Landmark-Based User Location Inference in Social Media

Yuto Yamaguchi*
University of Tsukuba, Japan
yuto_ymgc@
kde.cs.tsukuba.ac.jp

Toshiyuki Amagasa
University of Tsukuba, Japan
amagasa@
cs.tsukuba.ac.jp

Hiroyuki Kitagawa
University of Tsukuba, Japan
kitagawa@
cs.tsukuba.ac.jp

ABSTRACT

Location profiles of user accounts in social media can be utilized for various applications, such as disaster warnings and location-aware recommendations. In this paper, we propose a scheme to infer users' home locations in social media. A large portion of existing studies assume that connected users (i.e., friends) in social graphs are located in close proximity. Although this assumption holds for some fraction of connected pairs, sometimes connected pairs live far from each other. To address this issue, we introduce a novel concept of *landmarks*, which are defined as users with a lot of friends who live in a small region. Landmarks have desirable features to infer users' home locations such as providing strong clues and allowing the locations of numerous users to be inferred using a small number of landmarks. Based on this concept, we propose a landmark mixture model (LMM) to infer users' location. The experimental results using a large-scale Twitter dataset show that our method improves the accuracy of the state-of-the-art method by about 27%.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.3.5 [**Information Storage and Retrieval**]: On-line Information Services—*Web-based services*

Keywords

location inference, user profiling, twitter, social graphs, landmarks

1. INTRODUCTION

As the use in mobile devices grows, the amount of location-related information from social media users increases. For example, Facebook users share which places they like, Twitter users transmit what happens where they are, and Foursquare users share where they visit. In such situations, users' home location profiles, which are the focus of this paper, become more important due to their usefulness in various applications (e.g., disaster warnings, location-aware recommendations, and advertisements).

However, most of users do not explicitly provide their location profiles. According to Cheng et al. [5], 76% of Twitter users do not make public their home locations at the city level. Similarly, Backstorm et al. [2] also reported that 94% of Facebook users do not provide their residential locations. These limitations reduce the usefulness of such location-aware services. Herein we deal with the problem of inferring users' home locations in social media.

In recent yeas, users' home location inference is a well-studied problem [5] [2] [14] [15]. Most major location inference methods employ user-generated contents (e.g., tweets) and/or social graphs. User-generated contents provide *local words* like the name of popular venues. These data can be utilized as clues for location inference [5]. For example, that a user who posts "rockets" may live near Houston.

Social graphs provide different clues. Major graph-based inference approaches [2] [14] [15] assume that connected users on social graphs are located near each other. Connected users in Facebook are friends where every edge is mutual, or in Twitter are either friends or followers where each edge has a direction. Although this assumption holds for some fraction of connected pairs, a significant percentage are geographically distant from each other.

Figure 1 shows the distribution of distances between home locations of mutually connected users in Twitter. 60% or more of connected pairs are at least 100km apart from each other. Hence, this *closeness assumption* between connected users may not provide us with strong clues for location inference.

Herein to improve the accuracy of the graph-based inference method, we introduce a novel concept of *landmarks*. Landmarks are users who have the following two characteristics: 1) a lot of friends and 2) the home locations of these friends are near each other. For example, if user u has a lot of friends whose locations are mostly in Boston, we regard user u as a landmark in Boston. After identifying this landmark, if another user v follows this landmark u, we infer that the user v lives in Boston. In this case, the city of Boston is this landmark's *dominance location*.

Landmarks have two desired features for location inference:

1. **Strong clues**: Due to the friends' geographical proximity, solid inferences can be made utilizing landmarks.

*Research Fellow of the Japan Society for the Promotion of Science, DC1

COSN'13, October 7–8, 2013, Boston, Massachusetts, USA.

http://dx.doi.org/10.1145/2512938.2512941.

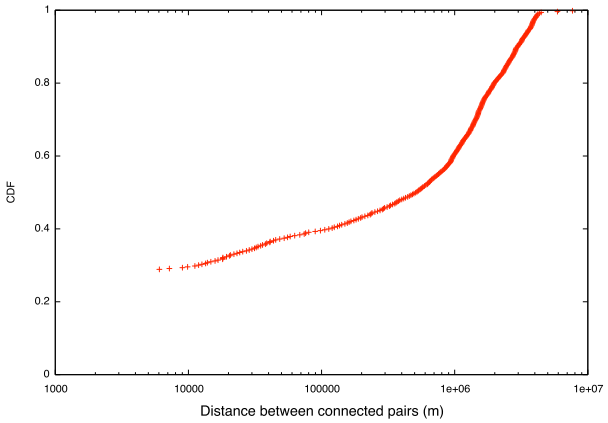

Figure 1: Distance distribution between home locations of mutually connected users in Twitter. 60% or more of connected users are at least 100km apart.

2. **Wide coverage**: Due to the large *centrality* of landmarks in a social graph, only a few landmarks are necessary to infer the most part of users in the graph.

Landmarks provide strong clues. Suppose that 80% of a Boston landmark ' s friends live in Boston, and these friends provide a clue with an 80% confidence level that a new friend of the landmark also lives in Boston even if the new friend ' s location is unavailable. Herein this is called the *concentration assumption*. Based on this concept, we propose a *landmark mixture model* (LMM) to infer users ' home locations. LMM models both dominance locations of landmarks and the home locations of users as continuous probability distributions over a geographical space. Specifically, the distributions of home locations are modeled as mixtures of the distributions of dominance locations of landmarks.

Because LMM employs the concentration assumption instead of the closeness assumption, it has advantages for location inference. First, LMM allows the trade-off between precision and recall to be exploited. In this context, precision refers to the ratio of correctly inferred users, while the recall is the ratio of inferred users versus all users. Because home locations are modeled as probability distributions, decisions can be made based on the distribution shape. If the home location distribution of a user has a clear peak at certain location, the user ' s location can be confidently determined. On the other hand, if the distribution lacks a clear peak, the user location can be excluded to avoid an incorrect inference. This is achieved by imposing the *confidence constraint*.

Second, LMM allows the trade-off between computational cost and recall to be exploited. Because finding the mode point in the mixture model is inherently costly, location inference based on LMM may also be costly. However, LMM reduces the cost based on the observation that user locations can be inferred by using only a small number of landmarks due to their wide coverage. In other words, by imposing a *centrality constraint*, we can make reasonable inferences using a mixture model with a small number of mixed components.

