Inferring Crowd-Sourced Venues for Tweets

Bokai Cao*, Francine Chen†, Dhiraj Joshi† and Philip S. Yu*‡
*Department of Computer Science, University of Illinois at Chicago, IL, USA; {caobokai, psyu}@uic.edu
†FX Palo Alto Laboratory, Palo Alto, CA, USA; {chen, dhiraj}@fxpal.com
‡Institute for Data Science, Tsinghua University, Beijing, China

***Abstract*—Knowing the geo-located venue of a tweet can facilitate better understanding of a user's geographic context, allowing apps to more precisely present information, recommend services, and target advertisements. However, due to privacy concerns, few users choose to enable geotagging of their tweets, resulting in a small percentage of tweets being geotagged; furthermore, even if the geo-coordinates are available, the closest venue to the geo-location may be incorrect.**

In this paper, we present a method for providing a ranked list of geo-located venues for a non-geotagged tweet, which simultaneously indicates the venue name and the geo-location at a very fine-grained granularity. In our proposed method for Venue Inference for Tweets (VIT), we construct a heterogeneous social network in order to analyze the embedded social relations, and leverage available but limited geographic data to estimate the geo-located venue of tweets. A single classifier is trained to estimate the probability of a tweet and a geo-located venue being linked, rather than training a separate model for each venue. We examine the performance of four types of social relation features and three types of geographic features embedded in a social network when inferring whether a tweet and a venue are linked, with a best accuracy of over 88%. We use the classifier probability estimates to rank the candidate geo-located venues of a non-geotagged tweet from over 19k possibilities, and observed an average top-5 accuracy of 29%.

***Index Terms*—data mining; venue inference; location-based social networks; Twitter.**

I. Introduction

Over 500 million tweets are generated per day by Twitter users for sharing activities, emotions and opinions. Inferring the location of tweets has emerged to be a critical and interesting issue in social media, since only a small percentage of users chooses to publicize their location when they tweet, and tweets with specific venues associated with them are even sparser. Specifically, it has been estimated that less than 1% of tweets are geotagged [27]. Inferring the location of non-geotagged tweets can facilitate better understanding of users' geographic context, which can enable better inference of a geographic intent in search queries, more appropriate placement of advertisements, and display of information about events, points of interest, and people in the geographic vicinity of the user [14], [8], [12], [13], [19]. Knowledge of tweet locations can also be used for profiling business locations [3].

Prior work on geo-location in social networks can be categorized roughly into two groups: content analysis of tweets, and inference via social relations of users. Depending on the objects being inferred, different studies focus on the home location of users or individual tweets. We have summarized related work on location inference in Table I. Most existing studies infer the location of a user or a tweet at a coarse level of granularity, such as country, state or city levels, which may not be good enough to identify potential recipients for location-driven advertising. Unlike previous work, we attempt to infer the location of a tweet as a *geo-located venue*, exploring both content-based features and features extracted from a user's friendship network to infer both location and venue name.

Despite its value and significance, the problem of inferring geo-located venues for tweets, to the best of our knowledge, has not used social network information for location inference at such a fine granularity. There are two major difficulties in inferring the venue and geo-location where a tweet was posted, as follows:

Noisy geotags: Other than location-based services, *e.g.*, Foursquare, that explicitly let users choose a point of interest/venue for their checkins, many social media applications on mobile devices, *e.g.*, Twitter, provide geotagging in the form of associating a latitude-longitude pair with a tweet. However, geotagging in the form of coordinates is often not very precise, especially within a confined geographic area. For example, it can be ambiguous to determine from geotags whether a tweet was posted at an Apple Store or a Starbucks next door. Hence, simply assigning a tweet to the nearest venue or point of interest can be error-prone. The problem becomes even harder in scenarios where users choose not to use location-based services or where users tweet about food on the way home after they have explored a good restaurant, although it would be desirable to associate such tweets with the restaurant.

Ambiguous text: For non-geotagged tweets, the most explicit information that can be used for location inference is the textual content of tweets, which can be a mix of a variety of daily activities (*e.g.*, food, sports, emotions, opinions) without clear location signals. Tweets are usually short and informal, implying that traditional gazetteer terms may not be present in the vocabulary of the tweets at all. Even if proper place names are contained in tweets, it can still be a tough problem, especially for chain stores. For example, there may not be a significant difference between content of tweets that are associated with the Starbucks at UC Berkeley and the Starbucks at Stanford Univeristy. Therefore, it is not easy to tell based on the content of a tweet from which store branch the tweet was posted.

To address the above problems, this paper presents our attempts to estimate the geo-located venue of tweets. Following

This work was done while the first author was doing internship at FX Palo Alto Laboratory.

TABLE I
SUMMARY OF RELATED WORK IN LOCATION INFERENCE IN SOCIAL NETWORKS.

Methods	Objects	Granularities	Features	Networks
Mahmud et al. [21]	home location	time zone/state/city	textual content+time stamps	Twitter
Wang et al. [27]	home location	province/city	textual content+social relations	Sina-Weibo
Cheng et al. [4]	home location	city	textual content	Twitter
Han et al. [10]	home location	city	textual content	Twitter
Davis et al. [7]	home location	city	social relations	Twitter
Backstrom et al. [1]	home location	coordinates	social relations	Facebook
Compton et al. [6]	home location	coordinates	social relations	Twitter
Sadilek et al. [23]	users	clustered coordinates	textual content+social relations	Twitter
Kinsella et al. [14]	tweets	country/zipcode	textual content	Twitter
Flatow et al. [9]	tweets	hyper-local location	textual content	Twitter
Li and Sun [18]	tweets	venue name	textual content	Twitter+Foursquare
Lee et al. [17]	tweets	geo-located venue	textual content	Twitter+Foursquare
VIT	tweets	geo-located venue	textual content+social relations	Twitter+Foursquare

the approaches of focusing on a geographic region [17], [9], [18], [23], [24], we collected a sample of public tweets that originated from a predefined geographic region. In addition, research on estimating the home location of a person [4], [10] could be used to pre-filter non-geotagged tweets and identify those with a home location of interest so that our method would in theory apply to a larger area, but that is outside the scope of this paper; our focus is on investigating a method of inferring geo-located venues in a metropolitan area where each tweet is assumed to be associated with one of the geo-located venues.

