Percimo: A Personalized Community Model for Location Estimation in Social Media

Guangchao Yuan, Pradeep K. Murukannaiah, and Munindar P. Singh *Fellow, IEEE*
Department of Computer Science, North Carolina State University, Raleigh, NC 27695-8206, USA
gyuan@ncsu.edu, pmuruka@ncsu.edu, m.singh@ieee.org

***Abstract*—User location is crucial in understanding the dynamics of user activities, especially in relating their online and offline aspects. However, users' social media activities, such as tweets sent, do not always reveal their location. We consider the problem of estimating geo-tags for tweets and develop a comprehensive approach that incorporates textual content, the user's personalized behavior, and the user's social relationships. Our approach, Percimo, considers the two major kinds of communal attachment, which have distinct computational ramifications.**

We evaluate Percimo via three geo-social graphs based on the mutual-follow relationships of Twitter users, their geographical distance (computed from their geo-tagged tweets), and their preferences for location categories (collected from Foursquare). We find that Percimo yields a smaller prediction error than the two state-of-the-art approaches we compare with.

I. Introduction

With the increasing prevalence of location sharing services on social media, understanding the location from which online content originates is valuable. It helps in characterizing the interplay between a user's online and offline activities [1]. We define a *geo-tag* as a representation of location, e.g., city, neighborhood, or latitude-longitude (lat-lon) coordinate. Geo-tagged messages provide meaningful real-time information for monitoring regional health [2], detecting local emergencies [3], observing linguistic differences across regions [4], and so on.

We address the problem of *location estimation* of user messages. We focus on tweets because of their prominence in social media and popularity over mobile devices: thus, our problem is one of determining a tweet's location or geo-tag. Although a GPS-enabled phone can geo-tag outgoing tweets, only about 2% of tweets are geo-tagged [5]. Thus, a vast majority of tweets can be geo-tagged via location estimation.

Most approaches for location estimation focus on identifying spatial usage of words or phrases in content [6], [7], [8] (content-based techniques). They assume that tweets encode location via place names or other location words and rely on word distributions over geo-tags. Although these approaches can identify words that are associated with a certain location from a global perspective (e.g., "rockets" is used frequently near Houston, the home of NASA), they do not consider that some words may be related to some locations from an individual's perspective. For example, if a user likes reading political news and tweeting them at home, then words related to political news are associated with her home location, although these words are not location related from the broader perspectives of meaning or usage.

Individualized techniques infer geo-tags by assuming that a user's textual content is correlated to her locations. Chen et al. [9] map a tweet's content to its sender's interests and associate interests with locations: a user who tweets from one museum may tweet similar content from another museum. However, their approach fails for users having insufficient historical geo-tagged tweets. Most users have sparse geo-tagged histories and some have no geo-tagged messages at all.

We are interested in solving the location estimation problem by exploiting the correlation between content and locations, not only from an individual perspective, but also from multiple users' perspectives. Grabovitch-Zuyev et al. [10] examine the correlation between users' tweets and their locations via statistical tests, and find that users who send tweets from nearby locations are correlated with users who are similar in textual content. Further, motivated by a basic assumption in recommender systems—friends share more interests than non-friends [11]—we posit that users who are friends are also similar in content (interests). In addition to the importance of friends in estimating a user's location [12], we posit that the user's content is also correlated with her friends' locations. Therefore, the challenge lies in (1) how to select other users that might be geographically or socially related to a user? (2) how to exploit the correlation between content and locations to estimate the location of a user's tweet? That is, *how are one's interests related to another's locations*?

We propose *Percimo—Personalized Community Model*—for solving the location estimation problem. Percimo contrasts with prior work in two ways. First, Percimo considers not only content and individual interests, but also how users' interests are correlated with locations of geo-socially related users. Second, inspired from social psychology regarding attachment to a community, Percimo employs *geo-social communities* to estimate locations of tweets.

Prentice et al.'s theory [13] posits that a user may attach to a community via a combination of *common bonds*—attachment to specific members of the community and *common identity*—attachment to the identity of the community, independently of its members. Sassenberg [14] validates Prentice et al.'s theory empirically for online behavior by associating participation in "on-topic" and "off-topic" chats, respectively, with common identity and bonding. A topic serves as a seed for communal identity independent of who else is interested in that topic.

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00 © 2015 IEEE

When there is no fixed topic, the participants relate more to the other participants: the communal identity is weak but the bonds are strong. Grabowicz et al. [15] find that communities based on interpersonal connections emphasize bonding over identity, which corroborates Sassenberg's idea.

To address the first challenge, we lift Sassenberg's distinction to the geo-social setting. Participation in a physical space (being near each other) indicates common identity. For example, people living in Manhattan have a common identity. In contrast, social linkages with others indicate common bonds. Percimo synthesizes the effect of both kinds of attachment in location estimation. To address the second challenge, Percimo assumes that users in a geo-social community have similar communal interests, and it modulates the communal interests with a user's personal interests. Percimo assigns the most likely geo-tag to a tweet by balancing *historical* (user's prior geo-tags, suited to a tweet about personal interests) and *social* (geo-tags of others in the user's community, suited to a tweet about community interests) effects.

We exploit the correlation between users' tweets and locations of geo-socially related users via a simple technique—mapping a user's tweets to her locations [9]. However, as more advanced techniques for relating a user's tweets to her locations become available, we believe that our approach would yield a better accuracy. Nonetheless, investigating the correlation between content and locations of geo-socially related users will benefit location estimation as well as location recommendation applications.