The contributions of this paper can be summarized as follows:

- We introduce a novel concept of *landmarks*, which have desirable features for location inference.
- Instead of the widely adopted closeness assumption, we introduce the *concentration assumption*.
- We propose the *landmark mixture model* (*LMM*) to infer users' home locations. This model can adjust the trade-offs between precision and recall and between computational cost and recall.

Experimental results using a large Twitter dataset show that LMM successfully achieves 75.4% precision, while preserving 85.0% recall, which improves the precision of the state-of-the-art method by about 27%. The results also demonstrate that LMM flexibly adjusts the abovementioned trade-offs; raising the precision to about 90% while preserving 60% of the recall; and the cost is reduced to 10% while preserving 85% of recall.

The rest of this paper is organized as follows. Section 2 overviews related works with an emphasis on location-based social network analysis, user location inference, and applications of location-related information. Section 3 states the problem addressed in this paper and defines the terminology. The concept of landmarks are introduced in Section 4, and then our LMM is proposed in Section 5. Section 6 describes the experiments conducted to verify the effectiveness of our method compared to the other existing methods including the state-of-the-art one. Finally, Section 7 concludes the paper.

2. RELATED WORK

2.1 Location-based Social Network Analysis

Location-based social networks are attracting a lot of attention due to their large-scale location-related data and potential applications. Volkovich et al. [21] studied the relationship between the structural properties of social graphs and spatial distance. They reported that 1) users connected by strong ties are more likely to be in close proximity of each other, and 2) users in a densely connected subgraph are more likely to be located near each other.

Quercia et al. [16] investigated the geographical proximity of users in *ego-networks*, which are subgraphs composed of a user and his/her immediate neighbors. They analyzed several types of ego-networks, and found that strongly connected ego-networks (i.e., users in the ego-network are connected by strong ties or communicate with each other frequently) show a high geographical proximity.

Gao et al. [7] studied the influence of social and historical ties over users' check-ins. They proposed a method to predict users' next check-ins using a state-of-the-art language model.

Cho et al. [6] analyzed user movements in location-based social networks and proposed a method to infer a users ' trajectories where trajectories are modeled as a multi-state probabilistic generative model. They reported that both geographical constraints and social influence affect users.

The observations in these works are utilized in the user location inference tasks and other location-aware applications.

2.2 User Location Inference

Most major location inference methods on social media employ user-generated contents (e.g., tweets) and/or user-

relationships on social graphs. These methods can be classified into three categories: content-based, graph-based, and integrated approaches.

Content-based approaches. Content-based approaches take advantage of user-generated contents. Cheng et al. [5] inferred residential locations of Twitter users based on the *local terms* contained in tweet texts. Local terms are mainly posted by users in a specific geographical region. They are extracted using handmade training data. For example, their paper stated that the term "rockets" is a local term because it is frequently posted near Houston, Texas. Consequently, their method infers that users who post texts containing "rockets" are located near Houston.

Chang et al. [4] developed the location distributions of terms based on a *GMM* (*Gaussian Mixture Model*) to infer user locations. Unlike Cheng et al.'s method, Chang et al.'s method extracts local terms without handmade training data (i.e., unsupervised learning). Chang et al.'s experiments showed that their method, which is based on GMM, achieves a better accuracy than [5].

Kinsella et al. [10] developed a language model for each city using geotagged tweets, and proposed a location inference method. Because their language model utilizes geotagged tweets, it is more robust against user movements than methods that employ only user location profiles.

Chandra et al. [3] also dealt with the location inference task by focusing on users' *conversations*. They reported that all tweets in a conversation are on the same topic, and developed a language model where all terms in the same conversation belong to the conversation initiator. They experimentally confirmed that their inference method, which is based on their language model, shows better accuracies than models that do not consider conversations.

Unlike these methods, which utilize user-generated content, our method employs landmarks in social graphs.

Graph-based approaches. Graph-based approaches utilize user-relationships on social graphs. Backstorm et al. [2] proposed a method to infer locations of Facebook users based on the closeness assumption that connected users (i.e., friends) are likely to be located near each other. Their method calculates the likelihood of obtaining a given social graph. Locations with the largest likelihood are assigned to users as their residential locations. They achieved a better accuracy than the IP-based approach.

Clodoveu et al. [8] also proposed a graph-based method to infer location by simply considering that user u's location is where the dominant fraction of u's friends live. Users with a small number of friends may not be inferred accurately due to an insufficient number of clues. Similarly, users with many friends (e.g., celebrities, commercial accounts) may also be inferred inaccurately because their locations often lack locality.

Sadilek et al. [17] dealt with a slightly different problem for user trajectory inference. They considered that user trajectory inference and the link prediction on social graphs are mutually complementary. Their method infers user trajectories using a social graph and performs link predictions using trajectory data based on the fact that friends tend to move together.

These methods are all based on the closeness assumption. In contrast, our method is based on another observation in which users with a lot of friends in a small region (i.e., landmarks) exist. Our observation provides stronger clues than the closeness assumption.

Integrated approaches. Integrated approaches, which are the state-of-the-art methods to the best of our knowledge, use both user-generated contents and user relationships on social graphs. Li et al. [15] proposed a unified discriminative influence model (*UDI*) to infer users' home locations. UDI models user-generated contents and user-relationships as a heterogeneous graph, and assumes each node (i.e., user or venue) has its own influence scope. Nodes with larger influence scope (e.g., Lady Gaga) are more likely to be followed by distant users. Consequently, these types of nodes do not provide good clues for location inference. They have also proposed two inference methods based on UDI model, which maximize the likelihood of obtaining a heterogeneous graph.

Li et al. [14] also proposed another model, *multiple location profiling model* (*MLP*). They argue that some users have more than one locations, for example, their home location, their work location, and their former home location. Hence, MLP deals with a problem of inferring users' multiple locations. This model, which is based on the probabilistic generative model, determines its parameters are inferred using the Gibbs sampling method.

2.3 Applications of location-related information

Several studies are based on location-related information. Specifically, local event detection and location-aware searches and recommendation are receiving more attention.

Local event detection. Sakaki et al. [18] proposed a method to detect events such as earthquakes and typhoons using users' location profiles and geotagged tweets. Their method estimates event trajectories (e.g., typhoons) using the particle filter.

Lee et al. [12] proposed a local event detection method using Twitter where tweets are clustered based on Incremental DBSCAN in real time. Then the location of each cluster is inferred using users' timezones.

