FIND YOU FROM YOUR FRIENDS: GRAPH-BASED RESIDENCE LOCATION PREDICTION FOR USERS IN SOCIAL MEDIA

Dan Xu[*]*, Peng Cui, Wenwu Zhu, Shiqiang Yang*

Department of Computer Science and Technology, Tsinghua University, Beijing, China
xu09@mails.tsinghua.edu.cn, {cuip, wwzhu, yangshq}@tsinghua.edu.cn

ABSTRACT

As a bridge between social media and physical space, location information will potentially make the internet smarter, and release the real power of social media to address the serious and significant problems in the real world. However, in terms of privacy and security, most of the users are unwilling to make their locations public. To address the problem, an algorithm is necessary to predict the users' residence locations based on the public profiles. We define location propagation probability of users, leverage a semi-supervised learning algorithm, and introduce a novel method of *location propagation* to predict users' residence locations based on users' social relationships, textual and visual contents and a small amount of known users' residence locations. The experimental results on a large scale real data set in *Tencent Weibo* demonstrate that our location propagation algorithm outperforms the state-of-the-art approaches in both accuracy and scalability.

Index Terms— Social media, user profiling, social graph, location prediction

1. INTRODUCTION

As online social media grows, the amount of location information from social media users is increasing. Location sharing is becoming more and more prevalent. Location bridges the gap between our online and offline activities. In such situations, users' residence locations, which are the focus of this paper, become more important due to their vital roles in applications such as friend recommendation, personal advertisement, disaster warning and local news feeding [1], [2], [3].

However, the disclosure of location raises serious privacy and security concerns. Updating location-aware status, and sharing location may enable attackers to identify users' trajectories, or even cause users to be theft or robbed. For privacy and security concerns, more and more users are unwilling to publish or describe exactly their locations in social media. Users tend to submit more fuzzy and generalize locations. According to [4], only 6% of Facebook users publish their residence locations at the city level. These limitations reduce the location-aware services applications. Therefore, it is important to identify users' locations from the public information such as contents, user profiles, and social relationships.

According to the characters of locations, location prediction is mainly classified into the prediction of residence location and current location. In this situation, residence location is defined as the location where user's most activities happen. It captures user's statical geographic range rather than a real-time spatial point. Many applications based on location trend to leverage residence location as an symbolic feature. Therefore, we focus on users' residence locations in social media in this paper.

To predict a user's residence location, some researches have been studied [2], [4], [5]. Most prediction methods leverage user-generated contents (content-based) or social relationships (social-based), even combine both methods (combination approach) together. Content-based methods predict locations by identifying *local* words from user-generated contents. These methods do not perform well in social media because of their lack of the relationship between the user's true geographic position and the location mentioned in the content. Moreover, the high computing cost leads these methods weak scalability. Social-based approaches assume that friends in social media are located near each other, and leverage the user's social graph to predict location. Most major previous graph-based approaches leverage inductive machine learning model, whose results depend heavily on the training samples and plentiful ground truth locations.

Herein we face several key challenges. (1) The sparsity problem. As mentioned before, only few users' residence locations are shared. How to make full use of them to infer the remaining major part? (2) The noisy problem. Users' following behaviors and their generated contents are often casual and uncertain. We cannot directly apply our existing prior to solve the sparsity problem. (3)The heterogeneous information. Here we face both social and content information. They play different roles in predicting residence locations. How do we balance them?

Considering all these challenges, we first get insightful observations from massive data, and then use these observa-

[*]This work is supported by National Natural Science Foundation of China, No. 61370022, No. 61303075 and No. 61210008. International Science and Technology Cooperation Program of China, No. 2013DFG12870; National Program on Key Basic Research Project, No. 2011CB302206.

tions to guide the predictive model design. The model translates the label propagation term on location. The contributions of this paper are as follows:

- We propose a novel framework for residence location inference in social media, by jointly considering social and content information.
- We propose a data-driven approach unveil the phenomena of friendship locality, social proximity and content proximity for geographically nearby users.
- We propose a location propagation algorithm to effectively infer residence location for social media users. In our experimental setting, we achieve 21% relative improvement over state-of-the-art approaches.