Contributions: Percimo's novelty lies in, first, how it exploits the correlation between users' textual content and their locations for location estimation. Percimo is the first approach to solve the problem by incorporating correlation between users. Second, Percimo models the correlation in location estimation by integrating a user's personal and community interests. Third, it employs communities and investigates the effect of different geo-social relationships in location estimation with inspiration from sociology, specifically, the common-bond and common-identity theory.

Main Findings: We evaluate Percimo via a dataset consisting of geo-tagged tweets collected over two months from two US states. By exploiting the correlation between multiple users' interests and locations, Percimo reduces prediction error over baseline models that rely purely on personal history, and predicts geo-tags even for users without historical geo-tags. By reducing the size of location-candidate sets through communities, Percimo greatly reduces the prediction error compared to a purely content-based state-of-the-art technique, and the synthesized geo-social attachment reduces the prediction error compared geo and social attachments, considered separately.

II. Data, Problem, Framework

We evaluate our approach based on data from Twitter and Foursquare. This data includes all tweets with geo-tags in bounding boxes approximating two US states: Maryland (MD) and North Carolina (NC) from August 5 to October 8, 2013. Considering two states helps ensure geographical dispersal of users. We removed users with fewer than five tweets and tweets whose geo-tag was not lat-lon coordinates (some geo-tags are a city or neighborhood). This yielded 1,066,327 tweets from 23,897 distinct users (accounts). Using the Twitter API, we created a mutual-follow graph of users: an edge connects two users who follow each other, and we treated the mutual-follow graph as a social graph.

To mitigate sensing errors, we discretized locations into $30\,\mathrm{m} \times 30\,\mathrm{m}$ cells on a spatial grid, generating 106,927 nonempty cells. We removed cells that were visited fewer than five times, yielding a total of 23,858 cells, each with an assigned ID. We treat each cell as a location.

We posit that point of interest (POI) information provides a conceptual meaning of a location, and helps in relating a user's tweets to her locations. Thus, we collected POI information from Foursquare [16]. For each tweet, we collected POIs (and each POI's top-level venue category) within a $500\,\mathrm{m}$ radius [9], and labeled the tweet with the closest POI.

Our final dataset contains 23,858 unique locations, 54,062 representative POIs, and 695,636 tweets from 12,500 users (6,824 in MD, 4,984 in NC, and 692 elsewhere).

A. Problem and the Percimo Framework

Let $U = \{u\}_{u=1}^{N}$ be a set of N users and $L = \{l\}_{l=1}^{M}$ a set of M locations. Given a time T, each user has a tweet log $X_u^T = \{x_u^t\}_{t=1}^{T}$, where x_u^t represents user u's tweet at time t. A tweet x_u^t may optionally be tagged with location l_u^t representing where the tweet originated. Let $L_u^T = \{l_u^t\}$ be the set of all such locations for user u until T. And, $G(U, E)$ be a social graph, where E is the set of bonds (friendships).

Now, our research task is: Given the tweet and location log of all users until time T, social graph G, and a user u's tweet x_u^{T+1}, determine its associated location $l_u^{T+1} \in L$.

$$l_u^{T+1} = \underset{l \in L}{\arg\max}\, P(l|X^T, L^T, G, x_u^{T+1}) \qquad (1)$$

Figure 1 shows Percimo's major steps: the first two involve offline and the third step involves online processing.

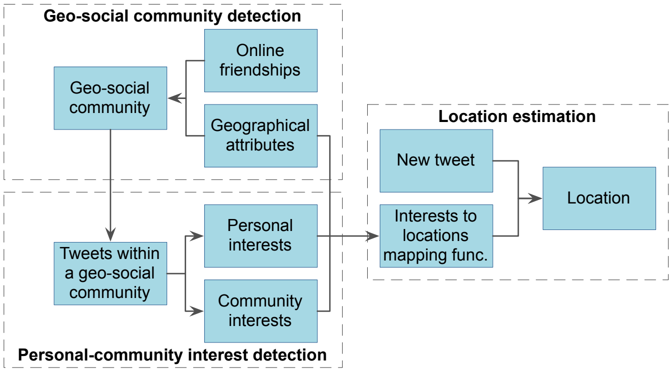
Geo-social community detection
Online friendships
Geo-social community
Geographical attributes
Location estimation
New tweet
Location
Interests to locations mapping func.
Personal interests
Tweets within a geo-social community
Community interests
Personal-community interest detection

Fig. 1: The Percimo framework.

Geo-social community detection involves detecting communities of users with similar location-visiting behavior. This step exploits both kinds of attachment (bond and identity).

Communities help overcome location data sparsity by limiting geo-tags to those of related, as opposed to all, users.

Personal-community interest detection learns the interests relationship between users within a geo-social community from the contents of their tweets. The output is a personal-interests distribution and a community-interests distribution for each user. We assume that users in a geo-social community have similar community interests.

Location estimation involves constructing a mapping $L_c^T = f(X_u^T)$ from user u's interests to location candidates, which include historical locations of u and of other users in u's geo-social community. A user's interests are correlated to her historical locations [9], [17]. Percimo models the correlation between users by relating one user's communal interests to *other* users' locations.

III. Percimo: Proposed Approach

We now describe the three major steps of Percimo.

A. Geo-Social Community Detection

We explore three kinds of *geo-social graphs* to investigate the three corresponding geo-social attachments.

- Social (G_S): an edge between users indicates a common bond.
- Local (G_L): an edge between users indicates common identity (geographical distance is below a certain threshold).
- Local-social (G_{LS}): an edge between users indicates a common bond as well as a common identity.