Walther et al. [22] developed a geospatial event detection system. They discussed what types of features are useful for event detection, and concluded that events can be detected if we examine the number of users who posts tweets in the same location and their themes. Their method uses such effective features and employs some ML methods such as decision trees to detect geospatial events.

Location-aware search and recommendation. Levandoski et al. [13] developed a location-aware recommender system (*LARDS*) after observing that users tend to prefer geographically close items (e.g., restaurants). LARDS creates a spatial grid, and provides different recommendations for each grid cell. To recommend items to moving users, LARDS developed the pyramid structure which efficiently maintain the structure of the grid.

Shaw et al. [19] proposed a spatiotemporal search scheme to support Foursquare users to *check-in* venues. To overcome the poor accuracy of GPS and the high density of urban area locations, their method exploits users' own check-

in histories, friends' check-in histories and the popularity of places.

3. PROBLEM STATEMENT

This section defines the terminology and describes the problem addressed in this paper. Social graph $G = (V, E)$ is a directed graph, where each edge $e = (v_i, v_j) \in E$ is directed. Similar to Twitter's vocabulary, we adopt the terms of *follower* and *friend*. When a user *follows* another user, then the former is called a follower of the latter user, while the latter user is called a friend of the former user. The vertex set is composed of two types of user set $V = U^L \cup U^N$ where U^L is a set of *labeled users* whose home locations are known in the form of latitude and longitude pairs[1], and U^N is a set of *unlabeled users* whose home locations are unavailable. Home locations are denoted as $\boldsymbol{l} = (lat, longi)$

In this notation, the problem of user location inference is stated as:

PROBLEM 1 (USER LOCATION INFERENCE). *Given a social graph $G = (V, E)$, infer the home location of each unlabeled user $u \in U^N$ so that the inferred location $\hat{\boldsymbol{l}_u}$ is close to the actual location $\boldsymbol{l_u}$.*

Instead of the widely adopted closeness assumption, we employ the concentration assumption to tackle this problem. Section 4 introduces the concept of landmarks, while Section 5 describes our landmark mixture model (LMM) to solve the user location inference problem.

4. LANDMARKS

To introduce landmarks, two measurements of landmarks, *centrality* and *dispersion*, need to be defined. Centrality, which is well-known in graph theory, is a measurement to determine the relative importance of a vertex within a graph. The dispersion of user u means how far u's neighbors (i.e., friends or followers) are located from each other. Note that dispersion does not depend on user u's own home location.

Based on centrality and dispersion, we define landmarks below.

DEFINITION 1 (LANDMARKS). *A landmark is a user account u with a large centrality c_u and a small dispersion d_u.* □

In this paper, we employ the degree centrality as centrality value c_u, and 2-dimensional spatial variance with respect to latitude and longitude as dispersion value d_u. Detailed definition is described in Section 5.2.

In the context of the social media where a social graph is directed like Twitter, there are two types of landmarks, *in-landmarks* and *out-landmarks*. To deal with these two types of landmarks, we also introduce the terms of *in-centrality* c_u^{in}, *out-centrality* c_u^{out}, *in-dispersion* d_u^{in}, and *out-dispersion* d_u^{out}. In-centrality and in-dispersion are measured using a user's followers (i.e., vertices with edge directed toward the user). Inversely, out-centrality and out-dispersion are measured using a user's friends.

Preliminary experiment. To demonstrate the presence of landmarks, we investigate the centrality and the dispersion of users in our Twitter dataset, which is described in Section 6.1. We employ the degree centrality for the centrality measure, and the variance of neighbors' locations for the dispersion measure (see Section 5.2).

[1]Practically, latitude and longitude pairs can be obtained by geocoding their location profiles.

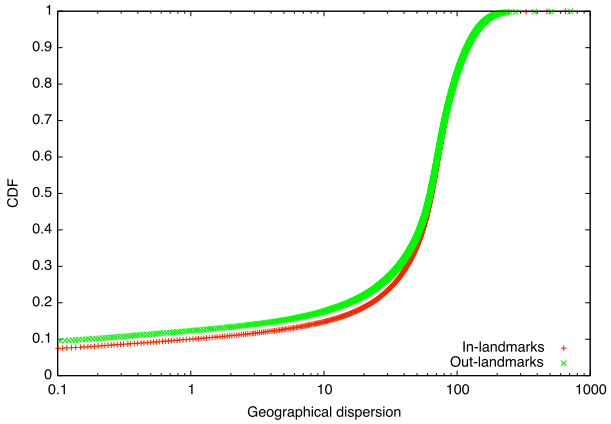

Figure 2: CDF of users over their dispersions for the top 5% users with high centralities. This figure indicates that there are y% of users have dispersions lower than x. Although most users have a large dispersion, there exist users with relatively small dispersion, which are regarded as landmarks.

If there exist landmarks in target dataset, we can say that the concentration assumption holds. In other words, there are some user groups whose locations are near each other and these users follow the same user (i.e., landmark). In this case, we can infer home locations of these users if their home locations are unknown, by propagate home locations of location-known users in the same group.

Figure 2 plots the CDF of users over their dispersions. This figure indicates that there are y% of users have dispersions lower than x. Note that the plotted users are limited to top 5% of users whose centralities are high in the dataset.

If we define users with dispersions less than 10 as landmarks, then 14% of plotted users in 2 can be regarded as in-landmarks and 17% as out-landmarks. Although the number of landmarks is rather small, landmarks do exist.

Figure 3 maps the above landmarks (i.e., users in the top 5% centrality, and less than 10 dispersion value). Red dots represent the home locations of all users in our dataset, while blue dots represent dominance locations of landmarks. First of all, most users, including landmarks, are located in the eastern part (e.g., east of the Mississippi River) of the United States, which is consistent with most of other works. Second, the distributions of both all users and landmarks are similar. Metropolises with a lot of users also have a lot of landmarks. Although most landmarks lie east of the Mississippi River, some cities west of the Mississippi river (e.g., Denver, Phoenix, and Salt Lake City) have a relatively large number of landmarks. Thus, landmarks can cover large segments of the user population; that is, most users can find landmarks near their home locations.

