Modeling geo-homopholy in online social networks for population distribution projection

Yuanxing Zhang
EECS, Peking University, Beijing, China

Zhuqi Li
Department of Computer Science, Princeton University, Princeton, New Jersey, USA

Kaigui Bian
School of Electronics Engineering and Computer Science, Peking University, Beijing, China

Yichong Bai
Shenzhen Maxwell Data Tech Inc, Ltd, Shenzhen, China, and

Zhi Yang and Xiaoming Li
School of Electronics Engineering and Computer Science, Peking University, Beijing, China

Received 2 August 2017
Revised 24 August 2017
Accepted 25 August 2017

Abstract

Purpose – Projecting the population distribution in geographical regions is important for many applications such as launching marketing campaigns or enhancing the public safety in certain densely populated areas. Conventional studies require the collection of people's trajectory data through offline means, which is limited in terms of cost and data availability. The wide use of online social network (OSN) apps over smartphones has provided the opportunities of devising a lightweight approach of conducting the study using the online data of smartphone apps. This paper aims to reveal the relationship between the online social networks and the offline communities, as well as to project the population distribution by modeling geo-homophily in the online social networks.

Design/methodology/approach – In this paper, the authors propose the concept of geo-homophily in OSNs to determine how much the data of an OSN can help project the population distribution in a given division of geographical regions. Specifically, the authors establish a three-layered theoretic framework that first maps the online message diffusion among friends in the OSN to the offline population distribution over a given division of regions via a Dirichlet process and then projects the floating population across the regions.

This work was supported by National Natural Science Foundation of China under grant number 61572051 and 61632017 and by the National 973 Grant under grant number 2014CB340405.

International Journal of Crowd Science
Vol. 1 No. 3, 2017
pp. 249-269
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-08-2017-0008

Findings – By experiments over large-scale OSN data sets, the authors show that the proposed prediction models have a high prediction accuracy in characterizing the process of how the population distribution forms and how the floating population changes over time.

Originality/value – This paper tries to project population distribution by modeling geo-homophily in OSNs.

Keywords Dirichlet process, Geo-homophily, Population distribution

Paper type Research paper

1. Introduction

The study of population distribution in fixed geographical regions (e.g. states, provinces) is of paramount importance for the government to enhance the public safety in certain places with a large floating population (FP) or for the business to launch marketing campaigns in densely populated areas (Harris and Todaro, 1970). The data relevant to people's trajectory are conventionally collected through offline sources. For instance, it is feasible to predict the FP in the transportation systems by analyzing the origin-destination data of passengers (Chi, 2010); Customers' bank notes can be used for modeling the human trajectories as a continuous-time random walk (Gonzalez *et al.*, 2008). The government (e.g. statistical bureau) collects the demographic data to investigate the correlation between human migration patterns and geographic labor demand and supply (National Bureau of Statistics of China, 2010). The challenges of conducting these studies are attributed to the high cost of the data collection methods (regarding time, manpower and money), and the restriction of accessing such data sets due to security and privacy concerns.

The wide use of online social network (OSN) apps over smartphones has accumulated a rich set of geographical data that describe anonymous user trajectories (Kido *et al.*, 2005) and habits (Gao *et al.*, 2017) in the physical world, which holds the promise of providing a lightweight means to study the population distribution (Li *et al.*, 2016). For example, many OSN applications, such as Facebook, Weibo, allow users to "check-in" and explicitly show their locations (Guo *et al.*, 2013; Nazir *et al.*, 2008; Liao *et al.*, 2015); some other applications have implicitly recorded users' geo-related information such as GPS coordinates, IP address (Backstrom *et al.*, 2010; Zheng *et al.*, 2010). Existing research has shown the feasibility of using the OSN data to predict users' offline locations as well their mobility patterns (Cho *et al.*, 2011; Li *et al.*, 2010). Moreover, the online relationship between friends can affect their social ties in the physical world (Zheng, 2012): the "close" friends in the OSN are also physically close to each other (Cho *et al.*, 2011).