The rest of this paper is organized as follows. In section 2 we overview related work with an emphasis on user location prediction. Section 3 defines the terminologies and formulates the problem addressed in this paper. The location propagation model is proposed in Section 4. Section 5 describes the experiments conducted to verify the accuracy of our approach compared to the state-of-the-art approaches. Finally, Section 6 draws the conclusion.

2. RELATED WORK

In recent years, user's residence location prediction in social media has become an increasingly active research field. In this section, we briefly review the three primary clues of related work, including content-based approaches, graph-based approaches and the combination of the both approaches.

Content-based approaches. These approaches leverage user-generated contents. Cheng et al. [5] focused on the residence location inference in Twitter by leveraging the local terms posted in a specific geographic region. Chandra et al. [6] developed a language model based on users' conversations. In the model, all terms in the same conversation belong to the conversation initiator. Chang et al. [7] inferred user locations without training data by proposing the location distributions of terms based on a Gaussian Mixture Model. The experiments confirmed that the method could achieve a better accuracy.

Social-based approaches. These approaches utilize user relationships on social graphs. Backstorm et al. [4] introduced a location estimation method for Facebook by probabilistic inference based on a user's friends. They firstly assign the probability of friendship versus the users' geographic distance, and then evaluate users' locations by employing maximum likelihood estimation. Sadilek et al. [2] predicted users' trajectories based on social graph. However, the existing graph-based approaches assume the probability of friendship at the same distance is the same. In fact, it is usually invalid. As a result, these models can not differentiate users with different influence.

Combined approaches. These approaches focus on both user-generated contents and social graphs. Li et al. [8] developed a unified discriminative influence model to profile users' residence locations. Based on both user generated contents and social relations, they integrated signals from both tweets and friends in a unified probabilistic framework to address the problem of sparsity and noise. Base on the model, the multiple locations profiling model [9] also proposed to address the problem of multi-locations.

3. PROBLEM STATEMENT

To ease our further description, this section defines the terminologies and describes the problem addressed in this paper.

Notation. In a social media platform such as *Tencent Weibo*, given a user, we detect the following signals: users' locations and following relationships between the users. A user v_i follows v_j does not indicate that v_j follows v_i, i.e., the following is a one-way relationship. Accordingly, we classify relationships into followings and followers. Specially, if v_j and v_i follows each other, we define the relationship between v_i and v_j as friends.

We summarize a social media as a directed graph $G = G(V, E)$, where V is the user set of v_i and E is the relationship set of $e(v_i, v_j)$ from v_i to v_j. Generally, every user v_i is related to a location ℓ_i. We view ℓ_i as a coordinate point $(longitude, latitude)$ on the geographic space. Our goal is to predict the missing locations. We denote the users whose locations are known as located users $V_l = \{v_1, \ldots, v_l\}$, and the rest users as unlocated users $V_u = \{v_{l+1}, \ldots, v_{l+u}\}$. If some of a user's social neighbors publish their locations, their locations can be propagated to him. In this notation, the problem of user location prediction is stated as:

Location Prediction Problem Given a social graph $G(V, E)$, predict the residence location of each unlocated user $\{v \in V_u\}$ so that the predicted location ℓ_v is close to the true location ℓ_v^{true}.

4. LOCATION PROPAGATION

4.1. Motivation

We study about $200,000$ sampled users from *Tencent Weibo*(http://t.qq.com/). including with their ID, followers, followees and residence locations. Also we collect one month tweets that are generated or shared by these sampled users. We observe three phenomena as following.

Friendship Locality. The first observation comes from one hypothesis. Actually, the online social graph somehow reflects peoples' offline social relations. Considering the spatial limitation in physical world, we assume that geographically nearby users are more probable to establish friendship relations. Then we hope to validate this hypothesis in real data. We plot the log-log figure for probability of friendship