Examples. Table 1 shows some exsample landmarks in the dataset. The top two user accounts are regarded as in-landmarks, the middle two user accounts are out-landmarks, and the bottom two user accounts are both in-landmarks

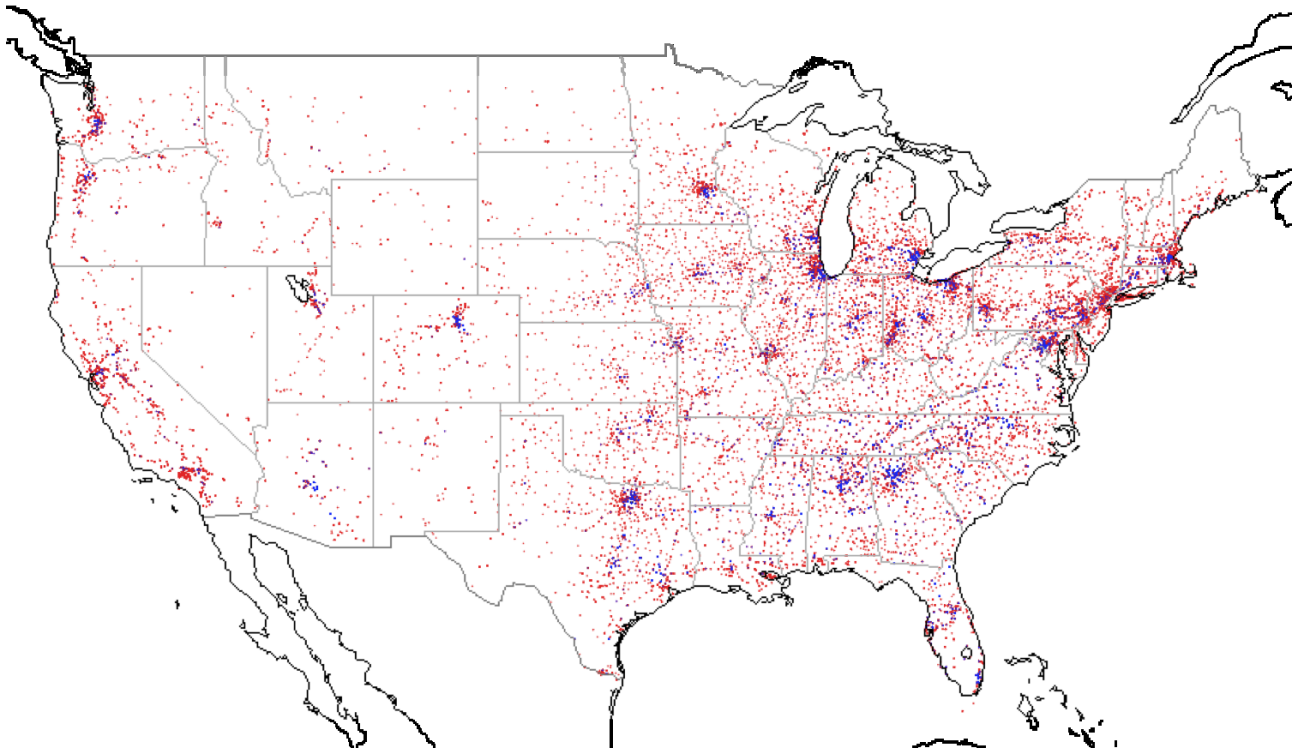

Figure 3: Distributions of both all users and landmarks. Red and blue dots represent home locations of all users in our dataset and dominance locations of landmarks, respectively. Both populations have similar distributions. Landmarks can cover a large segment of the user population; that is, most users can find a landmark near their home location.

and out-landmarks. These in- and/or out-landmarks tend to be local news accounts or commercial accounts, which are *bots* rather than *human accounts*. Specifically, most in-landmarks are local news accounts, which post about their local area. Although there are fewer out-landmarks, they tend to be commercial accounts. However, it should be noted that some user accounts can be regarded as both in-landmarks and out-landmarks. This type of landmark has a lot of followers and friends in a small region, indicating that it follows its followers back.

Our observations suggest that in-landmarks are authoritative user accounts that post useful tweets about their dominance location. Thus, in-landmarks can provide useful information about local locations. This observation provides another motivating factor to utilize in-landmarks to extract useful local information, but is beyond the scope of this paper and will be examined in the future.

On the other hand, out-landmarks are commercial accounts, including spammers, who want more followers in a small region. Although these landmarks do not post useful tweets, we can utilize them to address the home location inference problem.

5. LANDMARK MIXTURE MODEL

This section proposes the *landmark mixture model* (*LMM*) to address the user location inference problem. LMM models both the dominance locations of landmarks and the home locations of users as continuous probability distributions. Section 5.1 formulates the model, while Section 5.2 proposes a location inference method based on this model. Finally, Section 5.3 introduces the constraints to adjust the trade-offs.

5.1 Model Formulation

According to the definition, landmarks have small dispersions, leading to clear dominance locations. Hence, LMM estimates all users' dispersions and dominance locations, and then regards users with small dispersions as landmarks.

Dominance distribution. Similar to several other studies that model the probability distribution over a geographical space [20] [23] [15], we model the dominance location as a Gaussian distribution. We call this distribution the *dominance distribution*. The underlying idea is that the Gaussian distribution has two parameters, mean and variance, which represent the dominance location and the dispersion, respectively. The value of the probability density for each location point indicates how the likelihood of a user's dominance location. A dominance distribution with a large variance (i.e., large dispersion) does not have a clear peak, indicating that the user lacks a dominance location. Consequently, the user is not a landmark.

Based on the above idea, we assign a Gaussian distribution $N(\boldsymbol{\mu_u}, \boldsymbol{\Sigma_u})$ for each user u. The mean parameter $\boldsymbol{\mu_u}$ denotes the dominance location of user u, while the covariance matrix parameter $\boldsymbol{\Sigma_u}$ denotes the dispersion of u. Herein we assume that the shape of the dominance distribution is