However, it is still unclear that which type of OSNs can assist determining the population distribution in given geographical regions. Intuitively, there are two observations:

(1) It is easy to draw a population distribution over geographical regions that are *stable* – most people in a region do not travel distantly; and

(2) Acquaintances in the same geographical region have a strong desire to communicate with each other through the OSN (Girvan and Newman, 2002).

Therefore, we seek to answer the following questions in this study:

- Is there a way of measuring the stability of geographical regions by observing the online message diffusion among people in those regions?
- How to derive the offline population distribution over a stable division of regions?
- Given a population distribution, how to project the FP across regions?

Our research findings indicate that a division of geographical regions is stable only if the OSN users in these divided regions show a strong geo-homophily; people in each region

prefer communicating with others in the same region more than those in other regions, and the Dirichlet process (DP) (Neal, 2000) provides a viable way of modeling the distribution of OSN users across offline regions. These inspire us to investigate the relationship between the online information diffusion, i.e. users' communication in OSN, and the population distribution over a fixed division of offline regions.

In this paper, we present a systematic approach that projects the offline population distribution in fixed geographical regions by modeling the geo-homophily of OSNs. Specifically, we establish a three-layered theoretic framework that first maps the online message diffusion among friends in the OSN to the offline population distribution via a DP and then projects the FP across geographical regions given the derived population distribution. The contributions of this work are summarized as follows:

- *Connecting online data to stability of geographical regions*: We establish the correlation between online message diffusion and the stability of geographical regions by modeling the geo-homophily of an OSN with geographical attributes. We derive the condition for a division of geographical regions to have a non-decreasing stability.
- *DP-based prediction models*: We formulate the population distribution problem from the perspective of DP and present a theoretical framework to project the population distribution over fixed geographical regions by casting online message diffusion into the established framework. Based on the derived population distribution, we propose a prediction model that utilizes the message diffusion graph in OSNs to infer the FP across geographical regions.
- *Experiments using large real-world data sets*: By experiments over the real-world data sets, we validate the efficacy of the model in projecting the population distribution over fixed regions and meanwhile show that the proposed prediction models have a high prediction accuracy in characterizing the process of how the FP changes across regions upon the occurrence of societal events (the mass human migration caused by the Chinese Spring Festival 2016).

The rest of this paper is organized as follows. We introduce the related work and technical background on DP in Section 2. We provide the system model and formulate the problem in Section 3. In Sections 4 and 5, we show the approach of projecting the population distribution and present the model of predicting the FP, respectively. We validate the proposed model by experiments in Section 6, and conclude the paper in Section 7.

2. Related work

2.1 Geographical views of online social networks

Prior work on geographical aspects of OSNs has mostly focused on prediction and analytics of various properties in OSN by leveraging the location-related information.

2.1.1 Predicting mobility patterns using online social network data. Users' locations can be predicted by mining their periodic behaviors in social network, given that the observed movement is associated with certain reference locations (Li *et al.*, 2010). Cho *et al.* (2011) show that human movement and mobility patterns have a high degree of freedom and variation, but they can still exhibit structural patterns due to geographical and social constraints, on basis of two observations:

(1) short-ranged travel is periodic both spatially and temporally and not effected by the social network structure; and

(2) long-distance travel is more influenced by social network ties.

Thus, the historic data can be used to predict where a user might travel.

2.1.2 Data dissemination in a geographical perspective. Wang *et al.* (2014) pose a three-layered architecture to model the data dissemination in OSNs, present a density function of general social relationship distribution and derive the tight lower bounds on traffic load of data dissemination in the OSNs under the assumption that every source sustains a data generating rate of a constant order.

2.1.3 Online and offline social behaviors. Zheng (2012) propose a location-based social network (LBSN), which consists of the new social structure made up of individuals connected by the "interdependency" derived from their locations in the physical world as well as their location-tagged media content, such as photos, video and texts. Hristova *et al.* (2014) experimentalized on a data set with 74 college students as volunteers by observing evidence of homophily with regard to many factors within the online and offline social networks. They found that the social tie among students at the same educational institution was strongly affected by residential sector and year in college, but it exhibited diversity in other online aspects, leading to the affirmation saying diversity online is relative to diversity offline.