In this paper, we present Venue Inference for Tweets (VIT) that utilizes social, geographic and textual information for ranking geo-located venues as the location of a non-geotagged tweet. The main contributions of this paper include:

- Defining the problem of ranking geo-located venues of mobile social media posts that differentiates among branches of a chain with multiple stores;
- Constructing a heterogeneous social network modeling users' social relations and activities, textual information (Section III-A), and leveraging the geographic context (Section III-B);
- Proposing a single trained model for ranking, rather than the more common approach of creating a separate model for each location, allowing for better generalization to new venues (Section III-C);
- Evaluating the proposed ranking model using datasets collected from Twitter and Foursquare (Section II) on predefined geographic regions and on chain stores for geo-located venue accuracy and geographic distance accuracy, as well as evaluating the utility of the features individually and combined (Section IV).

II. DATASET

In this section, we describe the social media data that we collected from two sources. We also describe how we preprocess this data and create labeled ground truth data for our experiments.

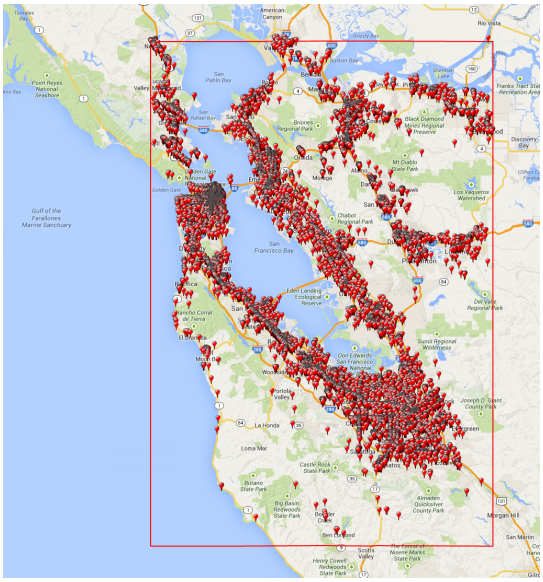


Fig. 1. Spatial distribution of verified venues in the San Francisco Bay Area, collected from Foursquare. The frame indicates the bounding box.

A. Data Collection

Data used in this paper was collected from Twitter and Foursquare. First, we defined a bounding box in terms of latitude and longitude for the San Francisco Bay Area, as shown in Figure 1. Using the geotag filter option of Twitter's streaming API[1], we collected tweets within the bounding box from June 2013 to April 2014. We then invoked the Twitter REST API[2] to collect each user's list of followers and followees. Friendship in Twitter is defined between users who mutually follow each other. In total, we obtained 10,080,973 tweets generated by 251,660 Twitter users, with 3,276,724 friendship links between them.

Using the Foursquare API[3] which allows for access to information from both Foursquare and Swarm, we collected all the non-private venues within the bounding box, then collected all tips associated with each of these venues from

[1]https://dev.twitter.com/docs/api/streaming
[2]https://dev.twitter.com/docs/api/1.1
[3]https://developer.foursquare.com/docs

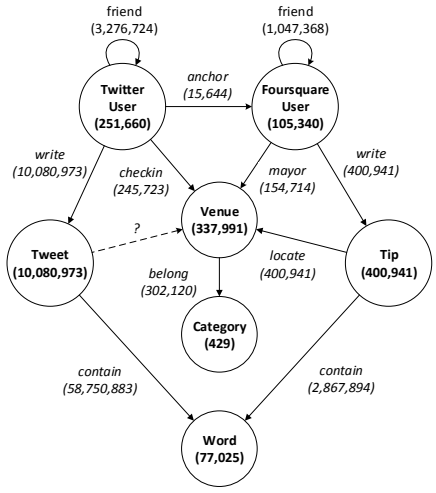


Fig. 2. The network schema of our combined Twitter and Foursquare data. Each circle represents a type of node in the network, and each line represents a type of link. Each number under the node/link represents the total number of nodes/links of the corresponding type in the social network.

February 2009 to June 2014. Moreover, we collected the friendship information between users who posted these tips. Our final dataset contained 400,941 tips generated by 105,340 Foursquare users associated with 84,338 venues, in addition to 253,653 venues without any tips. For our experiments, we considered only those venues verified by the business owners in Foursquare, a total of 19,084, a sample of which is shown in Figure 1. The Foursquare API also provides the corresponding Twitter account of a Foursquare user, if one exists. This information was collected in order to identify the same user across different social networks. Around 15% of the Foursquare users had a Twitter account linked to their Foursquare account. Because of privacy concerns, checkin records are not explicitly available from Foursquare. Instead, we collected the mayorship information, which denotes the user who had the most checkin records at a particular venue. Additionally, we used tweets sourcing from Foursquare as a sample of checkin records that users positively want to share with their friends.

A heterogeneous social network can be constructed based on the multiple types of entities and relationships we collected. We represent each type of entity as a type of node in the network schema, *e.g.*, *users*, *tweets*, *tips*, *venues*. Relationships between these entities can be represented as different types of links, *e.g.*, *write*, *locate*, *anchor links*. The constructed network schema is illustrated in Figure 2.

Note that *words* are also represented as a type of node in the network schema. For text processing, we simply removed stop words using NLTK[4] and filtered out those words appearing in less than ten tweets. A *contain* link is added between a tweet/tip and a content word if the word appears in the tweet/tip.

[4]http://www.nltk.org

B. Data Preprocessing and Labeling

To provide ground truth for training and evaluation, we need the geo-location and venue of each tweet. Providing geo-located venues for non-geotagged tweets, either manually with trained labelers or by crowdsourcing, can be a very difficult task since location-specific knowledge is needed. For example, the first tweet in Table II refers to "Lemondrop", a common type of candy. However, in this case it is the name of an albino python at the California Academy of Sciences. So although "Lemondrop" is a good clue for geo-locations near the California Academy of Sciences, especially for a user who tweets from science museums or has friends who do, many labelers would not be familiar with the python name nor know about each user's interests. Thus, we decided to use geotagged tweets for experiments. The geographic information of a tweet was only used as ground truth for evaluation; it was *not* used as a feature of training or test tweets. Similar use of geotagged tweets for developing and evaluating systems for geo-location inference was adopted by [17], [9], [23] for their experiments.

Tweets with Foursquare as the source were removed from the training and testing data because most are in the format "I'm at <somewhere>", which makes venue inference trivial. However, these Foursquare tweets are represented in the constructed social network as *checkin* links.