Table 1: Examples of landmarks

User Name	*Profile*	*Center Point*	*Centrality (in : out)*	*Dispersion (in : out)*
denvernews	Denver-specific news from The Denver Post. ...	39.73, -105.0 (Denver,CO)	20,526 : 7,418	0.0013 : 20.36
BostonFire	Official Twitter Boston Fire. Spring starts outdoor grilling. ...	42.32, -71.09 (Boston,MA)	34,111 : 49	0.6523 : 56.60
HomeTheaterMI	Genesis Electronics is a family-owned business ...	42.39, -83.13 (Detroit,MI)	2,331 : 1,841	16.66 : 0.0067
alabamanews1	All Alabama News!	33.52, -86.81 (Birmingham,AL)	836 : 1,523	10.51 : 0.3815
komonews	The latest breaking news, traffic, and weather from Seattle ...	47.63, -122.3 (Seattle,WA)	29,989 : 2,102	0.0107 : 0.0029
OWHnews	Updates from Omaha.com and the Omaha World-Herald ...	41.26, -96.01 (Omaha, NE)	17,400 : 9,662	0.0306 : 0.0306

symmetric, that is,

$$\boldsymbol{\Sigma_u} = \begin{pmatrix} d_u & 0 \\ 0 & d_u \end{pmatrix}, \tag{1}$$

where the diagonal components are the dispersions. It should be noted that users have two types of dispersions. We assign two Gaussian distributions for each user: in-dominance distribution $N(\boldsymbol{\mu_u^{in}}, \boldsymbol{\Sigma_u^{in}})$ and out-dominance distribution $N(\boldsymbol{\mu_u^{out}}, \boldsymbol{\Sigma_u^{out}})$.

LMM. Using the dominance distributions, landmark mixture model models the home locations of users as continuous probability distributions. Following a landmark provides a strong clue for inferring a user's location because most of the landmark's neighbors are in close proximity. On the other hand, following an ordinal user (i.e., a non-landmark) does not provide a good clue because the locations of neighbors of an ordinal user are geographically dispersed.

Based on this idea, LMM is modeled as a *Gaussian mixture model* (*GMM*) where the dominance distributions are mixed. Specifically, a distribution of user u's home location is modeled as a GMM where each Gaussian component is the dominance distribution of u's neighbors. We call this the *location distribution*. The location distribution is denoted as:

$$\begin{aligned} P_u(\boldsymbol{x}) &= \sum_{v \in N_u^{out}} \pi_v^{in} N(\boldsymbol{x}|\boldsymbol{\mu_v^{in}}, \boldsymbol{\Sigma_v^{in}}) \\ &+ \sum_{w \in N_u^{in}} \pi_w^{out} N(\boldsymbol{x}|\boldsymbol{\mu_w^{out}}, \boldsymbol{\Sigma_w^{out}}), \end{aligned} \tag{2}$$

where N_u^{in} is the set of followers of u, N_u^{out} is the set of friends of u, and π_v is the mixture weight. Mixture weights are defined as:

$$\pi_v^{in} \propto \log c_v^{in}, \tag{3}$$

$$\sum_{v \in N_u^{out}} \pi_v^{in} + \sum_{w \in N_u^{in}} \pi_w^{out} = 1, \tag{4}$$

where c_v^{in} is in-degree centrality of user v. The reason that we employ the logarithm of degree centrality is degree of users in social graphs follows the *power law*. In social graphs, some users have huge degree values, which requires to moderate these values.

The probability density at a location represents the likelihood of a user's home location. Hence, if a user's location distribution has a clear peak at a specific locale, we can confidently state that identify the user's home location.

LMM does not explicitly differentiate landmarks from ordinal users in its location inference process. Instead, it imposes weights (i.e., mixture weights and variances) on all users to implicitly differentiate them. A Gaussian component with a small variance and large mixture weight, which corresponds to a dominance distribution of a landmark, strongly affects the shape of the overall location distribution.

Consequently, our model mostly uses landmarks to determine a user's home location.

5.2 Inference Method

Given a social graph G, we initially estimate the parameters of the dominance distributions for all users. Based on the maximum likelihood criteria, the parameters are estimated using the location points of users' neighbors as

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{u}}^{\boldsymbol{in}} = \frac{1}{|N_u^{in}|} \sum_{v \in N_u^{in}} \boldsymbol{l_v}, \tag{5}$$

$$\hat{\boldsymbol{d}}_{\boldsymbol{u}}^{\boldsymbol{in}} = \frac{1}{2|N_u^{in}|} \sum_{v \in N_u^{in}} (\boldsymbol{l_v} - \hat{\boldsymbol{\mu}}_{\boldsymbol{u}}^{\boldsymbol{in}})^2. \tag{6}$$

The parameters for the out-dominance distributions are estimated in a similar manner.

However, noises strongly influence the mean. We found that some of the neighbors are located far from the other neighbors. To suppress the noise effect, we employ the median because the median is more robust against noises than the mean.

After the parameters of the dominance distributions are set, we can construct the users' location distributions. In this paper, we use the degree centrality for the centrality measurement because landmarks with a large number of immediate neighbors can be utilized to infer more users' locations. The estimated location distributions can be simply written as

$$\begin{aligned} P_u(\boldsymbol{x}) &= \sum_{v \in N_u^{out}} \pi_v^{in} N(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_{\boldsymbol{v}}^{\boldsymbol{in}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}}^{\boldsymbol{in}}) \\ &+ \sum_{w \in N_u^{in}} \pi_w^{out} N(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_{\boldsymbol{w}}^{\boldsymbol{out}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{\boldsymbol{out}}). \end{aligned} \tag{7}$$

It should be noted that statistical inference methods (e.g., the EM algorithm [7]) are unnecessary because LMM simply mixes the dominance distributions. This substantially reduces the parameter estimation cost.

Based on this model, the user location is inferred as the location with the largest probability density (i.e., mode point)

$$\hat{\boldsymbol{l_u}} = \arg\max_{\boldsymbol{x}} P_u(\boldsymbol{x}). \tag{8}$$

The computational complexity of finding the mode point of GMM is $O(k^2)$, where k is the number of Gaussian components of GMM. This can be explained as follows. The candidates for the mode point of GMM are limited to the center of each component because only the derivative at the center is 0. For each candidate point, we sum up the probability densities of all components at the point to determine the candidate with the largest probability density. This process has a relatively high computational cost. However, we

can reduce this by imposing centrality constraint described in the next section.

5.3 Constraints to Adjust the Trade-offs

LMM can adjust the trade-offs between precision and recall, and between computational cost and recall by imposing the *confidence constraint* and *centrality constraint*, respectively.

Confidence constraint. The process to find the mode point also gives the probability density p at that point, which indicates how likely the corresponding user's home location is at that point. If p is small, the confidence of the inference is low. To avoid making an unconfident inference, we impose the *confidence constraint*.

