (l)^{(n)}$
  end for
until $\theta^{(n)}$ and $\theta^{(n-1)}$ converge

6. EVALUATION

In this section, we conduct experiments to study the performance of the *IGT-EM* in comparison with a set of state-of-the-art baselines through four real world case studies using data collected from Foursquare. We show that the *IGT-EM* scheme can classify local people from non-local people more accurately and efficiently than the compared baselines.

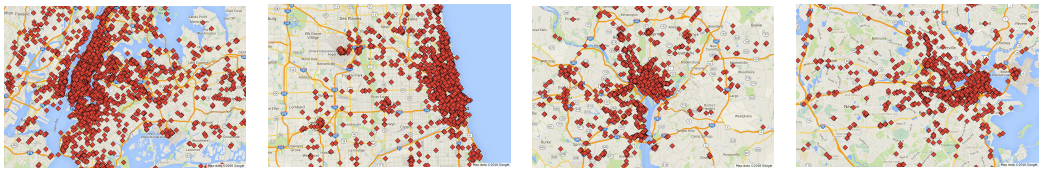
6.1. Experimental Setups

6.1.1. Dataset Statistics. In this paper, we study the performance of our proposed scheme using data traces collected from Foursquare, a widely used LBSN. In the evaluation, we selected four data traces from cities in U.S with user home location information available [1]: New York, Chicago, Washington D.C and Boston. Figure 3 shows the venue maps of the four cities [2]. Figure 4 and Figure 5 further show the distributions of check-in points per venue (density) and per users (frequency) in the four datasets respectively. In these figures, we can observe that the check-in points of the datasets are "incomplete" in the sense that check-in points at many venues are sparse and many users check in at venues with a low frequency.

Table I. Data Traces Statistics

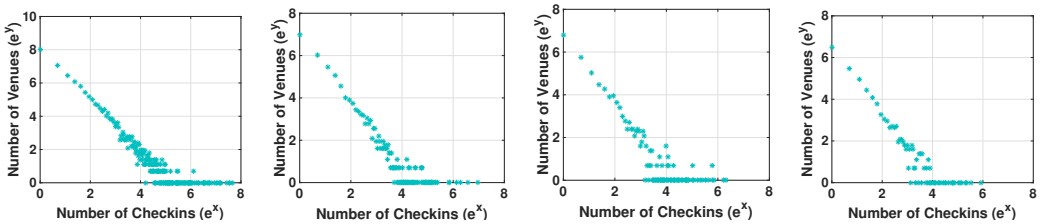
Data Trace	New York	Chicago	Washington D.C	Boston
Number of Users	75,680	31,965	17,231	12,946
Number of Venues	7,663	2,529	1,932	1,478
Number of Check-ins	144,179	48,605	25,722	18,296

[1] https://archive.org/details/201309_foursquare_dataset_umn
[2] Many venues are so close together that they overlap with each other in the map



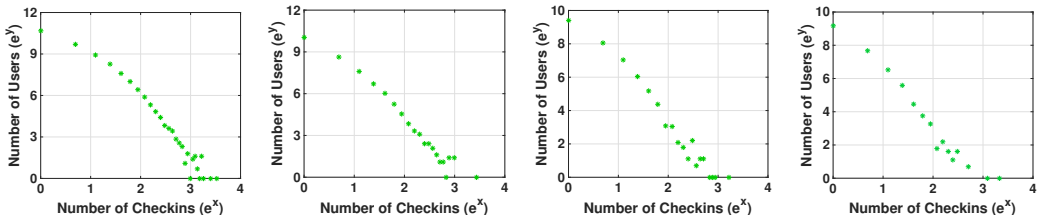
(a) New York (b) Chicago (c) Washington D.C (d) Boston

Fig. 3. Maps of Venues in Four Cities



(a) New York (b) Chicago (c) Washington D.C (d) Boston

Fig. 4. Distribution of Check-in Points Per Venue (Density)



(a) New York (b) Chicago (c) Washington D.C (d) Boston

Fig. 5. Distribution of Check-in Points Per User (Frequency)

6.1.2. Data Pre-Processing. We designed a few data pro-processing steps to generate the inputs to the IGT-EM scheme.

Generating Check-in Matrix: We generate the VU Matrix by examining which user check in at which venue. For example, if user U_j made a check-in point at venue V_i, we set the element V_iU_j in VU to 1 and 0 otherwise.