TABLE II
EXAMPLES OF GEOTAGGED TWEETS FROM OUR DATASET. THE SOURCE OF EACH TWEET IS INDICATED IN BRACKETS.

t_1	meet Lemondrop y'all @jeremysasson @ California Academy of Sciences [Instagram]
t_2	BEST BURGERS EVER WITH @GraciCarvalho ?? @ Smashburger [Instagram]
t_3	New insurance = Massive headaches at the pharmacy. ? (at @walgreens) [Path]

The remaining geotagged tweets were filtered by a method similar to that described in Chen et al. [3] to identify those tweets with text containing mention of a Foursquare *venue name* geo-located in the neighborhood of a tweet. To allow for shortened names, a tweet containing at least half the content words in a venue name was considered a match. Neighborhood was defined to be a radius of .0008 degrees, or about 290 ft.

The filtered tweets were then processed to remove geo-location information. Similar to Foursquare, several other popular mobile apps (*e.g.*, Instagram, Path) enable users to tag their posts with venue, as in the tweets in Table II. These tags were removed from the tweet text by removing words following "@". Note that in contrast to [3], which assigned *geotagged* tweets to venues, the tweets we use for cross-validation training and testing do not contain geo-location or @mentions, including @venue. For example, after processing, the text of t_1 in Table II would be "meet Lemondrop y'all" and the geo-coordinates are null. The venue information serves only as **ground truth** labels for our experiments.

After filtering, we obtained the actual venue for 126,917 tweets for our cross-validation experiments. We note that

geotagged and non-geotagged tweets may have a different distribution of text and locations. But similarly to Flatow et al. [9] who used geotagged tweets with geo-coordinates removed for their experiments, we believe that use of geotagged tweets with geo-coordinates and @mentions removed can provide some insight into the utility of using social network information for inferring the geo-located venue of a tweet. In addition, as described in Section III, we do not build separate classifiers for each location; instead a single classifier is used, so that the actual distribution of tweet locations in the dataset is somewhat less important.

III. Venue Inference Method

The problem of *venue inference for tweets* can be formalized as: given a non-geotagged tweet t_i and a candidate venue v_p, estimate the tweet's probability of being posted at the venue, $P(y(t_i, v_p) = 1)$. Similarly to the object recognition task where a fixed set of objects are ranked according to how likely each is to be in an image [16], we rank the set of candidate venues by $P(y(t_i, v_p) = 1)$. And also similarly to object recognition, our evaluation will focus on the correct venue being highly ranked.

As we have noted, inferring venues for tweets is a difficult and challenging task. In this paper, we present an approach to solve the problem by analyzing the social activities embedded in our constructed heterogeneous social network and leveraging available but limited geographic data. We investigate the task of inferring the specific geo-located venue a tweet was posted at within a confined geographic region. In this section, we first consider how to exploit information from our constructed network to make venue inference for tweets in Section III-A, assuming no geographic data is available. Next, we explore how to take into account the users' geographic context, and the geotagged tweets of their friends in Section III-B. Our solution to the problem of venue inference for tweets is formalized in Section III-C.

A. Exploiting the Social Network

Before exploiting the correlations among multiple types of entities in the constructed network, we first briefly review the notion of *meta-path* following previous work [2], [15], [26].

In general, a meta-path corresponds to a type of path within the network schema, containing a certain sequence of link types. For example, in Figure 2, a meta-path "tweet $\xrightarrow{contain}$ word $\xrightarrow{contain^{-1}}$ tip $\xrightarrow{locate}$ venue" denotes a composite relationship from tweets to venues, where $contain^{-1}$ represents the inverted relation of *contain*. The semantic meaning of this meta-path is that the tweet and the venue share common content words via tips.

Different meta-paths usually represent different relationships among linked nodes with different semantic meanings. For example, the meta-path "tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{anchor}$ Foursquare user $\xrightarrow{mayor}$ venue" denotes that the tweet was posted by a Twitter user who is a mayor of the venue in Foursquare, while the meta-path "tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{friend}$ Twitter user $\xrightarrow{checkin}$ venue" indicates the tweet was posted by a Twitter user whose friends checked in at the venue. In this way, relationships between tweets and venues can be described by different meta-paths with different semantics. Thus, we extract four types of meta-paths from our constructed network, as follows:

Ego Path directly relates a user's tweets to venues. Given a tweet-venue pair, say, (t_i, v_p), we denote the user who posted the tweet t_i as u_i. To infer the existence probability of the link (t_i, v_p), *i.e.*, $P(y(t_i, v_p) = 1)$, it is crucial to know if u_i has any type of direct interactions with the venue v_p, *e.g.*, *check in at*, *writing a tip about*, *being a mayor of*, which we refer to as *social activities*. As described above, the meta-path "tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{anchor}$ Foursquare user $\xrightarrow{mayor}$ venue" can detect if t_i was posted by u_i who is a mayor of v_p in Foursquare. Obviously, t_i should be more likely to be associated with the venue v_p if there exists such a meta-path from t_i to v_p than those venues without such connections. Similarly, we extract other meta-paths, denoted as EgoPath, to capture the correlations between t_i and v_p via u_i, as summarized in Table III.

Friend Path relates a user's tweets to venues through their friends. Although EgoPath is expected to be important for representing the correlations between t_i and v_p by leveraging explicit social activities of u_i across Twitter and Foursquare, we observe that only a small number of tweets can be inferred in this way, which is especially hard for the users who do not have linked Foursquare accounts. It was observed in [5] that social relationships can explain about 10% to 30% of all human movement. Inspired by the idea of the homophily principle in social science [22], [20], in addition to looking at the social activities of u_i, we can also exploit the activities of u_i's friends. We consider that if a friend u_j has any social activities at the venue v_p, the user u_i is more likely to post the tweet t_i at v_p than those venues without such connections. For example, the meta-path "tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{friend}$ Twitter user $\xrightarrow{checkin}$ venue" can tell whether any friends of u_i have tweeted Foursquare checkins at the venue v_p. We denote the meta-paths leveraging friends' information as FriendPath, as summarized in Table III.

Interest Path expands the relationship between tweets and venues through Foursquare categories. Taking into consideration the user interests, we assume that users tend to tweet at similar venues (based on their interests). For example, suppose v_p is *Chef Chu's* in Los Altos, v_q is *Cooking Papa* in Mountain View, and both of them belong to the category *Chinese restaurant*. Intuitively, if the user u_i has checkins at v_q, indicating s/he likes Chinese food, then t_i is more likely to be posted by u_i at v_p than those venues without such connections. In our collected data from Foursquare, each venue is associated with one of the 429 categories[5], as illustrated by the link type

[5]A few venues have a category value of null.

TABLE III
EXAMPLES OF META-PATHS USED IN OUR METHOD.