We first assign to each user a representative (likeliest) location m_u. We divide users into two sets: *users with a history* (at least one prior geo-tagged tweet) and *users without a history* (all other users). For each user with a history, we compute the latitude and longitude of m_u as the mean values of the latitudes and longitudes of her historical geo-coordinates L_u^T, respectively. For each user without a history, we infer her m_u via Jurgens' spatial label propagation [18] algorithm. This algorithm proceeds iteratively, where in each round, a user's location is the geometric median of all her friends' locations. No closed form solution exists for Jurgens' algorithm, but the experiment results show that propagation usually converges after four iterations. Next, we compute the distance between each pair of users to decide whether the two users live locally based on a threshold (Section V varies the threshold to investigate Percimo's prediction error).

Prentice et al. [13] describe that people form communities spontaneously. Since acquiring ground truth on user-formed communities is not feasible, we apply a community-detection technique. We adopt the Clauset-Newman-Moore [19] algorithm, which works by greedily optimizing the modularity and runs faster than many competing algorithms on large networks. However, Percimo is not restricted to this algorithm.

B. Personal-Community Interest Detection

For each geo-social graph, we learn the interests of each user in the same community. We assume a tweet's content captures (some of) a user's interests. The main idea is to mimic a user's decision making, for example, in deciding the location she wants to visit, and the set of words she wants to include in a tweet depending on her current location.

We make three assumptions about user behavior. First, a user's location visiting behavior is driven by her interests. For example, a user interested in *socializing* would go to a bar whereas a user interested in *classical music* would visit a concert hall. Second, users in a community might have similar interests (*community interests*). Third, a user's interest is based either on her *personal interest* or her community's interest. A user's interests may differ from her community's, especially when a community is not formed of common interests. For example, a student's interest in shopping malls may be higher than her residential hall community's, which presumably isn't based on a common interest in shopping.

Our interest detection model is based on Latent Dirichlet Allocation (LDA) [20]. We run our model once for each community, providing the set of tweets of all members of the community as input. We assume that each tweet has only one hidden interest label—generated by either a user's personal or her community's interests, similar to having a single label for each message [9], [21].

Figure 2 shows a graphical representation of our interest detection model and Table I shows important notations. The generative process is as follows. A user u first decides whether to go to a location from her community interests or personal interests. If she chooses the former, she selects an interest from φ_c; otherwise, she selects an interest according to η_u. With the chosen interest, words in the tweet are generated from her interest-word distribution ϕ_u. A Bernoulli distribution governs whether a user will choose her community interests or her personal interests.

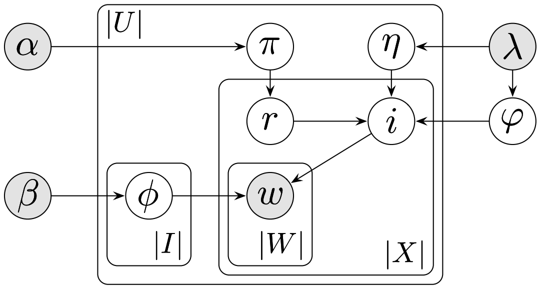

Fig. 2: Percimo's interest-detection model.

The following steps describe the generative process:

1) For a community c, draw $\varphi_c \sim Dirichlet(\lambda)$.
2) For each user u in community c,
   a. Draw $\eta_u \sim Dirichlet(\lambda)$;
   b. For each interest, draw $\phi_u \sim Dirichlet(\beta)$;
   c. Draw $\pi_u \sim Beta(\alpha)$.
3) For each tweet of a user u,
   a. Sample an indicator $r \sim Bernoulli(\pi_u)$;
   b. Sample an interest i: if $r = 1$, $i \sim Multinomial(\varphi_c)$, else $i \sim Multinomial(\eta_u)$.
4) For each word, sample $w \sim Multinomial(\phi_u)$.

TABLE I: Important notations

α, β, λ	Priors of Dirichlet distributions
U, L, X, R, I, W	Set of users, locations, tweets, indicators, interests, and words, respectively
u, l, x, r, i, w, c	Instance of a user, location, tweet, indicator, interest, word, community, respectively
φ	Community-interest distribution
η	Personal-interest distribution
π	Bernoulli distribution over indicators
ϕ	Multinomial distribution over words
n^{-x}	The counter calculated by excluding tweet x
$n_{r,u}$	Number of times r is observed in u's tweets
$n_{i,u}$ ($n_{i,c}$)	Number of tweets by u (any user in c) that are assigned to i
$n_{w,i,u}$	Number of times that w is generated by i for u
$Y_{w,x}$	Count of word w in tweet x
Y_x	Total number of words in tweet x
μ	Parameter that controls the weight of historical effect
A	Category of a location

The posterior probability of the latent variables in the model, given the observed data, can be factorized as follows:

$$P(\mathbf{w},\mathbf{r},\mathbf{i}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\lambda}) = P(\mathbf{r}|\boldsymbol{\alpha})P(\mathbf{i}|\mathbf{r},\boldsymbol{\lambda})P(\mathbf{w}|\mathbf{i},\boldsymbol{\beta}) \qquad (2)$$

We adopt collapsed Gibbs sampling [22] to approximate the latent variables. For a tweet x, we know it is from user u. The Gibbs sampler jointly samples r_x and i_x based on the values of all other hidden variables. Here, i_x represents u's interest for the tweet x; $\mathbf{i}^{-x}$ denotes all i except i_x; W_x denotes the set of words in tweet x (other variables have similar symbols). For each user u, the Gibbs update equation is:

$$\begin{aligned} P(r_x, i_x|\mathbf{r}^{-x}, \mathbf{i}^{-x}, \mathbf{w}) &\propto \frac{n_{r,u}^{-x}+\alpha_r}{\sum_{r\in R}(n_{r,u}^{-x}+\alpha_r)} \\ &\cdot \frac{n_{i,k}^{-x}+\lambda_i}{\sum_{i\in I}(n_{i,k}^{-x}+\lambda_i)} \cdot \frac{\prod_{w\in W_x}\prod_{y=0}^{Y_{w,x}-1}(n_{w,i,u}^{-x}+\beta_w+y)}{\prod_{y=0}^{Y_x-1}\left(\sum_{w\in W}(n_{w,i,u}^{-x}+\beta_w)+y\right)}, \end{aligned} \qquad (3)$$

where $k = u$ when $r_x = 0$, and $k = c$ when $r_x = 1$.