2.1.4 Social tie inference. Sociological phenomena can be also observed within OSNs. Although the OSN platform has facilitated people's communication, the volume of OSN communications between OSN friends (the strength of the social tie between them) is inversely proportional to the geographical distance, following a Power Law (Goldenberg and Levy, 2009). Considering the co-occurrence in time and space (Crandall *et al.*, 2010), Crandall *et al.* (2010) present a probabilistic model to prove that even a very small number of co-occurrences can result in a high empirical likelihood that the two people know each other – a social tie between them, which tells us a way to infer the social network structure only by capturing individual physical location over time.

2.2 Dirichlet distribution and Dirichlet process

Dirichlet distribution is the conjugate prior of multinomial distribution, which can be seen as a distribution over distribution. The probability density function is written as:

$$p = (P = \{p_i\}|\alpha_i) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma\left(\sum_i \alpha_i\right)} \prod_i p_i^{\alpha_i - 1}.$$

There are two parameters:

(1) The scale $\alpha = \sum_i \alpha_i$: a small scale α favors extreme distributions, but this prior belief is very weak and is easily overwritten by data, whereas an extremely large α makes the samples be more consistent with the base measure.

(2) The base measure $(\alpha_1', \alpha_2', \ldots), \alpha_i' = \alpha_i/\alpha$: The base measure determines the mean distribution.

One popular application of Dirichlet distribution is latent Dirichlet allocation on topic discovery in natural language processing. It is a generative statistical model aiming at describing sets of observations by connotative groups why some parts of the data are similar.

DP is a class of Bayesian nonparametric models, and DP generalizes Dirichlet distribution (Neal, 2000). DP is a distribution function in a space of infinite but countable number of elements, which also requires a scale parameter α and a base measure G_0,

denoted as DP(α, G_0). DP is an important method in Bayesian inference to identify the prior distribution of random variables, and it is widely used for density estimation, semiparametric modeling and sidestepping model selection/averaging. One important implication is that DP helps find the number of active components which is much less than the number of samples. In this paper, we investigate how to use DP to model the process that OSN users are distributed into geographical regions.

3. System model

In this section, we propose a three-layer framework that analyzes the message diffusions in the OSN to determine the stability of geographical regions. This problem is equivalent to the determination of whether the OSN has a strong geo-homophily – more specifically, whether the structure of the message diffusion graph is similar to that of the divided regions. We extend the concept of modularity (Newman, 2004) to quantify the degree of the geo-homophily of an OSN, and meanwhile we specify the condition on the geo-homophily of an OSN for the stability of underlying geographical regions to remain non-decreasing.

In Figure 1, we show a three-layered framework consisting of: Layer 1 that captures the message diffusion graph in an OSN; Layer 2 that seeks to derive the user population distribution from the geo-location of OSN users in Layer 1; and Layer 3 that predicts how the FP will change given the distribution derived in Layer 2.

From the top to the bottom layer, we first investigate how the messages diffuse among groups of people that have similar geo-locations. If people in the same geo region communicate frequently, it is highly likely that the structure of the message diffusion graph is similar to that of the underlying division of regions – the strong geo-homophily exists between the OSN and the offline regions. As a result, we can use the geo-location of messages among OSN users to derive the user distribution over the given regions. Then, the FP across regions can be further inferred based on the derived distribution.

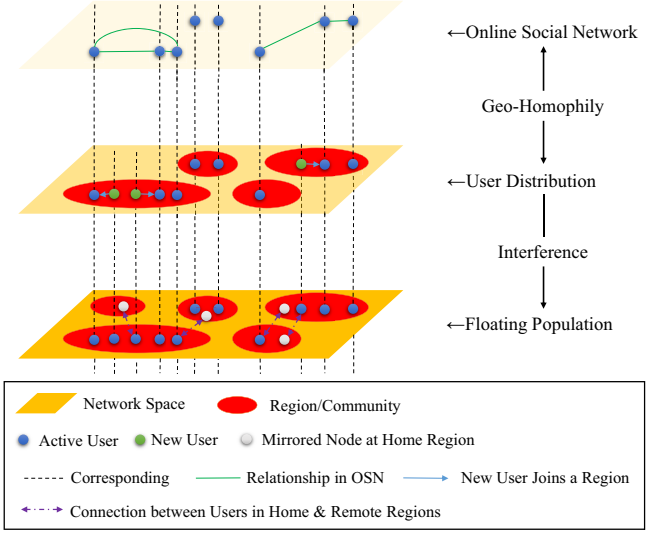

Figure 1. A three-layered analytical framework that defines the geo-homophily of OSNs to map the OSN message diffusion (Layer 1) to the offline user distribution (Layer 2) and infers the FP (Layer 3) based on the derived distribution (Layer 2)