Types	Meta-paths
EGOPATH	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{checkin}$ venue
	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{anchor}$ Foursquare user $\xrightarrow{mayor}$ venue
	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{anchor}$ Foursquare user $\xrightarrow{write}$ tip $\xrightarrow{locate}$ venue
FRIENDPATH	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{friend}$ Twitter user $\xrightarrow{checkin}$ venue
	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{friend}$ Twitter user $\xrightarrow{anchor}$ Foursquare user $\xrightarrow{mayor}$ venue
	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{anchor}$ Foursquare user $\xrightarrow{friend}$ Foursquare user $\xrightarrow{mayor}$ venue
	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{friend}$ Twitter user $\xrightarrow{anchor}$ Foursquare user $\xrightarrow{write}$ tip $\xrightarrow{locate}$ venue
	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{anchor}$ Foursquare user $\xrightarrow{friend}$ Foursquare user $\xrightarrow{write}$ tip $\xrightarrow{locate}$ venue
INTERESTPATH	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{checkin}$ venue $\xrightarrow{belong}$ category $\xrightarrow{belong^{-1}}$ venue
	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{anchor}$ Foursquare user $\xrightarrow{mayor}$ venue $\xrightarrow{belong}$ category $\xrightarrow{belong^{-1}}$ venue
	tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{anchor}$ Foursquare user $\xrightarrow{write}$ tip $\xrightarrow{locate}$ venue $\xrightarrow{belong}$ category $\xrightarrow{belong^{-1}}$ venue
TEXTPATH	tweet $\xrightarrow{contain}$ word $\xrightarrow{contain^{-1}}$ tip $\xrightarrow{locate}$ venue

belong in Figure 2. The meta-path "tweet $\xrightarrow{write^{-1}}$ Twitter user $\xrightarrow{checkin}$ venue $\xrightarrow{belong}$ category $\xrightarrow{belong^{-1}}$ venue" can effectively detect whether t_i was posted by a user who has checkins at venues sharing the same category as v_p. This type of meta-path is denoted as INTERESTPATH in Table III.

Text Path models the content words tweeted about venues. Unlike conventional approaches that focus on text processing for content analysis [18], [11], [4], [21], in this paper, we represent content words as a node type in our constructed network schema, as shown in Figure 2. Following the idea of meta-path, we can also define a meta-path via *word* to capture textual similarity between tweets and venues. The meta-path "tweet $\xrightarrow{contain}$ word $\xrightarrow{contain^{-1}}$ tip $\xrightarrow{locate}$ venue", denoted as TEXTPATH, can encode whether the tweet t_i and the venue v_p share common content words via tips. Words were used, following [17] which observed better performance of Foursquare unigrams over bigrams for location inference. We consider that t_i should be more likely to be associated with v_p sharing similar textual content than those venues without such correlations.

The path counts of these meta-paths are computed and used as elements of the feature vectors input to the classifier.

B. Leveraging the Geographic Context

In this part, we introduce how to make use of available geographic information contained in geotagged tweets.

Ego Geo: Although a tweet t_i posted by user u_i might be non-geotagged (or its geographic information was withheld for experiments), other tweets posted by u_i might be geotagged with their geo-coordinates available. To capture such information, let T_i denote the set of geotagged tweets posted by u_i. We define a geographic score between t_i and a candidate venue v_p as follows:

$$\text{EGOGEO}(t_i, v_p) = -\log\big(\min_{t_j \in T_i - t_i} d(t_j, v_p) + \epsilon\big)$$

where $d(\cdot,\cdot)$ denotes the distance between the geo-coordinates of a tweet and a venue, and ϵ is added to avoid underflow with a default value 10^{-9}. The formulation is to measure the closest distance between geotagged tweets of the user who posted t_i and a candidate venue v_p. Intuitively, t_i is more likely to be associated with v_p, if u_i has posted any geotagged tweet in the neighborhood of v_p. Therefore, the higher value of $\text{EGOGEO}(t_i, v_p)$, the higher existence probability of the link (t_i, v_p), *i.e.*, $P(y(t_i, v_p) = 1) \propto \text{EGOGEO}(t_i, v_p)$.

Friend Geo: In scenarios where users come to a new place and tweet, EGOGEO may not work. However, considering people usually hang out with friends and may tweet at some interesting places together, we further propose a measure based on a user's friends' geotagged tweets:

$$\text{FRIENDGEO}(t_i, v_p) = -\log\big(\min_{t_j \in T_k, u_k \in N_i} d(t_j, v_p) + \epsilon\big)$$

where N_i is the set of users who are friends of u_i, and T_k are the tweets by u_k. The formulation is to measure the closest distance between geotagged tweets of u_i's friends and a candidate venue v_p. Therefore, if u_i's friends have posted any geotagged tweet in the neighborhood of v_p, t_i is more likely to be associated with v_p than venues without such correlations. Therefore, we can say that the existence probability of the link (t_i, v_p) is likely to be positively correlated with $\text{FRIENDGEO}(t_i, v_p)$, *i.e.*, $P(y(t_i, v_p) = 1) \propto \text{FRIENDGEO}(t_i, v_p)$.

C. Ranking Geo-located Venues

Based on features extracted from our constructed social network and available geographic data, we can now estimate the existence probability of any given link between tweets

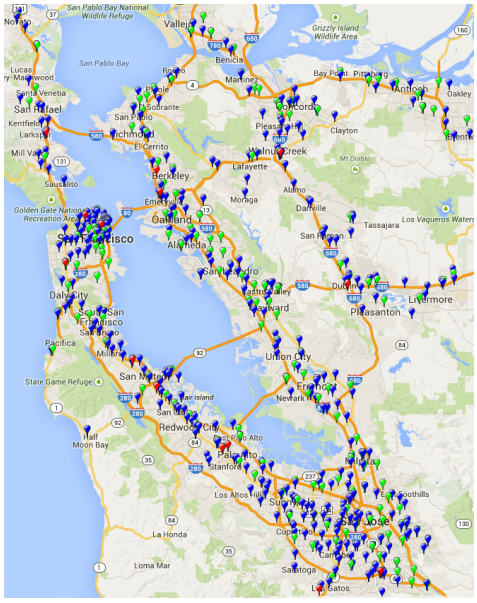


Fig. 3. Spatial distribution of Starbucks (blue pins), McDonald's (green pins), and Apple Stores (red pins) in the San Francisco Bay Area. Some stores are not visible because of overlapping marks.