Generating Time Vector: For simplicity, we divide all users into two sets based on the time length of their check-in trace. For example, if the time length of user's check-in points (i.e., the time difference between the earliest and latest check-in points) is larger than a predefined threshold (we used 10 days in our experiment), we set the corresponding element t_j in vector T as 1. Otherwise, we set the t_j as 0.

The above pre-processing steps generated all the inputs (i.e., VU Matrix and T vector) that are needed for the *IGT-EM* scheme.

6.2. Evaluation of Our Scheme

6.2.1. Baselines. We compared the *IGT-EM* scheme with state-of-the-art techniques that includes:

— *Regular-EM*: it uses a basic EM approach (that does not consider the time length of check-in trace) to solve the problem of inferring the loalness of people [Wang et al. 2012]. Particularly, the input to the Regular-EM algorithm is only Venue-User Matrix VU and it does not consider Time Vector T in the estimation model.

— *MLP*: it is a supervised learning scheme that leverages the location information of a user's online friends to infer the locations of the target user [Li et al. 2012a]. Partic-

ularly, it estimates a users location by taking average of home coordinates of people who have an online social connection with the user.

— *FM*: it is another supervised learning approach that solves the problem by leveraging the home locations of the people who have similar venue visit pattern as the target users [Backstrom et al. 2010].

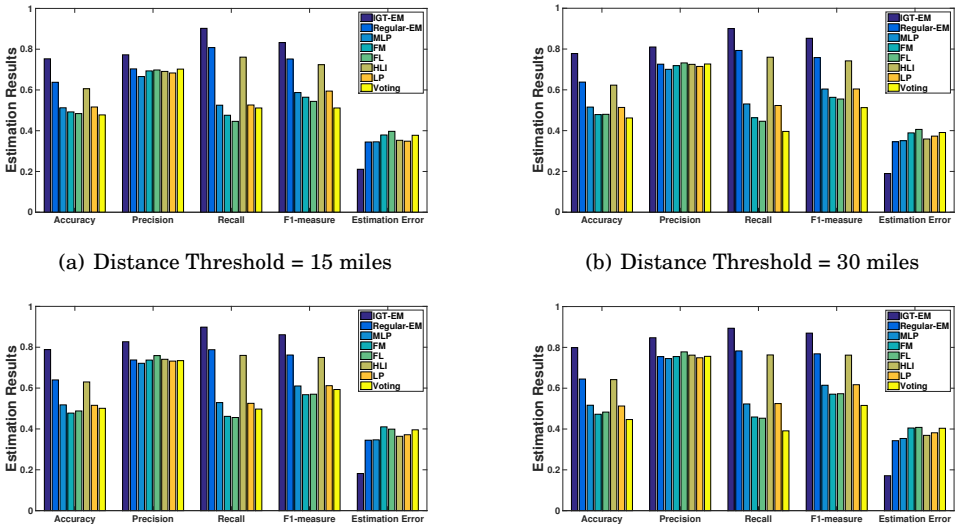
— *FL*: it arguments the MLP approach by explicitly considering the social tie strength between users [McGee et al. 2013].

— *HLI*: it develops a machine learning approach that estimates a user' location by assuming that users who have check-in points in the evening of the city are more likely to be local users [Hu et al. 2015]..

— *LP*: it presents an algorithm that leverages the geographic distribution of an individuals ego network (i.e., social network in one-hop) to infer their locations [Jurgens 2013]. Particularly, it estimates the location of a user by taking average of home coordinates of his/her online social friends in the ego network.

— *Voting*: it decides the localness of a user by assuming local people visit more venues than non-local people.

6.2.2. Results. In the evaluation, we use the home location of users as the ground truth information to decide whether a user is local or not. For example, if the user's home location is X miles away from the center of the city, the user is considered to be *local*. We also changed the value of X in our experiments to study the performance of our scheme and other baselines. One should note that we only used the ground truth information for the evaluation purpose and did not use it as the input to our *IGT-EM* scheme.