3.1 Geo-homophily of an online social network over divided regions

We define the *geo-homophily of an OSN* as the degree of similarity between the structure of the message diffusion graph in the OSN and that of a given division of regions.

We calculate *modularity* to quantify the geo-homophily of an OSN. Given the message diffusion graph of an OSN $G = (V, R, E, T)$, where V denotes the set of users, R denotes the set (or division) of regions and E denotes the set of edges e_{uv} with weights ρ_{uvT}. The weight ρ_{uvT} represents the number of views from user u to the content sharing by user v during time period T. The e_{uv} exists if the number of views from user u to the content sharing by user v during time period T is non-zero. Let Ω_T be $\sum_{u,v} \rho_{uvT}$. Each user $u \in V$ belongs to a specific region $r \in R$ given the division R, denoted as r_u.

We can easily transform E from a user-to-user perspective to a region-to-region one, recorded as ε, where $\forall \epsilon_{ij} \in \mathscr{E}$ has value:

$$\omega_{ijT} = \sum_{u \in V, r_u = i} \sum_{v \in V, r_v = j} \rho_{uvT}$$

which represents the number of views from nodes in region i to the content of nodes in region j during time period T.

Let p_{ijT} be the proportion of messages from i to j during T, namely, $\omega_{ijT} / \sum_{i',j'} \omega_{i'j'T}$. To quantify the geo-homophily of an OSN $G = (V, R, E, T)$, we define the modularity on R during T, Q_{RT}, as:

$$Q_{RT} = \sum_{i \in R} \left(p_{iiT} - \sum_{j \in R} p_{ijT} \sum_{j \in R} p_{jiT} \right) \tag{1}$$

It reflects the centrality of messages that are transmitted within same regions. Apparently, Q_{RT} ranges in $[-1, 1]$, where Q_{RT} reaches 1 if all diffusions take place inside the same regions and reaches -1 when none of the messages are transmitted between users from same regions. The greater Q_{RT} is, the higher geo-homophily the OSN has.

If an OSN shows a strong geo-homophily over the divided regions, most OSN users have more preference of communicating with others in the same region rather than with those in other regions, which implies each user is more attached/attracted by his current region instead of other regions, thereby leading to a high stability of each region. Next, we will show how to determine the change of the stability of a division R by imposing a condition over Q_{RT}.

3.2 Stability of a division of regions

The modularity quantifies the geo-homophily between an OSN and the underlying geographical regions. However, it is infeasible to foresee whether the regions will remain stable because the structure of the message diffusion graph is dynamically changing. For instance, a breaking news may reform the structure of the message diffusion graph and push people to move across regions, which may make the stability of the divisions vulnerable. Next, we will deduce: under what condition, the stability of a division of regions will remain non-decreasing.

Formally, given two time periods $T = [t_0, t_1]$ and $T' = [t_0, t_2]$, where $t_2 > t_1$, we need to find the distribution of messages in period $[t_1, t_2]$ that leads to an equal or higher modularity in at the end of T', i.e. $Q_{RT} \leq Q_{RT'}$.

We define the social-entropy of message diffusion inside and outside regions in the message diffusion graph G as:

$$H(G) = -\sum_{i \in R} \sum_{j \in R} p_{ijT} \times \log p_{ijT} \tag{2}$$

As the redistribution of message diffusion inside each region do not affect the modularity according to equation (1), we will only focus on those message diffusions (edges) across regions.