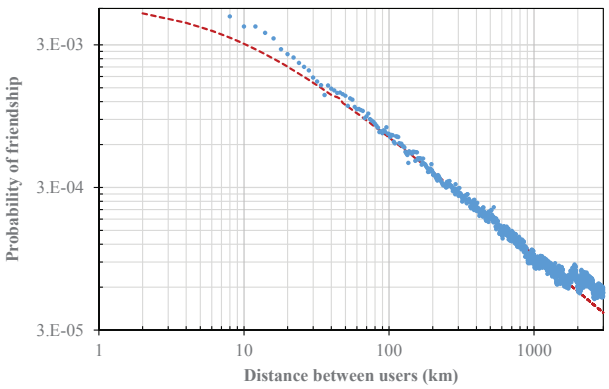


Fig. 1. Distance Distribution in Tencent Weibo

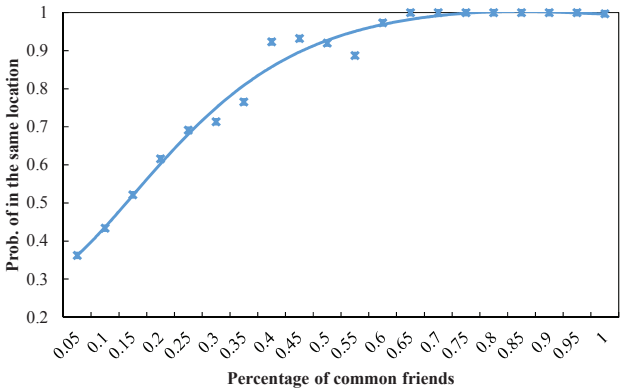
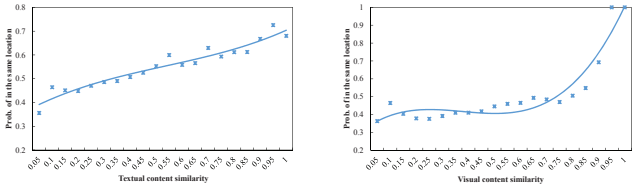


Fig. 2. The prob. of two users in the same location VS. the percentage of common friends



(a) Textual content similarity (b) Visual content similarity

Fig. 3. The prob. of two users in the same location VS. the content similarity

versus distance between users as Fig.1, so we can observe an obvious power law in this figure, which demonstrate the existence of friendship locality. Based on this observation, we can get an insight that residence location can be propagated along friendship relations, which fundamentally motivate us to adopt an label-propagation model for residence location inference.

Social Proximity. Based on the first observation, we further dig into the probability of two users sharing the same residence location. Enlightened by the friendship locality, we assume that geographically nearby users tend to have more common friends. We validate this hypothesis in real data, and plot the probability of two users sharing the same residence location versus their percentage of common friends as Fig.2. We can see that the probability of two users sharing the same residence location is monotonically increasing with the increase of the percentage of their common friends, which demonstrate the existence of social proximity. Hence, we can get the insight that the social proximity, which is quantified by the concept that the number of common friends is a key factor of location propagation probability between any pair of users.

Content Proximity. The third observation origins from the perspective of content. Considering that users' generated content in social media mainly comes from their life in physical world, so geographically nearby users tend to publish similar geo-related contents. In order to validate this, we plot the probability of two users sharing the same residence location versus the similarity of their generated contents as Fig.3. We can see that the probability of two users sharing the same residence location is monotonically increasing with the increase of content similarity in both modalities. So we get the third insight that the content proximity is another key factor of location propagation probability between any pair-wise users.

4.2. Location Propagation Probability

Based on all these observations and insights, we propose a social-content joint location propagation framework. From the raw data of social relations and profiles, and textual visual user generated contents, we extract the social graph with known locations as the propagation medium, calculate the content and social proximity to define the propagation probability, and integrate them into the location propagation algorithm to infer the residence locations for unknown users.