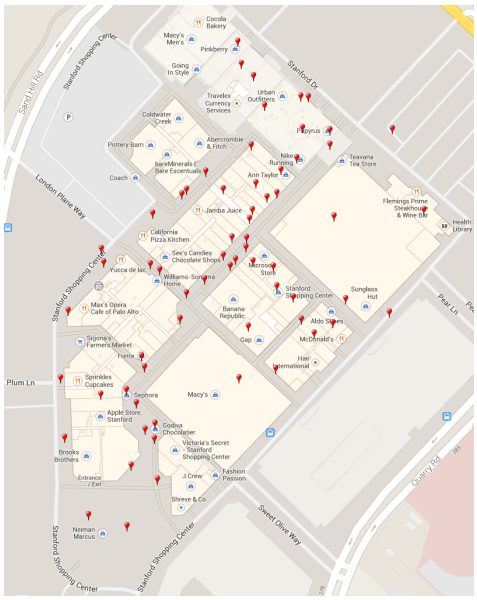


Fig. 4. Spatial distribution of verified venues in the Stanford Shopping Center, collected from Foursquare.

and venues. That is to say, we can estimate $P(y(t_i, v_p) = 1)$, the probability of t_i being posted at v_p. Then we consider that, given a tweet t_i, how its actual venue $v_{act}(t_i)$ *ranks* when venues are ordered by $P(y(t_i, v_p) = 1)$. To compute $P(y(t_i, v_p) = 1)$, we use a Support Vector Machine (SVM) (SCIKIT-LEARN[6]) with a linear kernel and default parameters as the classifier. Probability estimates are enabled as output and features are defined in Sections III-A and III-B.

A common use of SVMs is to train a separate one-against-all model for each class, as the ranking step in [17]. In our task, however, this would require training a separate SVM for each geo-located venue. Since the features of VIT implicitly encode a tweet-venue pair, one SVM model is trained to classify whether the link between a tweet and a venue is positive or negative. This approach has the added advantage of generalization to new venues.

Our focus is to estimate the probability of a venue being the location of a tweet. For this, we consider a closed-class task of inferring a tweet venue from a set of venues, such as the 19,084 venues described in Sections II-A and II-B as ground truth. Following the principle of maximum likelihood estimation, an intuitive idea is to compute $P(y(t_i, v_p) = 1), \forall v_p \in V$ for each tweet t_i, where V is the set of candidate venues, and the v_p with the maximum probability $P(y(t_i, v_p) = 1)$ is our estimated venue $v_{est}(t_i)$. In this way, the size of V, *i.e.*, $|V|$, would influence the efficiency of the inference process, if we enumerate all the venues in V. We examine whether efficiency can be improved by sampling the most relevant venues, for example, venues connected to t_i via FRIENDPATH, TEXTPATH, *etc.*

In this paper, we primarily investigate application scenarios inferring which store a tweet was posted at, from multiple geo-located venues of a chain of stores (*e.g.*, Starbucks, McDonald's, Apple Stores, as shown in Figure 3). We are also interested in determining which specific venue a tweet was posted at within a confined geographic area (*e.g.*, the San Francisco Bay Area, as shown in Figure 1, or the Stanford Shopping Center, as shown in Figure 4). In the former, the general area may be estimated using home location inference methods, and in the latter, a user may tweet that they are at the Stanford Shopping Center. These problems are very challenging due to similar topics shared by tweets at different Starbucks stores and the close proximity of different venues located in a shopping mall.

IV. EXPERIMENTS

We now introduce the experimental setup in Section IV-A, analyze experimental results in Section IV-B, and investigate the usefulness of different features in Section IV-C.

A. Experimental Setup

Experiments were conducted in the setting of 10-fold cross-validation. In each fold of the training data, we respectively sampled half of the known links between tweets and venues as positive links. For links in the other half, say (t_i, v_p), a venue v_q was randomly generated from $V - v_p$ to form a negative link (t_i, v_q). In this way, a balanced dataset was derived for the training process, containing the same number of positive links and negative links. This experimental setting will be justified in Section IV-C.

To evaluate the quality of a venue estimator, we compare the inferred venue of a tweet versus the actual venue. The first metric we consider is **ErrDist** which quantifies the distance in miles between geo-coordinates of the actual venue and the inferred venue. It is defined as follows [4]:

$$\textbf{ErrDist} = \frac{\sum_{t_i \in T} d(v_{act}(t_i), v_{est}(t_i))}{|T|}$$

[6]http://scikit-learn.org

where T is the set of test tweets.

A low **ErrDist** means that the model can geo-locate tweets close to their actual venue, but it can not directly provide us with a strong intuition about the distribution of venue inference errors. Therefore, **Accuracy** is considered to measure the percentage of tweets with their inferred venue correctly matched with the actual venue:

$$\textbf{Accuracy} = \frac{\sum_{t_i \in T} I(v_{act}(t_i), \{v_{est}(t_i)\})}{|T|}$$

where identity function $I(a, S) = \begin{cases} 1, & if\ a \in S \\ 0, & otherwise \end{cases}$ can check whether the actual venue can be matched within the set of inferred venues.

Since the venue estimator gives k venues for each tweet in decreasing order of confidence, we denote the **ErrDist** with k candidate venues as **ErrDist**@k, which applies the same **ErrDist** metric over inferred venues in the top-k and chooses the smallest error distance to the actual venue:

$$\textbf{ErrDist@}k = \frac{\sum_{t_i \in T} \min_{j=1 \cdots k} d(v_{act}(t_i), v_{est_j}(t_i))}{|T|}$$

where $v_{est_j}(t_i)$ is the j-th venue inferred for t_i in decreasing order of confidence. Similarly, we define the **Accuracy** with k candidate venues as **Accuracy**@k:

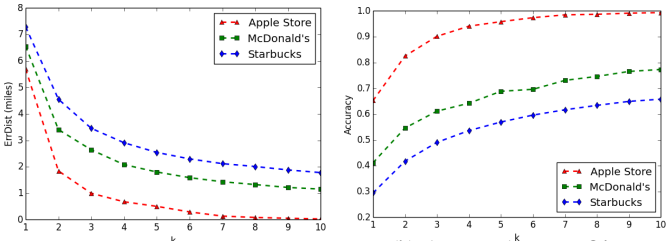
$$\textbf{Accuracy@}k = \frac{\sum_{t_i \in T} I(v_{act}(t_i), \cup_{j=1}^{k} \{v_{est_j}(t_i)\})}{|T|}$$

In this way, the metrics show the capacity of an estimator to infer a good candidate venue, even if the first candidate is in error.