(a) Distance Threshold = 15 miles

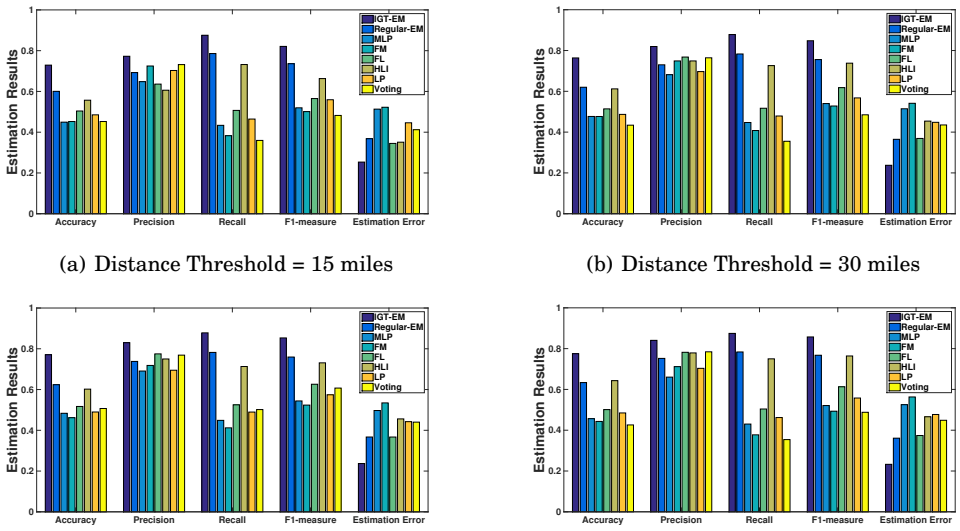
(b) Distance Threshold = 30 miles

(c) Distance Threshold = 50 miles

(d) Distance Threshold = 100 miles

Fig. 6. Estimation Results on New York Data Trace

Figure 6 shows the results of the *IGT-EM* and other baselines in New York City dataset. *IGT-EM* is observed to clearly outperform the compared baselines: it classifies local people from non-local people with smallest false positives and false negatives. It



(a) Distance Threshold = 15 miles

(b) Distance Threshold = 30 miles

(c) Distance Threshold = 50 miles

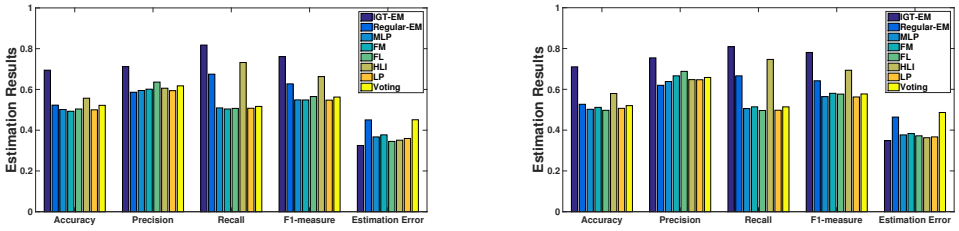
(d) Distance Threshold = 100 miles

Fig. 7. Estimation Results on Chicago Data Trace

also has a more accurate estimation on how attractive a venue is for local people compared to other schemes. For example, the *IGT-EM* outperforms the best baseline by 12% and 8% on accuracy and F1-measure respectively. We also observe that the performance gain of *IGT-EM* is consistent when we changed the values of X. Figure 7 shows the results of Chicago data trace. Similar results are observed: *IGT-EM* continues to outperform all baselines in both inferring the localness of people and the estimation error on the local attractiveness of venues. The results of Washington D.C and Boston data trace are shown in Figure 8 and Figure 9 respectively. In these two figures, similar evaluation results can be observed: *IGT-EM* is the best performed scheme over all evaluation metrics. The above results show that the *IGT-EM* can effectively identify the localness of people in a city and achieved significant performance gains compared to state-of-the-art techniques.