Definition 1 (*Content Proximity*) The content proximity is define as a linear combination of textual content similarity and visual content similarity. It can be represented as

$$\mathcal{P}_{con}(i,j) = \beta \mathcal{S}_{txt}(i,j) + (1-\beta)\mathcal{S}_{vis}(i,j), \quad 0 < \beta < 1. \tag{1}$$

Here we represent geo-related textual content by word vectors, and the geo-related images by visual word vectors, and finally use Consine distance to measure the content proximity.

$$\mathcal{S}_{txt}(i,j) = \frac{txt_i \cdot txt_j}{||txt_i|| ||txt_j||}. \tag{2}$$

$$\mathcal{S}_{vis}(i,j) = \frac{img_i \cdot img_j}{||img_i||||img_j||}. \quad (3)$$

Definition 2 (*Social Proximity*) We define the social proximity as the Jaccard distance between the target users. It can be represented as

$$\mathcal{P}_{soc}(i,j) = \frac{|F_i \cap F_j|}{|F_i \cup F_j|} \quad (4)$$

Definition 3 (*User Similarity*) The similarity $\mathcal{P}_{ij}$ is defined as a linear combination of social proximity $\mathcal{P}_{soc}(i,j)$ and content proximity$\mathcal{P}_{con}(i,j)$. It can be represented as

$$\mathcal{P}_{ij} = \alpha\mathcal{P}_{con}(i,j) + (1-\alpha)\mathcal{P}_{soc}(i,j), 0 < \alpha < 1. \quad (5)$$

Definition 4 (*Location Propagation Probability*) The location propagation probability $t_{i,j}$ is denotes probability of location propagation from user v_i to user v_j. It can be represented as

$$t_{i,j} = P(j \to i) = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{kj}}. \quad (6)$$

We use standard normal distribution to calculate the weight w_{ij}

$$w_{ij} = \exp\{-\frac{\mathcal{P}_{ij}^2}{2\sigma^2}\}. \quad (7)$$

4.3. Model formulation

In this paper, we extend label propagation algorithm [10] to location prediction, which is a semi-supervised, iterative algorithm designed to infer labels for items connected in a network. Usually, the true labels are known for only a small fraction of nodes in the network, which serve as a source of ground truth information for inference the labels of other nodes. The algorithm proceeds iteratively, where in each round, items receive the most frequent label from their neighbors. Herein we apply the label term to residence location term.

We define an $(l+u) \times (l+u)$ probability propagation matrix T to measure the probability of location propagation from a user to his friends.

$$T = \begin{pmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,l+u} \\ t_{2,1} & t_{2,2} & \dots & t_{2,l+u} \\ \vdots & \vdots & \ddots & \vdots \\ t_{l+u,1} & t_{l+u,2} & \dots & t_{l+u,l+u} \end{pmatrix} \quad (8)$$

We define the location label normalized matrix $\mathcal{L}_{(l+u)\times p}$, where p is the number of locations. Initialize the matrix as

$$\mathcal{L}_{ij}^{(0)} = \begin{cases} 1; & \ell_{ij} \text{ is the location of user } v_i \\ 0; & else \end{cases} \quad (9)$$

The location propagation algorithm is described in Algorithm 1:

Theorem 1 *The location propagation prediction algorithm converges.*

Algorithm 1 Location Propagation Algorithm

Input: $G = (V, E)$; $\{v_1, \ell_1\}, \dots, \{v_l, \ell_l\}$;
Output: $\{v_{l+1}, \ell_{l+1}\}, \dots, \{v_{l+u}, \ell_{l+u}\}$;
1: Calculate weight of user similarity w_{ij};
2: Calculate the propagation matrix T;
3: Initialize $\mathcal{L}^{(0)}$;
// The u bottom lows are assigned as 0;
4: **for** $t = 0$; $\mathcal{L}^{(t)}$ converges; $t++$ **do**
5: $\mathcal{L}^{(t)} = T\mathcal{L}^{(t-1)}$;
// In the t-th iteration, each user receives the location propagated by friends according to the similarity matrix, updates it's probability distribution;
6: Clamp the labeled data;
// Keep the initial locations of located users;
7: **end for**
8: Return location of unlocated user.