B. Experimental Results

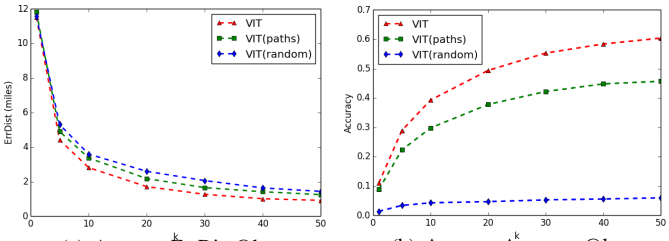
First, we are interested in inferring which specific branch of a store a tweet was posted at, from multiple venues of a chain store distributed over the San Francisco Bay Area. Such inference is useful for business analysis of chain stores. For example, inferring that a tweet was posted at the Starbucks at Berkeley or the Starbucks at Stanford can facilitate better understanding of user purchasing behavior at different campuses, or deciding on whether to conduct a campus promotion at Berkeley and/or Stanford.

Three chain stores are examined: Starbucks, McDonald's, and Apple Stores. As visualized on the Google Map in Figure 3, the numbers of verified venues of Starbucks, McDonald's, and Apple Stores are 409, 184, and 14, respectively, in our data collection area. Figure 5 illustrates the performance on inferring geo-located venues for tweets associated with these chain stores. It indicates that VIT can locate a branch store in the top-10 candidate venues within 2 miles around the actual venue for these three chains. We notice that the performance on Apple Stores is the best, because the problem difficulty decreases with fewer candidate venues. Similarly, VIT can correctly identify the actual venue in the top-3 candidate venues for almost 90% of the tweets about Apple Stores, and the **Accuracy**@10 for Starbucks and McDonald's is 66% and 78%, respectively.



(a) Average ErrDist@k. (b) Average Accuracy@k.

Fig. 5. Performance on inferring geo-located venues for tweets associated with Starbucks, McDonald's, and Apple Stores.



(a) Average ErrDist@k. (b) Average Accuracy@k.

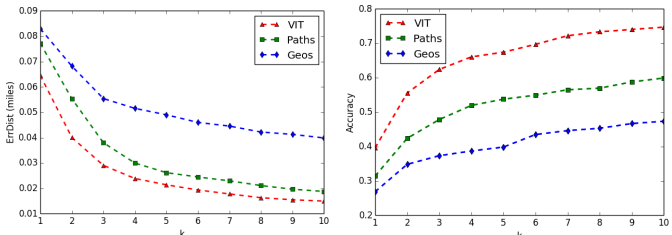
Fig. 6. Performance on using different strategies to rank over 19k geo-located venues.

Next, we investigate three strategies to enumerate geo-located venues for a tweet:

- VIT enumerates all the candidate venues;
- VIT($paths$) only enumerates the venues connected to a tweet through meta-paths as defined in Table III;
- VIT($random$) randomly samples the same number of venues as VIT($paths$) for each tweet.

Figure 6(a) indicates that by enumerating all candidate venues, VIT can locate a venue in the top-20 candidate venues within 2 miles around the actual venue; Figure 6(b) shows that the actual venue can be correctly identified in the top-20 candidate venues by VIT for almost 50% of the tweets. Note that being able to infer the actual venue among the top-20 is actually a very challenging task (considering there are a total of 19,084 candidate venues). By leveraging the meta-paths in Table III, VIT($paths$) achieves comparable results with VIT on **ErrDist**@k, and can infer the actual venue in the top-20 candidate venues for 40% of the tweets. The average number of venues to be enumerated for each tweet in VIT($paths$) is 1,571 in our dataset, which is an order of magnitude less than VIT. It reveals a trade-off for VIT between accuracy and efficiency, in terms of enumerating candidate venues. The enumeration process can be facilitated for most tweets, since the actual venue associated with a tweet is usually related to the user's social activities embedded in our constructed network. This can be further validated by VIT($paths$) significantly outperforming VIT($random$) on **Accuracy**@k.

We further consider how to infer which venue a tweet was posted at within a confined geographic area, *e.g.*, users may have tweeted that they are at the Stanford Shopping Center. As shown in Figure 4, there are 65 different venues located in the Stanford Shopping Center, including Starbucks, Apple Store, Macy's, *etc.* Compared with country-level or city-level



(a) Average ErrDist@k. (b) Average Accuracy@k.

Fig. 7. Performance on inferring geo-located venues in the Stanford Shopping Center.

location inference as studied in [11], [4], [14], such fine-grained venue inference in a shopping mall is very challenging due to the close proximity between stores.

Figure 7 shows the performance on inferring geo-located venues in the Stanford Shopping Center. We observe that VIT can correctly infer the actual venue in the top-10 candidate venues for 74% of the tweets in the Stanford Shopping Center. Figure 7 also includes results when only meta-path based features (PATHS) or features based on geographic data (GEOS) are used. Since venues are inferred within such a small-scale area, GEOS plays a less important role than PATHS in this task, which implies the importance of exploiting our constructed social network for venue inference. A detailed feature analysis is presented in Section IV-C.

C. Feature Analysis

In the experiments, we used a balanced dataset to train a classifier for link prediction, *i.e.*, inferring whether there exists a positive/negative link given a tweet-venue pair. We justify this experimental setting by varying the number of total exemplars and the number of positive exemplars used in training. For a given number of total exemplars, the set of test tweets was fixed.

Observing the red lines in Figure 8 where the training set contains 100 positive exemplars each, the difference in performance is relatively small as the number of total exemplars is varied. In contrast, the lines with round markers all have 100 total exemplars, and the performance varies quite a bit depending on the percentage of positive exemplars; and similarly for the cases with 500 total exemplars, indicated with triangular markers. Thus, we infer that it is better to train with 50% positive exemplars (indicted with "*" in the legend in Figure 8), which provides performance close to that of training on a larger dataset with the same number of positive exemplars but the true ratio of positive to negative exemplars. That is, if we only have 50 positive exemplars and the true positive/total ratio is 0.1, it's better to train on 50 positive and 50 negative exemplars (VIT_100_50), which provides performance close to training on 50 positive exemplars and a positive/total ratio of 0.1 (VIT_500_50), and much better than if we train on 100 exemplars with a positive/total ratio of 0.1 (VIT_100_10). Note that each test tweet in Figure 6 is compared against the 19084 candidate venues, so that the actual positive/total ratio is 1/19084; there would be very few positive exemplars if the true positive/total ratio was used.