DEFINITION 2 (CONFIDENCE CONSTRAINT). *If the probability density p_u of user u's location distribution at the mode point is less than the predefined threshold p_0, the location of user u is not inferred.* □

As the value of p_0 increases, the precision increases, but the recall decreases. On the other hand, as p_0 decreases, the opposite is true. This trade-off is examined in Section 6.3.

Centrality constraint. Because LMM does not explicitly discriminate between landmarks and ordinal users, it uses the dominance distributions of all users to infer the location. This causes a relatively expensive computational cost. Because landmarks provide the strong clues and have the wide coverage, we can infer the locations for a large segment of users using a small number of landmarks. To reduce the computational costs, we impose the *centrality constraint*.

DEFINITION 3 (CENTRALITY CONSTRAINT). *If user u's centrality c_u is lower than the predefined threshold c_0, the dominance distribution of user u is not used for the inference.* □

Because users with a low centrality are not regarded as landmarks, they do not provide good clues for location inference. The centrality constraint reduces the computational cost by eliminating the dominance distributions of these ordinal users in the location inference step.

Even if we exclude a significant fraction of users whose centralities are low, most users are connected to at least one landmark. This can be explained by the fact that the degree distribution of the Twitter social graph follows the *power law*[2] [11]. Based on *percolation theory*, in scale-free networks, the most vertices are connected even if a lot of vertices are removed as long as the degrees of removed vertices are small [1].

If the threshold c_0 is large, then it is expected that both the computational cost and recall decrease. On the other hand, if c_0 is small, then a decrease in computational costs is small and the recall remains high. This trade-off is verified in Section 6.4.

6. EXPERIMENTS

This section describes the experiments to:

1. Compare the precision and the recall of our proposed method to other existing methods.
2. Compare the precision and the recall between variations of our proposed method.
3. Evaluate the trade-offs of precision, recall, and computational cost of our method by varying the threshold values of the two constraints described in Section 5.3.

Section 6.1 explains the experimental conditions, while Sections 6.2-6.4 describe the results.

6.1 Experimental Setups

Dataset. We used the dataset from Li et al. [15]. This dataset is composed of 3,122,842 Twitter users in the United States with 284,884,514 edges. Similar to previous studies, we geocoded users' location profiles into latitude and longitude pairs using the 2010 census U.S. gazetteer[3]. Specifically, we converted location profile texts in the form of *cityName,stateName* or *cityName,stateAbbreviation* into latitude and longitude pairs. As a result, we obtained 464,794 (14.9%) labeled users. Note that misreports of location in location profiles can degrade the location inference. However, Jurgens et al. [9] experimentally show that there are not so many misreports of location. So we believe users' location profiles show their true home locations.

To evaluate the precision, we randomly divided the labeled users into a test set and a training set, where 10% were assigned to the test set and the rest were assigned to the training set.

Implementation. We implemented our proposed method and other existing methods as described in Section 6.2. Our code is available at `http://github.com/yamaguchiyuto/tomato`.

Evaluation metrics. We evaluated our method and existing methods using five metrics.

- *Precision*: The ratio of correctly inferred users versus all inferred users. If the error distance between the inferred and actual location is less than 160 km (100 miles), the inference is assumed to be accurate. This metric has been used in [4] [5] [15].
- *Recall*: The ratio of inferred users versus all labeled users in the test set.
- *F-measure*: The harmonic mean of the precision and recall, which is denoted as $F = (2 \cdot P \cdot R)/(P + R)$, where P and R are the values of precision and recall, respectively.
- *Mean E.D.*: The mean error distance between the inferred location and the actual location.
- *Median E.D.*: The median error distance between the inferred location and the actual location.

In addition, we employ *accumulative precision* at various distances, which shows that y% of users' error distances are within $x$km.

In our inference method, the dominant part in terms of computational complexity is determining the mode point of LMM, which is $O(k^2)$ where k is the number of mixed components. Hence, to evaluate the computational cost of our

[2]In fact, [11] reported that there are some Twitter users who have higher in-degrees than expected.

[3]`http://www.census.gov/geo/maps-data/data/gazetteer2010.html`

method, we use the average number of neighbors (i.e., the average number of mixed components in LMM) in Section 6.4.

Constraints. The threshold p_0 of the confidence constraint is varied in Section 6.3 to examine its effect on the trade-off between precision and recall. The threshold c_0 of the centrality constraint is used in Section 6.4 to examine its effect on the trade-off between recall and computational cost. If the threshold values are not clearly specified, c_0 is not used and p_0 is set to 0.003, which achieves the best F-measure.

6.2 The Performance Comparisons

This section shows the results of two experiments: one that compares our method to existing methods and the other compares variations in our method.

Comparison with existing methods. Our method is compared to a state-of-the-art method and a naive method. These two existing methods are based on the closeness assumption.

The state-of-the-art method proposed by Li et al.[15] is called UDI, which is described in Section 2. For UDI, we employ the *global prediction method*[4] as its inference method. Although their model can integrate user-generated contents and a social graph, we do not use user-generated contents because the objective of this experiment is to compare the performance of these graph-based methods.

The naive method infers user u's locations by simply calculating the medoid of locations of user u's neighbors. Note that this method uses both followers and friends as neighbors, which achieves better precision and recall than the case using only followers or friends.

Table 2 summarizes the results. Our method achieves the best precision and F-measure. Our method has an improved precision of 27% compared to the state-of-the-art method and preserves 85% of the recall. UDI shows approximately the same results as the original paper.

In addition, the mean E.D. and median E.D. of our method are substantially improved compared to the other two methods. Figure 4 shows the accumulative precision at various distances. Our method reduces the error distances. Specifically, about half of the users are located within a 1km error distance using our method.

These results indicate that our concentration assumption provides better clues than the closeness assumption. If we can find landmarks in social media, they provide the home locations of other users accurately.

The recall of our method drops to a lower value because if the probability density at the mode point is lower than p_0, our method can decide not to infer the user location. The effect of this constraint is verified in Section 6.3.

Comparison of variations of the proposed method. To examine which part of the method leads to the good results, five variations of our method are compared.

- *LMM*: Our method described as Section 5.

[4]The authors of [15] proposed two types of inference methods: global prediction method and local prediction method.The former achieved a higher accuracy than the latter.

Table 2: Summary of the comparison of our method to existing methods.