Hence, we combine all edges within a region into a new set. Let $I_T = \bigcup_{i \in R} \epsilon_{iiT}$, and $\omega_{I_T} = \sum_{i \in R} \omega_{iiT}$, $p_{I_T} = \omega_{I_T}/\Omega_T$. $H(G)$ can be rewritten as:

$$H(G) = -p_{I_T} \times \log p_{I_T} - \sum_{i \in R} \sum_{j \in R, i \neq j} p_{ijT} \times \log p_{ijT} \tag{3}$$

New message diffusions in time period $[t_1, t_2]$ will create new edges and construct a new message diffusion graph G' (that can be extended from G). Let l_{ij} be the number of new edges from region i to j in G', which are not included in G. Note that $\forall i, j \in [1, |R|], l_{ij} \geq 0$.

Let $\mathbf{L}_{|R| \times |R|}$ be the matrix of l_{ij}, and $\mathcal{L} = \sum_{i,j} l_{ij}$ where $\mathcal{L} \ll \Omega_T$). Let $l_I = \sum_{i \in R} l_{ii}$ be the number of new edges inside regions.

To measure the impact and the change to G caused by new message diffusion $\overline{\mathbf{L}}$, we define *Information Increment*, $\mathcal{G}(G, \overline{\mathbf{L}})$, as follows:

$$\mathcal{G}(G, \overline{\mathbf{L}}) = \frac{(\omega_{I_T} + l_I)^{(\omega_{I_T} + l_I)} \prod_{i,j \in R, i \neq j} (\omega_{ijT} + l_{ij})^{(\omega_{ijT} + l_{ij})}}{\omega_{I_T}^{\omega_{I_T}} \prod_{i,j \in R, i \neq j} \omega_{ijT}^{\omega_{ijT}}} \tag{4}$$

According to equation (3), the social-entropy becomes:

$$\begin{aligned} H(G') = &-\frac{\omega_{I_T} + l_I}{\mathcal{L} + \Omega_T} \log \frac{\omega_{I_T} + l_I}{\mathcal{L} + \Omega_T} + \Omega_T \\ &-\sum_{i \in R} \sum_{j \in R, i \neq j} \frac{\omega_{ijT} + l_{ij}}{\mathcal{L} + \Omega_T} \log \frac{\omega_{ijT} + l_{ij}}{\mathcal{L} + \Omega_T} + \Omega_T \end{aligned} \tag{5}$$

The following proposition prescribes the condition for the stability of divided regions to remain non-decreasing, based on the analysis of the OSN message diffusion graph.

P1. Given a message diffusion graph G over a division of regions, the geo-homophily will not decrease, if $\mathcal{G}(G, \overline{\mathbf{L}})$ is no smaller than $\Omega_T^{\mathcal{L}}$, where $\mathcal{L} \ll \Omega_T$.

Proof. The degree of the geo-homophily of an OSN will not decrease if the social entropy never had a tendency to increase – i.e. ΔH is non-positive, where:

$$\begin{aligned} \Delta H &= H(G') - H(G) \\ &= -\frac{\omega_{I_T}}{\Omega_T} \log \frac{\omega_{I_T} + l_I}{\omega_{I_T}} - \sum_{i,j \in R, i \neq j} \frac{\omega_{ijT}}{\Omega_T} \log \frac{\omega_{ijT} + l_{ij}}{\omega_{ijT}} \\ &\quad - \frac{l_I}{\Omega_T} \log \frac{\omega_{I_T} + l_I}{\Omega_T} - \sum_{i,j \in R, i \neq j} \frac{l_{ij}}{\Omega_T} \log \frac{\omega_{ijT} + l_{ij}}{\Omega_T} \end{aligned}$$

That is:

$$\Delta H = \frac{1}{\Omega_T}\left(\log\Omega_T^{\mathcal{L}} - \log\mathcal{G}(G,\overline{\mathbf{L}})\right) \tag{6}$$

Then we can substitute $\mathcal{G}(G,\overline{\mathbf{L}}) \geq \Omega_T^{\mathcal{L}}$ into equation (6) and we could conclude with this proposition. □

4. Population distribution projection

Given a division of regions, the geo-homophily is an indicator of the similarity between the structure of the OSN message diffusion graph and that of the division. The stronger geo-homophily an OSN has, more in-region communications occur between friends in the same region rather than across-region communications. Whenever a new user joins the OSN, he/she is highly likely to be distributed to the region where most of his/her friends reside. This is similar to the Chinese restaurant process (one representation of DP), which describes how guests are assigned to different tables in the restaurant according to the existing guest distribution.