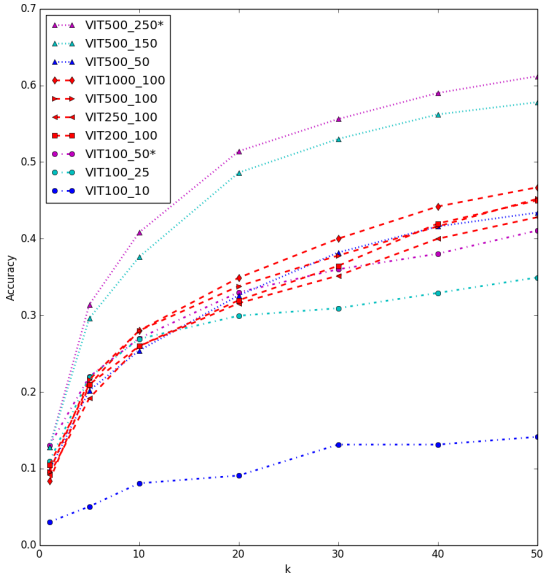


Fig. 8. Link prediction performance with different numbers of total and positive exemplars used for training VIT. Legend format: VIT[total]_[positive].

TABLE IV
LINK PREDICTION PERFORMANCE WITH DIFFERENT TYPE OF FEATURES USED IN VIT.

Features	Evaluations			
	Accuracy	Precision	Recall	F1
EGOPATH	0.5413	0.9981	0.0827	0.1527
FRIENDPATH	0.5694	0.9829	0.1413	0.2470
INTERESTPATH	0.5414	0.8600	0.0990	0.1775
TEXTPATH	0.7367	0.8998	0.5327	0.6692
PATHS	0.7731	0.9195	0.5998	0.7247
EGOGEO	0.8175	0.8644	0.7531	0.8049
FRIENDGEO	0.6862	0.7466	0.5638	0.6424
GEOS	0.8297	0.8715	0.7734	0.8195
VIT	0.8875	0.9327	0.8352	0.8813

To analyze the discriminative power of different types of features in the setting of link prediction, in addition to a balanced training dataset, we sampled the same amount of negative links for test data. Random guessing can therefore be regarded as a naive baseline with 50% accuracy on predicting whether a link exists. Feature analysis results are shown in Table IV, where performance is evaluated by accuracy, precision, recall, and F1-score.

EGOPATH is useful only when a tweet was posted at a venue which is exactly the same place where the user has other social activities, *e.g.*, *check in at*, *writing a tip about*, *being a mayor of*. As can be observed in Table IV, EGOPATH achieves a very high precision but a very low recall. This is reasonable because a tweet-venue pair would be inferred as positive if and only if the corresponding EGOPATH could be found, which is very sparse in the network.

FRIENDPATH achieves a higher recall but a lower precision than EGOPATH. By leveraging the social activities of a user's friends in Twitter or Foursquare, FRIENDPATH can detect correlations between tweets and venues in more cases, which

however are not as confident as those by EGOPATH.

INTERESTPATH performs comparably with EGOPATH and FRIENDPATH by taking into account user interests. It indicates that users tend to tweet at venues sharing the same category as their social activities.

TEXTPATH is used in our method to encode textual similarity between tweets and venues by matching the common content words in tweets and venue-related tips. By using this single meta-path, 74% of the tweet-venue pairs can be accurately classified. It validates that text is an important feature for location inference, and users implicitly reveal location information in the content of their tweets, with or without realizing it [11].

PATHS combines four types of meta-path based features and achieves a significant improvement over any single feature type on both accuracy and F1-score. It demonstrates the effectiveness of exploiting multiple types of meta-paths embedded in our constructed social network.

EGOGEO can exploit the distance between a user's geotagged tweets and a candidate venue when geographic data is available for some of the user's tweets. Its 82% accuracy indicates that users tend to tweet in the neighborhood of venues where they have tweeted before.

FRIENDGEO leverages the geographic information of a user's friends which especially compensates in the scenario where the user him/herself is not geo-active (*i.e.*, without any geotagged tweets). In agreement with earlier work [23], [25], our results also indicate that each Twitter user can be regarded as a sensor to estimate their friends' locations.

GEOS concatenates two types of features based on geographic data and performs quite well on inferring geo-located venues for non-geotagged tweets.

Table IV shows that by combining PATHS and GEOS, VIT outperforms any single type of features, and can achieve very good performance with an accuracy of almost 89%. Our results demonstrate that for the problem of venue inference for tweets, it is important to analyze the social relations embedded in our constructed heterogeneous social network and to leverage available geographic data simultaneously.

V. RELATED WORK

To our knowledge, this paper is the first effort in exploring whether social network information is useful for inferring the geo-located venue where a tweet was posted.

There is extensive research on estimating tweet location and user location from tweets. Prior work relevant to this topic can be roughly categorized into two groups based on the techniques used in geo-locating: inference via content analysis of tweets, and analysis of users' social relations. A number of works have focused on content analysis-based methods for inferring a user's home location, *e.g.*, Mahmud et al. [21], Cheng et al. [4] and Han et al. [10]. Methods for inferring home location can be used to pre-filter tweets to identify the general area of the tweet. Our interest is in inferring the fine-grained geo-location and venue of a tweet, given a general location, such as a home location.

Other works estimate the geo-location of *tweets* based on tweet content. Kinsella et al. ranked locations according to the probability that a tweet was generated by the proposed language model for a location. Experiments were conducted at different levels of granularity ranging from the zip code to the country level [14]. Li and Sun extracted venue names mentioned in tweets [18], and Flatow et al. used an iterative cluster-refinement-based method for identifying terms associated with a "hyper-local-geographic area" [9]. However, both of these models cannot distinguish different geo-located venues with the same name (*e.g.*, all Starbucks are treated as the same venue). Lee et al. created a language model from Foursquare tips and inferred geo-located venues only for locations that have at least 50 tips [17]. They did not utilize the information embedded in social relationships between Twitter and Foursquare users that our experiments indicate improved performance when inferring geo-located venues for all locations. In contrast to [9], [18], [17] which create a separate model for each target location, a single model estimating link probability for a tweet-venue pair is created in our proposed approach. In addition, we seek to differentiate among all the stores of a chain.

A number of works focus on geo-location of a user based on followee-follower relationships. Davis et al. presented a simple majority voting scheme to infer the location of a user based on the location of his/her friends at city-level [7]. Backstrom et al. introduced an algorithm that infers the location of an individual from a sparse set of located users by leveraging the probability of friendship as a function of distance [1]. Wang et al. proposed a location inference model for microblog users by maximizing the weighted location from their tweet content and friends, and conducted experiments for city-level and province-level geo-location [27]. Here we address a different task, which is to recover the venue and geo-location of a user (or a tweet) at the time the tweet was posted.