C. Location Estimation

Given a geo-social community, we now estimate the location of a new tweet by building the mapping function from users' interests to their historical locations L_c^T. We model the correlation between users' interests and their historical locations by balancing *historical* and *social* effects [23]. When a user's tweet is about her personal interests, we posit that her location is unrelated to locations of others in her community: the candidates are her historical locations (historical effect). When the user tweets about her community interests, based on the assumption that users who send similar textual content are correlated with the locations of geo-social related users, we posit that her location may be the same as another user's. In this way, historical locations of all other users are candidates (social effect). For example, colleagues sharing the interest *pizza* might go to a pizzeria for lunch on the weekdays.

The probability of selecting a candidate $l \in L_c^T$ is:

$$\begin{aligned} P(l|X^T, L^T, G, x_u^{T+1}) &= P(l, i_x|\eta_u, \varphi_c, \pi_u) \\ &= \mu \times P(l, i_x|\eta_u) + (1-\mu) \times P(l, i_x|\varphi_c), \end{aligned} \qquad (4)$$

where $\mu \in [0, 1]$ is a parameter that controls the weight between historical and social effects. We set $\mu = P(r_x = 0|u)$, where $P(r_x = 0|u)$ is learned from π_u.

$P(l, i_x|\eta_u)$ is the probability of selecting location l from user u's historical locations. Following Chen et al. [9], we posit that a user would visit locations of a category A driven by the same interest, even if the locations are distinct. For example, for a user u, if we detected that two of her tweets are labeled with the interest *eating*, it is likely that both tweets are sent from locations belonging to the *food* category.

$$P(l, i_x|\eta_u) = P(i_x|u)P(A|i_x, u)P(l|A, u), \qquad (5)$$

For a user without a history, her representative location m_u (computed from label propagation) is the only candidate for the historical effect, and $P(m_u, i_x|\eta_u)$ is always 1.

$P(l, i_x|\varphi_c)$ is the probability of selecting location l from user u's community's historical locations. We posit that users with the same interests and in the same community tend to visit locations with the same category, although their probabilities of visiting a location may differ.

$$\begin{aligned} P(l, i_x|\varphi_c) &= P(i_x|c)P(l|i_x, c) \\ &= P(i_x|c)\frac{\sum_{v\in c} s(u,v) \times P(A|i_x, v) \times P(l|A, v)}{\sum_{v\in c} s(u,v)}, \end{aligned} \qquad (6)$$

where $s(u, v)$ is the similarity between users u and v. We consider only a user having history as user v.

We compute $s(u, v)$ as follows. If u has a history, we set $s(u, v)$ to be her *check-in similarity*, defined as the cosine of their check-in vectors, whose i-th component is the number of times the user visited location i [23]. If u does not have a history, we compute $s(u, v)$ based on the distance between the representative geo-tags of the two users. Specifically, we set $s(u, v)$ to $s_{dist}(u, v) = a \times distance(m_u, m_v)^b$ [24], where $a = 0.0414$ and $b = -0.508$ are parameters values as determined by Ye et al. [11].

IV. Evaluation

Our objectives are to compare Percimo's prediction error (1) to that of the baseline models, (2) for three kinds of geo-social attachment, and (3) for different parameter settings.

A. Evaluation Strategy

We investigate the prediction error of Percimo on each geo-social graph. We vary the threshold defining local users from 5 km to 40 km. Table II summarizes the statistics of the geo-social graphs we study. The subscript indicates the threshold, e.g., G_{LS_5} represents the local-social graph with the threshold 5 km. In each graph, we ignore isolated users. Since the number of users varies across graphs, to compare Percimo's prediction error across graphs, we employ the 5,623 users appearing in G_{LS_5} because these users also appear in the other graphs. We construct local-social graph with the threshold $5\,\mathrm{km}$ on the state-level sub-datasets.

TABLE II: Statistics of the geo-social graphs

Graph	Users	Edges	Mean degree	Clustering coefficient
G_S	8,483	23,163	5.46	0.14
G_{L_5}	8,485	1,202,908	283.54	0.81
G_{LS_5}	5,623	9,350	3.33	0.19
G_{LS_10}	6,508	13,827	4.25	0.18
G_{LS_20}	7,106	16,523	4.65	0.18
G_{LS_40}	7,541	18,487	4.90	0.17
G_{LS_5} (MD)	2,930	5,441	3.71	0.17
G_{LS_5} (NC)	2,242	3,375	3.01	0.27

Parameters of Percimo. We set the total number of interests $|I|$ to 20, λ to $\frac{10}{|I|}$, and β to 0.01. We set these parameters based on guidance from previous studies [9], [25] and our preliminary experiments. A simple way to set α is to choose symmetric priors (i.e., $\alpha_1 = \alpha_0 = 0.5$) for each user, meaning that the user's historical locations and the locations of her community have equal influence in inferring a new location for the user. However, this may not be the case. Cho et al. [26] found that, on Brightkite (a location-based social network), there is a 53% chance that a user will check in at a location where she previously checked in, whereas only a 10% chance that she will check in at a location where a friend previously checked in. We set a user's α_1 as the user's betweenness centrality [27] in the subgraph of a geo-social graph induced by the user's community (and $\alpha_0 = 1 - \alpha_1$). Thus, the higher the user's betweenness centrality the greater the community's influence. We compute Percimo's interest-detection model for each geo-social graph via 500 iterations of Gibbs sampling. We take 25 samples with a gap of five iterations in the last 125 iterations to compute the values of all hidden variables.