To investigate the effect of check-in time length (the time difference between the earliest and latest check-in points of a user) threshold that we used to generate the time vector T, we studied the performance of our proposed *IGT-EM* scheme by varying the threshold of check-in time length (i.e., 6 days, 8 days, 10 days, 12 days, 14 days and 16 days). Particularly, *IGT-EM-t* represents the corresponding IGT-EM scheme with t days as the check-in time length threshold value. The evaluation results on all data trace are presented in Figure 10. We observe that the performance of the IGT-EM scheme is robust to the values of t in the evaluated range. The reason is that the ratio of non-local to local users in the evaluated threshold range (i.e., 6-16 days) is relatively stable (i.e., 2-1.8), which leads to the insensitiveness of the IGT-EM scheme to the check-in length threshold.

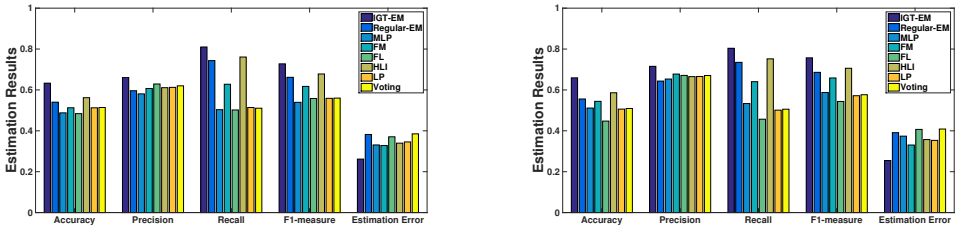
We further investigate the scalability of the IGT-EM scheme in several different dimensions. We first compared the performance of IGT-EM scheme on four cities with different number of check-in points (i.e., New York City, Chicago, Washington D.C and Boston). The results are reported in Figure 11. We can observe that the performance of IGT-EM improves as the number of check-in points increases. The reason is intuitive: more check-in points provide more evidence for the IGT-EM to accurately infer the localness of the users. Furthermore, we also investigate the effect of check-in point den-



(a) Distance Threshold = 15 miles (b) Distance Threshold = 30 miles

(c) Distance Threshold = 50 miles (d) Distance Threshold = 100 miles

Fig. 8. Estimation Results on Washington D.C. Data Trace

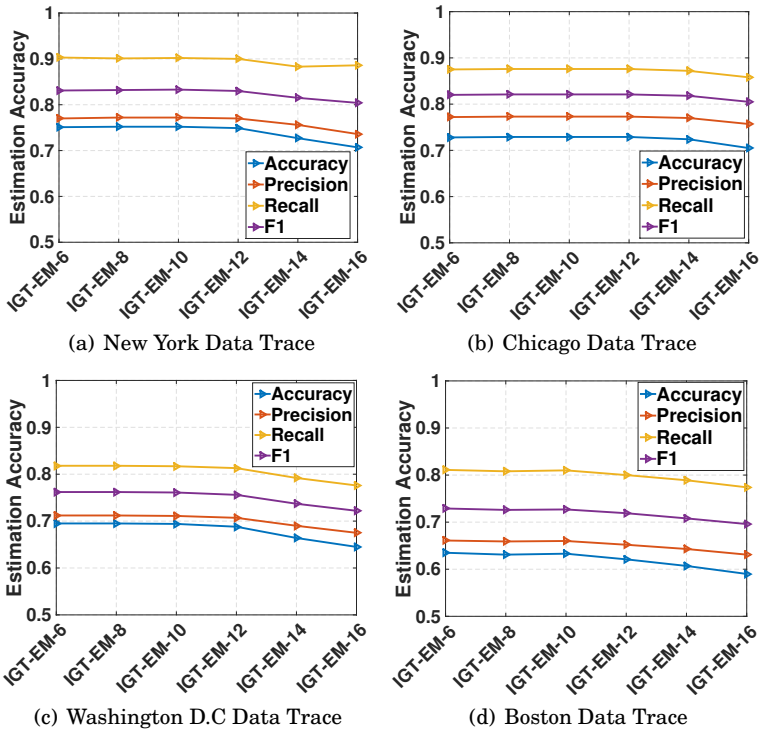


(a) Distance Threshold = 15 miles (b) Distance Threshold = 30 miles

(c) Distance Threshold = 50 miles (d) Distance Threshold = 100 miles

Fig. 9. Estimation Results on Boston Data Trace

sity (number of check-in points per venue) and frequency (number of check-in points per user) on the performance of the IGT-EM scheme. The results are reported in Figure 12 and Figure 13 respectively. In particular, we choose the city with the largest number of check-in points (New York) and the city with the smallest number of check-in points (Boston) in our datasets as two typical examples in the reported results. We observe that the performance of IGT-EM in general improves as the number of check-in points per venue and per user increases respectively.