Some studies have attempted to predict friendship links and location links in location-based social networks. Cho et al. stated that social relationships can explain about 10% to 30% of all human movement [5]. Sadilek et al. proposed a probabilistic model based on a Markov random field to infer social ties by considering patterns in friendship formation, tweet content and user location, and implemented a dynamic Bayesian network to infer the most likely location of a user at any time [23]. Zhang et al. presented an iterative approach to collectively predict social links and location links by transferring information across aligned source networks [28]. In contrast, our method ranks specific geo-located venues for inferring where a tweet was posted.

As summarized in Table I, unlike previous works, we represent heterogeneous entities from Twitter and Foursquare in a network, and analyze tweet content and social activities embedded therein. We model locations with individual tweets and infer geo-located venues as the granularity, which is much more fine-grained than country-level or city-level geo-location. Our proposed method ranks specific geo-located venues for a tweet, which simultaneously indicates the geo-location at

a very fine-grained granularity and the venue name that is associated with the tweet.

VI. Conclusions and Future Work

In contrast to previous works on geo-location inference, we present an initial investigation on utilizing information in a heterogenous social network to infer the venue as well as the geo-location of where a tweet was posted. This has potential applications in presenting information, recommending services, targeting advertisements and business analysis at the level of geo-located venues. We showed that by analyzing social activities embedded in our constructed heterogeneous social network and leveraging the geographic context, performance was improved over a simple text-only model. Furthermore, our proposed method, VIT, exhibited an average top-5 accuracy of 29% in inferring from which of over 19k possible geo-located venues the tweet originated.

Rather than training a separate model for each venue, we train one model to estimate the probability of a tweet and a venue being linked. For our experiments, we used this probability in a closed-class model to rank the candidate geo-located venues. To handle an open set of venues in the future, the method described in [17] of filtering tweets based on Foursquare keywords and POS tagging to identify those discussing a location in present tense could be used. This increases precision but the effect on recall was not presented; in addition, only locations with many Foursquare tips can be considered. Alternatively, the estimated probability could be thresholded, as has been done for other ranking systems, such as visual object recognition.

Our initial model can be extended in several other ways. More sophisticated text-based and language models of tweets have been found to be useful for geotagging, *e.g.*, [18], [9], especially for non-chain stores; we would like to examine enhancing our Foursquare-based text paths with such models. Another potential extension is to consider temporal information. For example, by exploring co-location of friends at the time a tweet was posted, the tweet is likely to be associated with the venues of nearby locations of the user's friends.

Recently, Twitter revealed plans to work with Foursquare to allow users to tag their precise locations in a tweet[7]. Through this potential integration, we can expect more geotagged tweets for our model training which can greatly alleviate the problem of data sparsity. However, given such added functionality, some users will still not turn on their geotagging, in which case our proposed technique is needed. On the other hand, this cooperation also reflects the commercial importance of inferring geo-located venues for tweets.

[7] http://www.valuewalk.com/2015/03/twitter-foursquares-location-tagging

References

[1] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *WWW*, pages 61–70. ACM, 2010.

[2] Bokai Cao, Xiangnan Kong, and Philip S Yu. Collective prediction of multiple types of links in heterogeneous information networks. In *ICDM*, pages 50–59. IEEE, 2014.

[3] Francine Chen, Dhiraj Joshi, Yasuhide Miura, and Tomoko Ohkuma. Social media-based profiling of business locations. In *GeoMM*. ACM, 2014.

[4] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating Twitter users. In *CIKM*, pages 759–768. ACM, 2010.

[5] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082–1090. ACM, 2011.

[6] Ryan Compton, David Jurgens, and David Allen. Geotagging one hundred million twitter accounts with total variation minimization. In *IEEE BigData*, pages 393–401. IEEE, 2014.

[7] Clodoveu A Davis Jr, Gisele L Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L Arcanjo. Inferring the location of Twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.

[8] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *EMNLP*, pages 1277–1287. ACL, 2010.

[9] David Flatow, Mor Naaman, Ke Eddie Xie, Yana Volkovich, and Yaron Kanza. On the accuracy of hyper-local geotagging of social media content. *arXiv preprint arXiv:1409.1461*, 2014.

[10] Bo Han, Paul Cook, and Timothy Baldwin. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

[11] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *CHI*, pages 237–246. ACM, 2011.

[12] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J Smola, and Kostas Tsioutsiouliklis. Discovering geographical topics in the Twitter stream. In *WWW*, pages 769–778. ACM, 2012.

[13] Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. Location inference using microblog messages. In *WWW(Companion)*, pages 687–690. ACM, 2012.

[14] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. I'm eating a sandwich in Glasgow: modeling locations with tweets. In *SMUC*, pages 61–68. ACM, 2011.

[15] Xiangnan Kong, Bokai Cao, and Philip S Yu. Multi-label classification by mining label and instance correlations from heterogeneous information networks. In *KDD*, pages 614–622. ACM, 2013.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] Kisung Lee, Raghu K Ganti, Mudhakar Srivatsa, and Ling Liu. When Twitter meets Foursquare: tweet location prediction using Foursquare. In *WWW*, pages 198–207, 2014.

[18] Chenliang Li and Aixin Sun. Fine-grained location extraction from tweets with temporal awareness. In *SIGIR*, pages 43–52. ACM, 2014.

[19] Wen Li, Pavel Serdyukov, Arjen P de Vries, Carsten Eickhoff, and Martha Larson. The where in the tweet. In *CIKM*, pages 2473–2476. ACM, 2011.

[20] Nan Lin. Social networks and status attainment. *Annual review of sociology*, pages 467–487, 1999.

[21] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? Inferring home locations of Twitter users. In *ICWSM*, 2012.

[22] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[23] Adam Sadilek, Henry Kautz, and Jeffrey P Bigham. Finding your friends and following them to where you are. In *WSDM*, pages 723–732. ACM, 2012.

[24] Adam Sadilek, Henry A Kautz, and Vincent Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *AAAI*, 2012.

[25] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860. ACM, 2010.

[26] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In *VLDB*, 2011.

[27] Xia Wang, Ming Xu, Yizhi Ren, Jian Xu, Haiping Zhang, and Ning Zheng. A location inferring model based on tweets and bilateral follow friends. *Journal of Computers*, 9(2):315–321, 2014.

[28] Jiawei Zhang, Xiangnan Kong, and Philip S Yu. Transferring heterogeneous links across location-based social networks. In *WSDM*, pages 303–312. ACM, 2014.