We infer the representative geo-tag of a user without history via Jurgens' *geometric median select* method [18] with seven iterations.

Evaluation Metric. We temporally order each user's geo-tagged tweets, and take the first six weeks of data (05 August 2013 to 21 September 2013) as the training set, and test on the last two weeks of data (22 September 2013 to 08 October 2013). For each user, we predict the location of every tweet in the test set. We compare Percimo and the baseline models via average error distance (*AED*) [9]. For a tweet, error distance (*ErrDist*) is the geographical distance between the tweet's actual location and its predicted location, and the error distance of a user (*ErrDist(u)*) is the average error distance (AED) over all of her test tweets:

$$AED = \frac{\sum_{u \in U} ErrDist(u)}{|U|} \tag{7}$$

B. Baseline Models

PIM (*Personal Interest Model*) [9] is most similar to Percimo among the existing works. PIM maps a user's interests detected from tweets to her historical locations and predicts the user's next location from her historical locations, not considering the social effect. We implement PIM and choose the parameters as Chen et al. do.

CM (*Content-Based Model*) Cheng et al. [6] predict a user's location purely based on her tweets' content. We adapt this approach to consider all tweets from a given location l: $P(l|S_{words}(X_l)) = \sum_{w \in S_{words}(X_l)} P(l|w)P(w)$, where $S_{words}(X_l)$ is the set of words in all tweets from location l. We compute $P(l|w)$ via maximum likelihood estimation and $P(w)$ as $\frac{count(w)}{|W|}$, where $count(w)$ is the number of occurrences of w. We implement two enhancements Cheng et al. suggested: (1) discarding nonlocal words, and (2) performing lattice-based neighborhood smoothing.

CommPIM combines PIM and communities in geo-social graphs. We apply Chen et al.'s [9] model to detect each user's interests distribution and hidden interest label. Similar to Percimo, the location candidates are L_c^T. Whereas Percimo learns η_u from the interest detection model, CommPIM learns it from PIM: $P(l|X^T, L^T, G, x_u^{T+1}) = \mu \cdot P(l, i_x|\eta_u) + (1-\mu) \cdot \frac{\sum_{v \in c} s(u,v) P(l, i_x|\eta_v)}{\sum_{v \in c} s(u,v)}$, where $P(l, i_x|\eta_u)$ is computed according to Equation 5.

URLM (*User Representative Location Model*) always uses the representative location of a user as the prediction.

CRLM (*Community Representative Location Model*) always uses the representative location of a user's community m_c as the prediction. We compute m_c by averaging the latitude and longitude of the community's users' representative locations (geo-coordinates).

V. Results

Assumption Validation: We first validate the assumption that a user's interests are correlated with categories of her visited locations. For each user and each category, we compute $P(i_x|A, u)$ from all her tweets based our detected interests labels. If the probability is high for a certain interest k, it indicates that when the user visits locations belonging to category A, she tends to have the same interest k. We choose three thresholds (30%, 50%, 80%). For each category, if there is one interest k that makes a user's $P(i_x = k|A, u)$ exceed the threshold, we call the user a *valid user*. Figure 3 reports the percentage of valid users for each category (number of valid users divided by the number of users who send tweets from the category). In comparison, we randomly select the interest label for each tweet, and report the percentage. We observe that (1) the percentage from detected interests is always higher than that from the random labels for each threshold, (2) when the threshold is 80%, the percentage from detected interests is above 50% or slightly below 50% for all categories. For each threshold, we conduct a paired t-test: each value in the first sample is the percentage of valid users from detected interests; each value in the second sample is the percentage from random interests. The p-value is 0.0001 for each of the three thresholds. Therefore, we claim that interests and location category are strongly correlated.

Percimo and Baseline Models: Table III shows AEDs for all models for users with and without a history, except PIM, which works only for users with a history. The type of a model

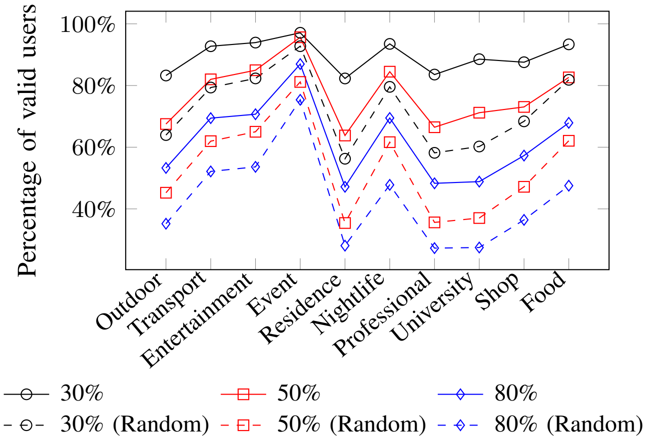

Fig. 3: Valid users based on detected and random interests.

indicates its main aspects: I and B for common identity and common bond, respectively and H for historical effect only (neither bond nor identity). In our dataset, 16.68% users have no history (no geo-tagged tweets in the training set). We set 5 km as the threshold defining local users (other thresholds below). On the local-social graph (G_{LS_5}), Percimo yields the least prediction error among the models compared.