(a) New York Data Trace (b) Chicago Data Trace

(c) Washington D.C Data Trace (d) Boston Data Trace

Fig. 10. Performance Evaluation of IGT-EM-t

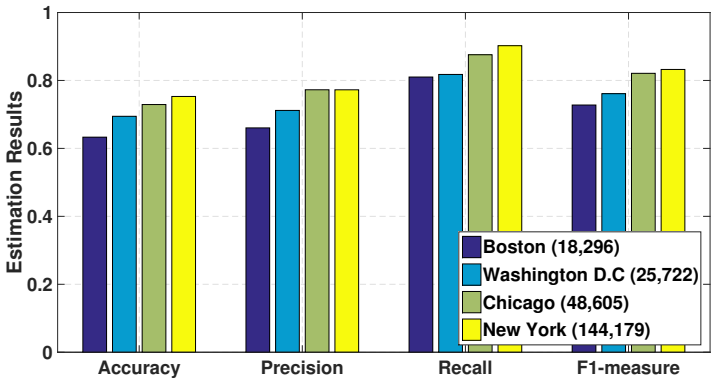
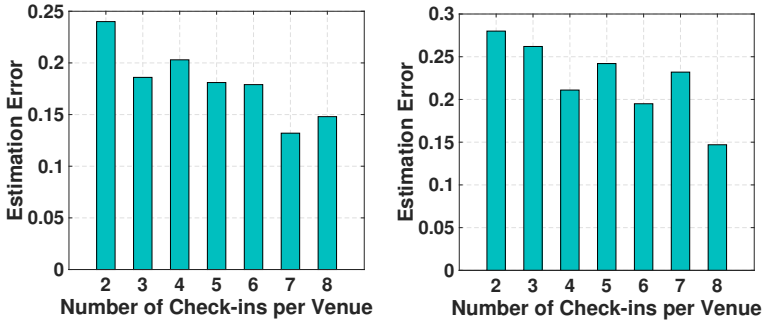


Fig. 11. Performance of IGT-EM on Four Cities with Different Number of Check-ins

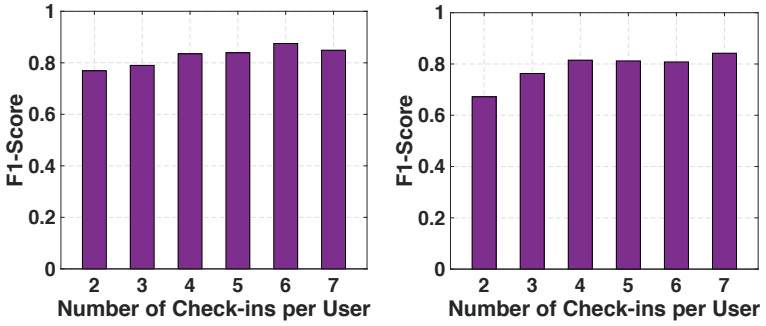
We also investigated the convergence of the IGT-EM scheme on the four data traces and the results are presented in Figure 14. The figure showed the negative likelihood function with respect to the number of iterations of IGT-EM scheme. We observe the IGT-EM scheme converges quickly on all data traces.

Finally, we evaluate the efficiency of the parallel IGT-EM implementation discussed in Section 5. We implement the parallel IGT-EM on a computer with Nvidia GeFore GPU (2496 cores and 1.25 GHZ for each core, 4GB memory). We compare the parallel IGT-EM with all baselines we discussed earlier. We run the Sequential IGI-EM (i.e.,