	LMM	*UDI*	*Naive*
Precision	**0.754**	0.594	0.445
Recall	0.850	**0.926**	0.900
F-measure	**0.799**	0.724	0.596
Mean E.D.	**297,739**	542,483	616,196
Median E.D.	**3,804**	37,363	249,982

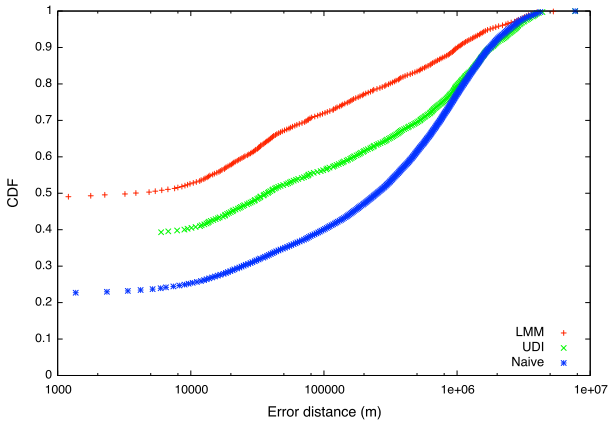

Figure 4: Accumulative precision of our method and existing methods at various error distances. Our method successfully locates about half of the users within a 1km error distance, and outperforms the other two methods, including the state-of-the-art method.

- *LMM w/o m*: Our method without mixture weights. This method uses the same value for all mixture weights.
- *LMM w/o mv*: Our method without mixture weights and variances. This method uses the same value for all mixture weights, and regards variances of all the Gaussian distributions as 1.
- *Medoid*: A method that simply calculates the medoid of neighbors' dominance locations. This method does not use the dominance distribution.
- *Centoid*: A method that simply calculates the centroid of neighbors' dominance locations. This method does not use the dominance distribution.

Table 3 summarizes the results. Contrary to our expectations, LMM and LMM w/o m give approximately the same results. These two methods form almost the same curve 5, indicating that the mixture weights do not improve the precision. There are three reasons that the mixture weights do not work well. 1) Even if users have relatively small centralities, they provide some clues as long as they have small dispersions. 2) Most of the users have small centralities, which results in discarding a substantial part of the clues by imposing small weights. 3) Users with large centralities are weighted heavily regardless of their dispersions. Hence, we have to carefully develop the mixture weight, and this is our future work.

Table 3: Summary of the comparison of variations of our method.

	LMM	*LMM w/o m*	*LMM w/o mv*	*Medoid*	*Centroid*
Precision	0.754	**0.757**	0.543	0.357	0.274
Recall	0.850	0.846	**0.996**	**0.996**	**0.996**
F-measure	0.799	**0.800**	0.703	0.526	0.429
Mean E.D.	297,739	**292,917**	587,857	698,769	705,689
Median E.D.	3,804	**2,694**	75,885	413,459	455,695

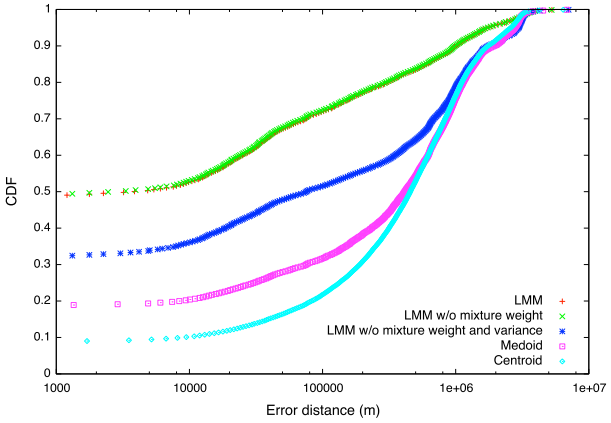

Figure 5: Accumulative precision for variations of our method at various error distances. Although considering the geographical dispersion improves the precision, considering the mixture weight does not. Medoid and Centroid do not work well because they simply use the dominant location as the points rather than as the distributions.

The other three variations do not show good results (3 and 5), indicating that employing dispersion leads to good results. Moreover, comparing LMM w/o mv and Medoid indicates that the considering the dominance location as the probability distribution rather than just a location point positively influences the results. Although not all users provide clues, users with a small dispersion provide significant clues for location inference.

6.3 Effect of the Confidence Constraint

LMM can adjust the trade-off between precision and recall by imposing the confidence constraint. This section shows the effect of the confidence constraint. Figure 6 shows the result by varying the value of threshold p_0. The x-axis denotes the value of p_0, and each line denotes the precision, recall, and F-measure. As the value of p_0 increases, the precision increases but the recall decreases. The F-measure achieves the best score around $p_0 = 0.003$. Note that even if we do not impose the confidence constraint ($p_0 = 0$), our method outperforms other methods for all these metrics.

If the probability density at the mode point is low, the overall probability distribution does not have a clear peak. In this case, we should not infer the home location of that user because there are insufficient clues to determine the location. Our method can select this option because it is based on the probability distribution.

If we require a high precision (e.g., in the case of sending disaster warnings), we can achieve that by imposing the large p_0 value. On the other hand, if we want a high recall

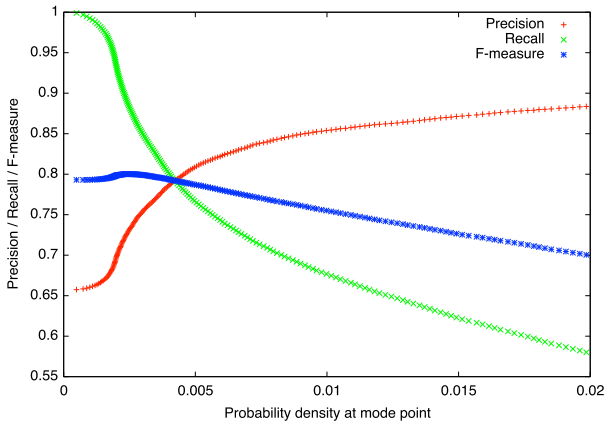

Figure 6: Effect of the confidence constraint. x-axis denotes the value of the threshold p_0. As the value of p_0 increases, the precision increases but the recall decreases. The F-measure achieves the best score around $p_0 = 0.003$, and the precision is about 0.88 at $p_0 = 0.02$.

(e.g., local advertisements), we can get that by imposing no constraint, or small p_0 value. The ability of this trade-off adjustment may expand the applications of users' home location profiles.