TABLE III: AEDs (km) of Percimo and baseline models

Model	Type	Users with a history	Users with no history	All users
Percimo (G_{L_5})	I	8.74	45.94	14.94
Percimo (G_S)	B	8.47	52.90	15.88
Percimo (G_{LS_5})	I+B	**6.77**	**45.02**	**13.15**
PIM	H	8.28	–	–
URLM	H	8.32	52.35	15.41
CRLM (G_{L_5})	I	12.36	46.94	18.13
CRLM (G_S)	B	63.39	79.94	66.15
CRLM (G_{LS_5})	I+B	8.90	46.37	15.15
CommPIM	I+B	7.21	46.06	13.69
CM	I	269.87	268.48	269.64

First, Percimo yields better results than PIM, suggesting that a community-based approach yields lower prediction error than individual-based approaches. Second, although Percimo and CommPIM both set μ as 1 minus a user's betweenness centrality, Percimo learns η_u via the interest-detection model. Thus, the lower prediction error of Percimo can be attributed to its interest-detection model, which effectively models the interests relationship between users, and effectively maps users' interests to their historical locations.

Among the models compared, CM's AED is the worst, supporting our claim that a large candidate pool increases the probability of a tweet's predicted geo-tag to be far from the actual. Also, CM's AEDs do not differ much for the two kinds of users as CM does not consider the historical effect.

Although URLM and CRLM baselines seem naïve, their AEDs are not bad (except CRLM (G_S)), suggesting that geographical influence is a crucial factor in location estimation. Percimo and CRLM both yield their best results on G_{LS} among the three geo-social graphs. However, the common-bond attachment performs much better in Percimo. These suggest (1) the synthesized attachment performs best and (2) common-bond attachment can play an important role if we properly relate one's interests to another's locations.

Finally, we observe that Percimo on G_{LS_5} yields the least prediction error for sub-datasets for each state (NC and MD). The other models follow a pattern similar to the entire dataset. We omit a detailed state-level analysis for brevity.

Threshold of Defining Local Users: We vary the threshold defining local users between 5 km and 40 km to study its effect on Percimo. We restrict our analyses to the local-social graph, which has the lowest AED for both kinds of users. Figure 4 shows that the lower the threshold the lower the AED, in general. The AED of G_{LS_20} is higher than that of G_{LS_40} for users with a history and the reverse for users without a history. A similar pattern arises for Percimo on G_{L_5} and G_S (Table III). That is, at higher thresholds, the identity effect (locality) fades, and the bonding effect (sociality) dominates.

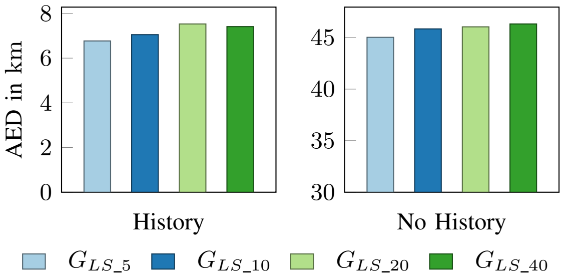

Fig. 4: Percimo's AEDs for four local-social graphs.

Social and Historical Effects: Percimo balances social and historical effects by learning μ (Equation 4). Setting $\mu = 1$ and $\mu = 0$ forces Percimo to consider historical and social effect only, respectively. Figure 5 compares Percimo's AEDs for the three settings of μ. The AED for $\mu = 1$ is less than that for $\mu = 0$: the historical effect is more important than the social effect for location estimation. However, Percimo's AED is least for learned μ, suggesting that both the historical and social effects contribute to reducing the AED.

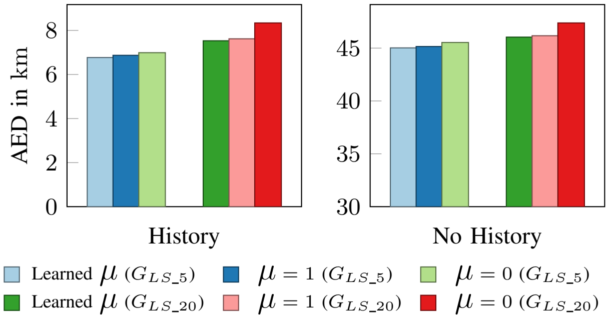

Fig. 5: Percimo's AEDs for $\mu = 0$ (social effect), $\mu = 1$ (historical effect), and learned μ (historical and social effects).

Symmetric Prior vs. Betweenness Centrality: Figure 6 compares Percimo's AED when α_1 is set as users' betweenness centrality or 0.5. The AED is higher for $\alpha_1 = 0.5$ on both graphs, whether a user has a history or not. Thus, we conjecture that setting α_1 as users' betweenness centrality is a better choice in Percimo than setting it to 0.5 (symmetric priors). Importantly, we are not suggesting that α_1 necessarily be bound to betweenness centrality; other metrics that estimate user's attachment to her community could also be good choices. We defer this analysis to future work.

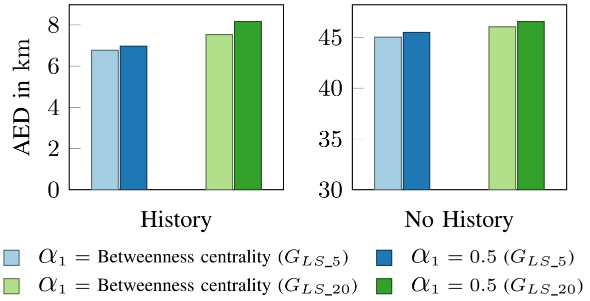

Fig. 6: Percimo's AEDs for different α_1 settings.