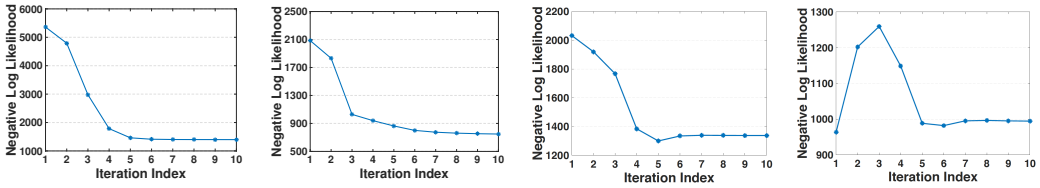
(a) New York Data Trace (b) Boston Data Trace

Fig. 12. Performance of IGT-EM by Varying the Check-in Points Per Venue



(a) New York Data Trace (b) Boston Data Trace

Fig. 13. Performance of IGT-EM by Varying the Check-in Points Per User



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

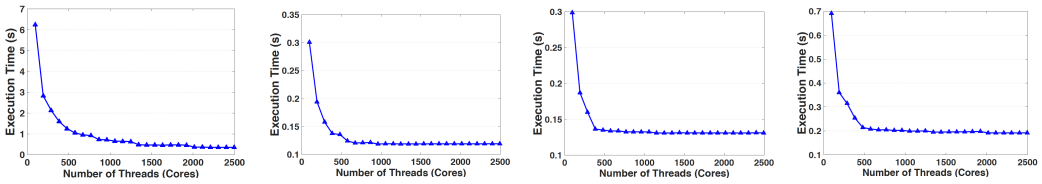
Fig. 14. Convergence Analysis of IGT-EM

the IGT-EM scheme without parallel implementation) and other baselines on a regular lab computer (4 cores and 2 GHZ for each core, 8GB memory). Table II presents the execution time required by all algorithms on four data traces. We observe that the parallel IGT-EM runs several orders of magnitude faster than the Sequential IGT-EM and other baselines. The performance gain is more significant on the New York trace since it has more users, venues and check-in points. The efficiency of the parallel IGT-EM is achieved by leveraging the computation powers from thousands of cores in the GPU. Figure 15 shows the execution time of the parallel IGT-EM with respect to varying number of threads (cores). Observe that the execution time drops quickly as the number of cores increases (i.e., more computational tasks in parallel IGT-EM run in parallel). We also examined the estimation performance of the parallel IGT-EM: the results are exactly the same as the ones shown in Figure 6 to Figure 9. The reason

is straightforward: the parallel IGT-EM is just a parallel implementation of the same IGT-EM algorithm we discussed in Section 4. We simply distribute the computation task of estimation parameters and hidden variables to different threads on different GPU cores, which should not affect the estimation performance (as verified by our experiments).

Table II. Execution Time Comparison

Algorithms	New York (s)	Chicago (s)	Boston (s)	Washington D.C (s)
Parallel IGT-EM	**0.46**	**0.119**	**0.192**	**0.131**
Sequential IGT-EM	909.90	34.47	26.20	65.01
Regular-EM	465.32	16.93	12.91	31.51
MLP	163.25	19.37	17.53	2.88
FM	266.09	4.15	17.53	13.71
FL	277.45	30.04	46.03	24.18
HLI	613.40	11.67	23.75	8.62
LP	148.55	16.46	19.28	3.42
Voting	21.69	0.62	0.47	1.25



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

Fig. 15. Effects of Number of Threads on Execution Time

7. DISCUSSION AND FUTURE WORK

In this paper, we develop an unsupervised learning framework to infer the localness of people in a city using incomplete geo-temporal data. Some directions exist for future work.

In this paper, we investigate the binary case on the localness of people: a person is either local or non-local. We also demonstrated our scheme is insensitive to the distance threshold used to decide whether a person is local or not. However, it might also be interesting to know the exact distance between the user's home location and the center of the city. Our current maximum likelihood estimation framework can be extended to solve such problem [Wang et al. 2014a]. In our next step, we plan to extend the binary hidden variable that represents the localness of a person to a continuous variable that could represent the exact distance between the person's home location and the center of the city. The key challenge would be how to develop the likelihood function of the MLE problem that could generate a close-loop solution.