Proof.

$$\begin{aligned} \mathcal{L}^{(t)} - \mathcal{L}^{(t-1)} &= T^{(t)}\mathcal{L}^{(0)} - T^{(t-1)}\mathcal{L}^{(0)} \\ &= T^{(t-1)}[T - I]\mathcal{L}^{(0)} \end{aligned} \quad (10)$$

Because every row of the probability propagation matrix T is non-negative and the sum is 1, so

$$\lim_{t\to\infty} T^{(t-1)} = 0 \quad (11)$$

Hence the matrix $\mathcal{L}$ must converge, so the algorithm must converge. □

5. EXPERIMENTAL RESULTS

5.1. Data set

We first sample 2 million users from the complete dataset of Tencent microblog, and their ID, followers, followees and residence locations are included. Also we collect one month microblogs that are generated or shared by these sampled users. For the residence locations, we extract both city-level(*city, province*) and province-level(*province*) locations from their profiles. So, in total, we have 355 cities and 31 provinces as the pool of residence locations, as Fig.4 showed.

By identifying city names and province names in the text format from location profiles, we get the corresponding latitude and longitude pairs. We calculate the geo-distance between cities of users, observe the relationship between probability of friendship. The curves in Fig.1 show that the probability distribution is power-law.

We select located users with city level locations, and then we randomly select $500,000$ users from them to make our test-bed. Among them, we randomly select $100,000$ located users, who have at least 20 microblogs and 20 labeled followers and followings.

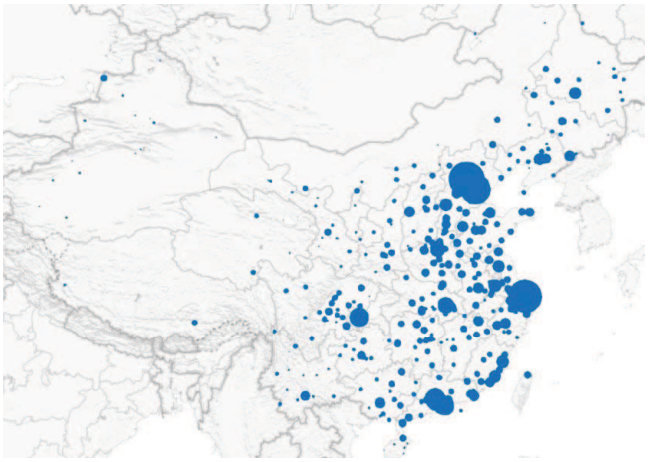


Fig. 4. User Geo Distribution of Crawled Tencent Weibo Data

5.2. Evaluation methods

We compare our method with the state-of-the-art methods based on social graph in [4] and [8].

FindMe approach is proposed in [4] to predict a user's location based on social graph, in which followers and followings are all treated as users' friends.

UDI approach is a unified discriminative influence model to infer users' residence locations in [8]. Here we do not consider user-generated contents because the objective of this experiment is to compare the performance of these social-based methods.

LPA approach is our location prediction approach, which is based on location propagation algorithm.

Our evaluation is designed with the following goals: For the evaluation, we adopt the held-out evaluation strategy. We held-out 20% known users as ground truth, and using the difference between predicted location and the ground truth location to evaluate the performance. comparing the accuracy of different approaches both at the city and the province level; showing the effectiveness of location propagation comparing to the baseline approaches.

For each test user $v_i \in V$, we calculate the Error Distance, Err_i, which represents the distance between the predicted location ℓ_i and the true residence location ℓ_i^{true}:

$$Err(v_i) = EarthDist(\ell_i, \ell_i^{true}) \tag{12}$$

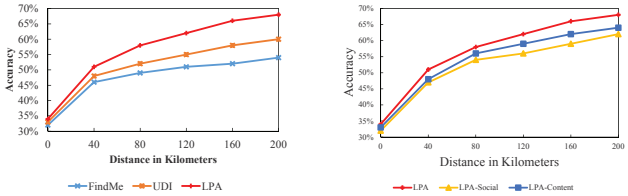
We define Average Error Distance(AED) and Accuracy (ACC) as

$$AED = \frac{\sum_{v_i \in V} Err(v_i)}{|V|} \tag{13}$$

$$ACC = \frac{|V \cap \{v_i | \ell_i = \ell_i^{true}\}|}{|V|} \tag{14}$$

Table 1. Accuracy comparison table

	FindMe approach	UDI approach	Location propagation
ACC@city	52.1%	56.2%	**68.2**$\%$
ACC@province	58.3%	60.4%	**73.7**$\%$
AED@80$\%$	320	255	**240**
AED	987	840	**800**



(a) LPA VS. Baselines (b) LPA tradeoff

Fig. 5. Accumulative Accuracy at Various Distance

5.3. Experimental result

We compare our LPA approach with FindMe and UDI approach. All of the three approaches profile users' locations based on social graphs. The performance of each method is shown in Table1. The results show that our approach outperforms the baseline approaches.