VI. Related Work

Works on location estimation estimate locations either of messages or of users. Percimo falls into the first category, which is based on the assumption that messages encode location-related information: either specific location names or words that are associated with the location. Cheng et al. [6] build a classifier for automatically identifying words in tweets with a strong geo-scope to estimate a user's city-level location. Chang et al. [28] estimate a user's home location via a Gaussian Mixture Model by assuming that each word has several centers where users tweet it more frequently. Besides messages, some researchers employ users' tweeting behaviors (volume of tweets per time unit, reply-tweet relationships) [7], [29]. Percimo differs from these works in that it estimates location in a fine-grained manner—at the level of every tweet.

For estimating locations at the fine-grained level, Kinsella et al. [30] estimate the location of every tweet by sampling the word distribution for that location. Instead of assuming the independence between words, some researchers [31], [8] model the location distributions of phrases (n-grams) and assign a location to a tweet by identifying n-grams associated with hyper-local regions. Schulz et al. [17] propose a multi-indicator approach with dedicated location entries and user profiles. Some researchers focus on recognizing textual references to geographical locations [32], [33]. In contrast to these approaches relying on spatial aspects of words in unstructured texts, Percimo exploits the correlation between users' textual content and their locations. Chen et al. [9] estimate location of a tweet by assuming that a user's interests are related to her locations. Their techniques apply to each user individually. Percimo is novel in that it exploits not only the correlation between a user's content and her locations, but also the correlation between a user's content and others' locations.

Approaches in the second category seek to predict the location of a user, not a message. Some works claim that the locations of a user's friends are helpful in predicting the user's location [12], [23]. Jurgens [18] infers users' locations by spatially propagating location through social network, given a small number of labeled locations. Sadilek et al. [34] propose a probabilistic model to infer a user's fine-grained location from her friends' locations. Some works estimate user location by mining mobility patterns from GPS trajectories [35], [36]. Song et al. [37] build a model that captures an individual moving to a new location and returning to a visited location. Murukannaiah and Singh [38] learn places of interest to a user from smartphone sensor data and user-provided labels. They also show that the places a user visits influences her social relationships [39]. In contrast, Percimo focuses on the content analysis of messages and the relationship between a message and its associated location. Content analysis is a rich source of knowledge for estimating locations [40] and inferring potential social relationships [41].

Some researchers build models to detect communities wherein users talk about similar topics, [42], [43], [25]. Percimo has a different motivation: it detects interests of users from communities for location estimation.

VII. Conclusions and Future Work

We estimate locations of user-generated messages such as tweets, made challenging by the sparsity of geo-tagged messages. Our approach, Percimo, addresses location estimation by exploiting (1) correlation between users' locations and their textual content, and (2) communities and different geo-social attachments. Percimo balances a user's personal and community interests to outperform a state-of-the-art technique that considers only personal interests. And, by reducing the pool of candidate locations, Percimo outperforms a state-of-the-art approach that relies solely on content information.

Percimo's parameters affect prediction error. We find that the synthesized attachment (bond and identity) yields least AED in location estimation, and a lower threshold of defining local users could reduce the prediction error. Percimo's effectiveness is limited when a user has neither geo-tagged tweets nor social relationships though it is better than traditional approaches in this respect. We defer modeling users' participation in overlapping and multiple communities to future work.

Acknowledgments: Thanks to the NCSU Laboratory for Analytic Sciences for partial support, and to the anonymous reviewers for helpful comments.

References

[1] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. Copenhagen, 2010, pp. 119–128.

[2] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: Detecting influenza epidemics using Twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK, 2011, pp. 1568–1576.

[3] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, “Chatter on the red: What hazards threat reveals about the social life of microblogged information,” in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. Savannah, GA, 2010, pp. 241–250.

[4] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, “Discovering geographical topics in the Twitter stream,” in *Proceedings of the 21st International Conference on World Wide Web*. Lyon, France, 2012, pp. 769–778.

[5] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E. Shook, “Mapping the global Twitter heartbeat: The geography of Twitter,” *First Monday*, vol. 18, no. 5, 2013.

[6] Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: A content-based approach to geo-locating Twitter users,” in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. Toronto, 2010, pp. 759–768.

[7] J. Mahmud, J. Nichols, and C. Drews, “Where is this tweet from? Inferring home locations of Twitter users,” in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. Dublin, 2012, pp. 511–514.

[8] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza, “On the accuracy of hyper-local geotagging of social media content,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. Shanghai, 2015, pp. 127–136.

[9] Y. Chen, J. Zhao, X. Hu, X. Zhang, Z. Li, and T.-S. Chua, “From interest to function: Location estimation in social media,” in *Proceedings of the 27th AAAI Conference on Artificial Intelligence*. 2013, pp. 180–186.

[10] I. Grabovitch-Zuyev, Y. Kanza, E. Kravi, and B. Pat, “On the correlation between textual content and geospatial locations in microblogs,” in *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data*. Snowbird, UT, 2007, pp. 3:1–3:6.

[11] M. Ye, P. Yin, and W.-C. Lee, “Location recommendation for location-based social networks,” in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. San Jose, CA, 2010, pp. 458–461.

[12] L. Backstrom, E. Sun, and C. Marlow, “Find me if you can: Improving geographical prediction with social and spatial proximity,” in *Proceedings of the 19th International Conference on World Wide Web*. Raleigh, NC, 2010, pp. 61–70.