The four datasets used in the experiments are collected from large cities in US (e.g., New York, Chicago, Washington D.C., and Boston). It is also interesting to study the performance of our method in small cities. We expect the performance of IGT-EM will be affected by the following factors in the small cities: (i) there are fewer number of users and venues in small cities so the IGT-EM algorithm is expected to converge faster; (ii) The small cities may have a different distribution of check-in points per venue and per user from large cities, which will affect the estimation accuracy of the IGT-EM scheme as well. Unfortunately, we do not have sufficient data from small cities

in our current datasets to rigorously evaluate the performance of our model. Therefore, we leave this as an important direction to further investigate in future study.

Our work also bears a resemblance to the works that estimate the locations of people by exploring the textual information from social media. For example, Cheng et al. developed a statistical model to estimate the location of Twitter users by analyzing the key phrases used in tweets [Cheng et al. 2010; 2013]. We can integrate the textual information into our IGT-EM scheme and further improve its estimation performance by exploring the content of user's texts. For example, a user might use words or languages that are specific to a city (e.g., local dialect) in the comments he/she submitted together with the check-in points at the visited venues. In such context, our model could be extended to incorporate the textual features as additional latent variables to characterize the localness of users. We expect this extension to further improve the estimation accuracy of our scheme.

The inputs to the IGT-EM scheme (i.e., VU Matrix and T Vector) can be readily obtained from the location-based online social network services. This minimal requirement on prior/external knowledge makes the proposed method robust and generally applicable to different application scenarios [Huang et al. 2015; Huang and Wang 2016b; Marshall and Wang 2016; Marshall et al. 2016]. However, we might still be able to improve the performance of IGT-EM if additional information about users and venues is known to the application. For example, knowing some of the venue's local attractiveness a priori (e.g., Statue of Liberty is more likely to attract tourists than local people), we can initialize the IGT-EM scheme with a better start point (compared to a random start point). This will greatly expedite the convergence process of the EM algorithm and improve the response time of IGT-EM in large cities. The key challenge is how to incorporate the additional information into the proposed model without sacrificing the rigidity of the analytical framework. The authors are actively working on the above extensions.

In summary, this paper solves the problem of identifying the local people in a city by using the incomplete online check-in points from location based online social services. We develop an Incomplete-Geo-Temporal Expectation Maximization (IGT-EM) scheme that classifies local people from non-local ones under a rigorous analytical framework. We also develop a parallel implementation of IGT-EM based on a GPU with 2496 cores to improve the computing peformance. We study the performance of our new approach through four real world case studies using data from Foursquare. We demonstrated that our approach can accurately identify the local people in a city and significantly outperform other state-of-the-art baselines. The results of our paper are important because they can directly contribute to many crowdsourcing-based smart city applications where training data is difficult or expensive to obtain.

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

Amr Ahmed, Yucheng Low, Mohamed Aly, Vanja Josifovski, and Alexander J Smola. 2011. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 114–122.

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*. ACM, 61–70.

Paul N Bennett, Ryen W White, Wei Chu, Susan T Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 185–194.

Giuseppe Cardone, Luca Foschini, Paolo Bellavista, Antonio Corradi, Cristian Borcea, Manoop Talasila, and Reza Curtmola. 2013. Fostering participaction in smart cities: a geo-social crowdsensing platform. *IEEE Communications Magazine* 51, 6 (2013), 112–119.

Xuefeng Chen, Yifeng Zeng, Gao Cong, Shengchao Qin, Yanping Xiang, and Yuanshun Dai. 2015. On Information Coverage for Location Category Based Point-of-Interest Recommendation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 759–768.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2013. A content-driven framework for geolocating microblog users. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4, 1 (2013), 2.

Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z Sui. 2011. Exploring Millions of Footprints in Location Sharing Services. *ICWSM* 2011 (2011), 81–88.

Jennifer Golbeck. 2009. Trust and nuanced profile similarity in online social networks. *ACM Transactions on the Web (TWEB)* 3, 4 (2009), 12.

Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779–782.

Tian-ran Hu, Jie-bo Luo, Henry Kautz, and Adam Sadilek. 2015. Home location inference from sparse and noisy data: models and applications. In *ICDM*. IEEE, 1382–1387.

Chao Huang and Dong Wang. 2016a. Exploiting spatial-temporal-social constraints for localness inference using online social media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 287–294.