Average Error Distance. On Table1, the AED results show such an improvement over the baseline approaches. Because AED is easily influenced by outliers, we report AED at different percentage. $AED@x\%$ denotes that the average error distance of the top $x\%$ of predictions. When we compare $AED@80\%$ and $AED@100\%$, the average error distance increases to $800km$ rapidly, because the average error distance is influenced by the users predicted inaccurately. Hence, we should not just pay attention to $AED@100\%$.

Accuracy. Table 1 shows that location propagation algorithm has a very promising accuracy. LPA improves by 15% in accuracy in city level, even 20% in province level than the baseline approaches. To describe our experimental results in detail, we plot curves of accumulative accuracy at distances for each approaches in Fig.5(a). A coordinate point (x, y) in the curve shows that $y\%$ users are correct in x kilometers. The curves show that LPA approach is more accurate than baseline approaches in different distances. The reason is that LPA approach assumes that more friends are close by, while the baseline approaches restrict that non-friends have to be further away, which may not always be true.

Tradeoff of social and content proximity. In order to get more insights on the proposed method, we implement two variants. One is purely social-based LPA, the other is content-based LPA. The experimental result is shown as Table2 and Fig.5(b) which demonstrates that both social proximity and

Table 2. Accuracy comparison table

	LPA Social	LPA Content	Location Propagation
ACC@city	62.5%	64.2%	**68.2**%
ACC@provice	67.1%	69.3%	**73.7**%
AED@80%	250	247	**238**
AED	810	800	**783**

content proximity contribute much to the residence location inference.

Efficiency. We also compare the efficiency between our approach and the baseline approaches. Our approach is almost constant while the baseline approaches is linear. When the number of users is slow, our approach and the two baseline approaches take around 2 seconds. With the increasing of the users, the running time of our approach almost keeps constant, while that of the baseline approaches is increasing rapidly. The reason is that our approach only considers friends of a user, whose number is almost constant, while the baseline approaches need to consider all users, including both friends and non-friends, so it is linearly correlated to the volume of the data set. So our approach is much more efficient and scalable.

6. CONCLUSION

We propose a novel framework for residence location inference in social media, by jointly considering social and textual information. The approach translates the label propagation term into location, and addresses several challenges, including location signal sparsity, user signal noisy and similarity between different users. A data-driven approach unveils the phenomena of friendship locality, social proximity and content proximity for geographically nearby users. We propose a location propagation algorithm to effectively infer residence location for social media users. In our experimental setting, we achieve 21% relative improvement over state-of-the-art approaches. The approach outperforms the baseline approaches in time and accuracy, so it is suitable for online applications.

7. REFERENCES

[1] Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik-Lun 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*. ACM, 2011, pp. 325–334.

[2] Adam Sadilek, Henry Kautz, and Jeffrey P Bigham, "Finding your friends and following them to where you are," in *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012, pp. 723–732.

[3] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[4] Lars Backstrom, Eric Sun, and Cameron 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*. ACM, 2010, pp. 61–70.

[5] Zhiyuan Cheng, James Caverlee, and Kyumin 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*. ACM, 2010, pp. 759–768.

[6] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya, "Estimating twitter user location using social interactions–a content based approach," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on social computing (socialcom)*. IEEE, 2011, pp. 838–843.

[7] Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee, "@ phillies tweeting from philly? predicting twitter user locations with spatial word usage," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 2012, pp. 111–118.

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

[9] Rui Li, Shengjie Wang, and Kevin Chen-Chuan Chang, "Multiple location profiling for users and relationships from social network and content," *Proceedings of the VLDB Endowment*, vol. 5, no. 11, pp. 1603–1614, 2012.

[10] Xiaojin Zhu and Zoubin Ghahramani, "Learning from labeled and unlabeled data with label propagation," Tech. Rep., Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.