[13] D. A. Prentice, D. T. Miller, and J. R. Lightdale, “Asymmetries in attachments to groups and to their members: Distinguishing between common-identity and common-bond groups,” *Personality and Social Psychology Bulletin*, 20(5):484–493, Oct. 1994.

[14] K. Sassenberg, “Common bond and common identity groups on the Internet,” *Group Dynamics: Theory, Research, and Practice*, 6(1):27–37, 2002.

[15] P. A. Grabowicz, L. M. Aiello, V. M. Eguiluz, and A. Jaimes, “Distinguishing topical and social groups based on common identity and bond theory,” in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. Rome, 2013, pp. 627–636.

[16] Foursquare, “Foursquare category hierarchy,” 2015. [Online]. Available: https://developer.foursquare.com/categorytree

[17] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser, “A multi-indicator approach for geolocalization of tweets,” in *Proceedings of the 8th International Conference on Weblogs and Social Media*. Ann Arbor, MI, 2013, pp. 573–582.

[18] D. Jurgens, “That’s what friends are for: Inferring location in online social media platforms based on social relationships,” in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*. Boston, 2013, pp. 273–282.

[19] A. Clauset, M. Newman, and C. Moore, “Finding community structure in very large networks,” *Physical Review E.*, 70(6):066111, Dec. 2004.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” *Journal of Machine Learning Research*, 3:993–1022, Jan. 2003.

[21] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, “Finding bursty topics from microblogs,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. 2012, pp. 536–544.

[22] J. S. Liu, “The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problems,” *Journal of the American Statistical Association*, 89(427):958–966, 1994.

[23] H. Gao, J. Tang, and H. Liu, “Exploring social-historical ties on location-based social networks,” in *Proceedings of the 7th International Conference on Weblogs and Social Media*. Boston, 2012, pp. 114–121.

[24] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee, “Exploiting geographical influence for collaborative point-of-interest recommendation,” in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Beijing, 2011, pp. 325–334.

[25] T. Hoang and E. Lim, “On joint modeling of topical communities and personal interest in microblogs,” in *Proceedings of the 6th International Conference on Social Informatics*. Barcelona, 2014, pp. 1–16.

[26] E. Cho, S. A. Myers, and J. Leskovec, “Friendship and mobility: User movement in location-based social networks,” in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, CA, 2011, pp. 1082–1090.

[27] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010.

[28] H.-W. Chang, D. Lee, M. Eltaher, and J. Lee, “@Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage,” in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. Istanbul, 2012, pp. 111–118.

[29] S. Chandra, L. Khan, and F. B. Muhaya, “Estimating Twitter user location using social interactions–A content based approach,” in *Proceedings of the IEEE 3rd International Conference on Social Computing*. Boston, 2011, pp. 838–843.

[30] S. Kinsella, V. Murdock, and N. O’Hare, “I’m eating a sandwich in Glasgow”: Modeling locations with tweets,” in *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*. Glasgow, 2011, pp. 61–68.

[31] R. Priedhorsky, A. Culotta, and S. Y. D. Valle, “Inferring the origin locations of tweets with quantitative confidence,” in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work*. Baltimore, 2014, pp. 1523–1536.

[32] M. D. Lieberman, H. Samet, and J. Sankaranayananan, “Geotagging: Using proximity, sibling, and prominence clues to understand comma groups,” in *Proceedings of the 6th Workshop on Geographic Information Retrieval*. Zurich, 2010, pp. 6:1–6:8.

[33] C. Li and A. Sun, “Fine-grained location extraction from tweets with temporal awareness,” in *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Gold Coast, Queensland, Australia, 2014, pp. 43–52.

[34] A. Sadilek, H. Kautz, and J. P. Bigham, “Finding your friends and following them to where you are,” in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*. Seattle, 2012, pp. 723–732.

[35] M. C. González, C. A. Hidalgo, and A. Barabási, “Understanding individual human mobility patterns,” *Nature*, 453(7196):779–782, 2008.

[36] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, “WhereNext: A location predictor on trajectory pattern mining,” in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris, 2009, pp. 637–646.

[37] C. Song, T. Koren, P. Wang, and A.-L. Barabási, “Modelling the scaling properties of human mobility,” *Nature Physics*, 6(10):818–823, Sep. 2010.

[38] P. K. Murukannaiah and M. P. Singh, “Platys: An active learning framework for place-aware application development and its evaluation,” *ACM Transactions on Software Engineering and Methodology*, 24(3):1–33, May 2015.

[39] P. K. Murukannaiah and M. P. Singh, “Platys Social: Relating shared places and private social circles,” *IEEE Internet Computing*, 16(3):53–59, May 2012.

[40] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, “Exploring millions of footprints in location sharing services,” in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*. Barcelona, 2011, pp. 81–88.

[41] G. Yuan, P. K. Murukannaiah, Z. Zhang, and M. P. Singh, “Exploiting sentiment homophily for link prediction,” in *Proceedings of the 8th ACM Conference on Recommender Systems*. Foster City, CA, 2014, pp. 17–24.

[42] K. Henderson, T. Eliassi-Rad, S. Papadimitriou, and C. Faloutsos, “HCDF: A hybrid community discovery framework,” in *Proceedings of SIAM International Conference on Data Mining*. Columbus, OH, 2010, pp. 754–765.

[43] Z. Yin, L. Cao, Q. Gu, and J. Han, “Latent community topic analysis: Integration of community discovery with topic modeling,” *ACM Transactions on Intelligent Systems and Technology*, 3(4):1–21, Sep. 2012.