Chao Huang and Dong Wang. 2016b. Topic-Aware Social Sensing with Arbitrary Source Dependency Graphs. In *2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 1–12.

Chao Huang, Dong Wang, and Nitesh Chawla. 2015. Towards Time-Sensitive Truth Discovery in Social Sensing Applications. In *Mobile Ad Hoc and Sensor Systems (MASS), 2015 IEEE 12th International Conference on*. IEEE, 154–162.

David Jurgens. 2013. That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships.. In *ICWSM*. AAAI, 273–282.

Kazuki Kodama, Yuichi Iijima, Xi Guo, and Yoshiharu Ishikawa. 2009. Skyline queries based on user locations and preferences for making location-based recommendations. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*. ACM, 9–16.

Cliff AC Lampe, Nicole Ellison, and Charles Steinfield. 2007. A familiar face (book): profile elements as signals in an online social network. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 435–444.

Rui Li, Shengjie Wang, and Kevin Chen-Chuan Chang. 2012a. Multiple location profiling for users and relationships from social network and content. *Proceedings of the VLDB Endowment* 5, 11 (2012), 1603–1614.

Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. 2012b. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1023–1031.

Chris YT Ma, David KY Yau, Nung Kwan Yip, and Nageswara SV Rao. 2013. Privacy vulnerability of published anonymous mobility traces. *Networking, IEEE/ACM Transactions on* 21, 3 (2013), 720–733.

Jermaine Marshall, Munira Syed, and Dong Wang. 2016. Hardness-aware truth discovery in social sensing applications. In *Distributed Computing in Sensor Systems (DCOSS), 2016 International Conference on*. IEEE, 143–152.

Jermaine Marshall and Dong Wang. 2016. Mood-Sensitive Truth Discovery For Reliable Recommendation Systems in Social Sensing. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 167–174.

Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. 2013. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 459–468.

CUDA Nvidia. 2008. Programming guide. (2008).

Moon-Hee Park, Jin-Hyuk Hong, and Sung-Bae Cho. 2007. Location-based recommendation system using bayesian users preference model in mobile devices. In *Ubiquitous Intelligence and Computing*. Springer, 1130–1139.

Ulrike Pfeil, Raj Arjan, and Panayiotis Zaphiris. 2009. Age differences in online social networking–A study of user profiles and the social capital divide among teenagers and older users in MySpace. *Computers in Human Behavior* 25, 3 (2009), 643–654.

Foster Provost, Brian Dalessandro, Rod Hook, Xiaohan Zhang, and Alan Murray. 2009. Audience selection for on-line brand advertising: privacy-friendly social network targeting. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 707–716.

Lakshmish Ramaswamy, P Deepak, Ramana Polavarapu, Kutila Gunasekera, Dinesh Garg, Karthik Visweswariah, and Shivkumar Kalyanaraman. 2009. Caesar: A context-aware, social recommender system for low-end mobile devices. In *Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on*. IEEE, 338–347.

Carlo Ratti, S Williams, D Frenchman, and RM Pulselli. 2006. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B Planning and Design* 33, 5 (2006), 727.

Dong Wang, Tarek Abdelzaher, and Lance Kaplan. 2014a. Surrogate mobile sensing. *Communications Magazine, IEEE* 52, 8 (2014), 36–41.

Dong Wang, Md Tanvir Al Amin, Tarek Abdelzaher, Dan Roth, Clare R Voss, Lance M Kaplan, Stephen Tratz, Jamal Laoudi, and Douglas Briesch. 2014b. Provenance-assisted classification in social networks. *IEEE Journal of Selected Topics in Signal Processing* 8, 4 (2014), 624–637.

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

Dong Wang, Lance Kaplan, Hieu Le, and Tarek Abdelzaher. 2012. On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proceedings of the 11th international conference on Information Processing in Sensor Networks*. ACM, 233–244.

Ryen W White, Peter Bailey, and Liwei Chen. 2009. Predicting user interests from contextual information. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 363–370.

Hongzhi Yin, Yizhou Sun, Bin Cui, Zhiting Hu, and Ling Chen. 2013. Lcars: a location-content-aware recommender system. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 221–229.

Jia-Dong Zhang and Chi-Yin Chow. 2015. GeoSoCa: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 443–452.