6.4 Effect of the Centrality Constraint

LMM can also adjust the trade-off between the computational cost and recall by imposing the centrality constraint. Varying the value of threshold c_0 demonstrates the effects of the cost, recall, and precision. The confidence constraint is not imposed in this section.

Figure 7 shows the results where the x-axis denotes the value of c_0. The left y-axis denotes the values of precision, recall, and cost, while the right y-axis denotes the average number of neighbors (or components). The ratio of the utilized landmarks means the ratio of users satisfying the centrality constraint (i.e., users with their centralities $c_u > c_0$) versus all users. The average number of neighbors means the average number of each user's neighbors, in other words, the average number of mixed components for each user's location distribution, which dominates the computational complexity of our method.

As the value of c_0 increases, the ratio of utilized landmarks decreases rapidly but the recall remains high. Hence, we conclude that our method can infer locations of almost all users with only about 5% of landmarks ($c_0 = 100$). This means only 5% of landmarks ' Gaussian distributions (i.e., mean and variance parameters) and following relationships need to be stored to infer user locations.

In terms of computational cost, we can reduce the average number of neighbors to approximately 30%, preserving 85% of the recall ($c_0 = 200$). This means that because the computational complexity of our method is $O(k^2)$, where k is the number of neighbors, the cost is reduced to about 10%.

From $c_0 = 0$ to 400, the precision remains about the same value or even decreases, but from $c_0 = 500$ to 1000, it increases. Two factors may lead to such a behavior. If our method utilizes landmarks with high centrality, good results

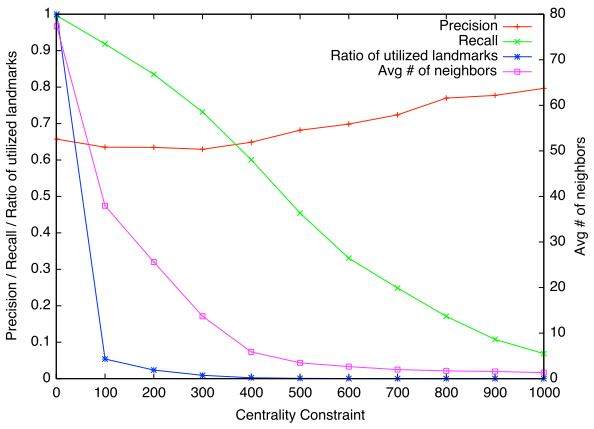

Figure 7: Effect of the centrality constraint. x-axis denotes the value of the threshold c_0. As c_0 increases, the ratio of utilized landmarks rapidly decreases, preserving the high recall value. The decrease in the average number of neighbors denotes the reduction of the computational cost of our method. The precision remains high or even increases when we utilize a small number of landmarks.

are achieved because their small dispersion values are statistically significant. On the other hand, because only a small number of landmarks have a high centrality, other users do not satisfy the centrality constraint and are not used for location inference. Ignoring these ordinal users, which is most of the users, may degrade the performance.

7. CONCLUSION

Hewein we introduce a novel concept of landmarks and propose a landmark mixture model (LMM) to address the user location inference problem. Landmarks have desirable features for location inference: strong clues and a wide coverage. LMM can adjust the trade-offs between precision and recall and between computational cost and recall. This capability may expand applications employing users' home locations. The experimental results show that our inference method outperforms other existing methods, including the state-of-the-art method. The results also demonstrate that imposing the two constraints allows our method to realize a high precision, reduce the computational cost, and preserve a high recall.

Our future work includes 1) refining our method to iteratively propagate landmark's clues to increase the inferenve coverage, 2) improving our method by integrating it with methods employing user-generated content, and 3) examining the other applications of landmarks such as recommending landmarks' tweets to travelers and searching local information utilizing landmarks.

Acknowledgements

This work was supported in part by JSPS KAKENHI, Grant-in-Aid for JSPS Fellows #242322.

8. REFERENCES

[1] R. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

[2] 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, 2010.

[3] S. Chandra, L. Khan, and F. B. Muhaya. Estimating twitter user location using social interactions-a content based approach. In *SocialCom/PASSAT*, pages 838–843, 2011.

[4] H.-W. Chang, D. Lee, M. Eltaher, and J. Lee. @phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *ASONAM*, pages 111–118, 2012.

[5] 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, 2010.

[6] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082–1090, 2011.

[7] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *ICWSM*, 2012.

[8] C. A. D. Jr., G. L. Pappa, D. R. R. de Oliveira, and F. de Lima Arcanjo. Inferring the location of twitter messages based on user relationships. *T. GIS*, 15(6):735–751, 2011.

[9] D. Jurgens. That ' s what friends are for: Inferring location in online social media platforms based on social relationships. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[10] S. Kinsella, V. Murdock, and N. O'Hare. "i'm eating a sandwich in glasgow": modeling locations with tweets. In *SMUC*, pages 61–68, 2011.

[11] H. Kwak, C. Lee, H. Park, and S. B. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600, 2010.

[12] C.-H. Lee, H.-C. Yang, T.-F. Chien, and W.-S. Wen. A novel approach for event detection by mining spatio-temporal information on microblogs. In *ASONAM*, pages 254–259, 2011.

[13] J. J. Levandoski, M. Sarwat, A. Eldawy, and M. F. Mokbel. Lars: A location-aware recommender system. In *ICDE*, pages 450–461, 2012.

[14] R. Li, S. Wang, and K. C.-C. Chang. Multiple location profiling for users and relationships from social network and content. *PVLDB*, 5(11):1603–1614, 2012.

[15] 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, 2012.

[16] D. Quercia, L. Capra, and J. Crowcroft. The social world of twitter: Topics, geography, and emotions. In *ICWSM*, 2012.

[17] A. Sadilek, H. A. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *WSDM*, pages 723–732, 2012.

[18] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[19] B. Shaw, J. Shea, S. Sinha, and A. Hogue. Learning to rank for spatiotemporal search. In *WSDM*, pages 717–726, 2013.

[20] S. Sizov. Geofolk: latent spatial semantics in web 2.0 social media. In *WSDM*, pages 281–290, 2010.

[21] Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo, and A. Kaltenbrunner. The length of bridge ties: Structural and geographic properties of online social interactions. In *ICWSM*, 2012.

[22] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *ECIR*, pages 356–367, 2013.

[23] Z. Yin, L. Cao, J. Han, C. Zhai, and T. S. Huang. Geographical topic discovery and comparison. In *WWW*, pages 247–256, 2011.