In this section, we present a Bayesian nonparametric model based on the DP, which predicts how users in a OSN with strong geo-homophily are distributed over a given division of regions. In contrast, the weak geo-homophily in the OSN over given regions fails to establish the link between OSN message diffusion and the user distribution, which leads to a low prediction accuracy.

4.1 User distribution model

We propose a user distribution model (UDM) on basis of the Dirichlet process mixture (DPM) model for learning the hyper-parameters of the gathering mode, which is defined as a distribution of a random probability measure u. A UDM has two parameters: base distribution u_0 which is considered as the mean of DP and the scale parameter α which is like an inverse-variance of the DP. Then we have:

$$\mathcal{U} \sim \mathrm{UDM}(\alpha, \mathcal{U}_0)$$

representing a draw of a random probability measure u over a given parameter space U from the corresponding DP. For every user $u \in V$, we can draw a relevant θ_u from u. Here, α affects the probability that $\theta_u = \theta_v$, $u \neq v$. Thus, sampling from UDM is executed by the following generative process:

$$\begin{aligned} &\mathcal{U} \sim \mathrm{UDM}(\alpha, \mathcal{U}_0) \\ &\theta_u \sim \mathcal{U} \\ &r_u \sim F(\theta_u) \end{aligned}$$

where F is the likelihood function determining which region user u belongs to. Due to the cluster property, the number of distinct θ's would be exactly $|R|$, far less than $|V|$. Let $\tilde{\theta}_r, r \in R$ be the non-redundant hyper-parameters.

We have u in $|R|$ dimensions where $\sum_{r\in R}\alpha_r = \alpha$, i.e.:

$$\mathcal{U} \sim \mathrm{Dir}\left(\{\alpha_r\}_{r\in R}\right)$$

Define n_r be the amount of r_u that equals to r for every user u, and we can deduce the posterior distribution as:

$$
\begin{aligned}
&P\left(\{\tilde{\theta}_r\}_{r\in R}|\{n_r\}_{r\in R}\right)\\
&\propto \text{Mult}\left(\{n_r\}_{r\in R}|\{\tilde{\theta}_r\}_{r\in R}\right)\text{Dir}\left(\{\tilde{\theta}_r\}_{r\in R}|\{\alpha_r\}_{r\in R}\right)\\
&\propto \prod_{r\in R}\tilde{\theta}_r^{\alpha_r-1}\frac{\Gamma\left(\sum_{r\in R}\alpha_r\right)}{\prod_{r\in R}\Gamma(\alpha_r)}\prod_{r\in R}\tilde{\theta}_r^{n_r}\\
&\propto \prod_{r\in R}\tilde{\theta}_r^{\alpha_r-1}\prod_{r\in R}\tilde{\theta}_r^{n_r}=\prod_{r\in R}\tilde{\theta}_r^{n_r+\alpha_r-1}\\
&=\text{Dir}\left(\{\tilde{\theta}_r\}_{r\in R}|\{\alpha_r+n_r\}_{r\in R}\right)
\end{aligned}
$$

Thus, the marginal probability would be:

$$
\begin{aligned}
&P\left(\{n_r\}_{r\in R}\right)\\
&=\int_{\{\tilde{\theta}_r\}_{r\in R}}P\left(\{\tilde{\theta}_r\}_{r\in R}|\{n_r\}_{r\in R}\right)\\
&=\frac{\Gamma\left(\sum_{r\in R}\alpha_r\right)}{\prod_{r\in R}\Gamma(\alpha_r)}\int_{\{\tilde{\theta}_r\}_{r\in R}}\prod_{r\in R}\tilde{\theta}_r^{n_r+\alpha_r-1}\\
&=\frac{\Gamma\left(\sum_{r\in R}\alpha_r\right)}{\prod_{r\in R}\Gamma(\alpha_r)}\frac{\prod_{r\in R}\Gamma(\alpha_r+n_r)}{\Gamma\left(|V|+\sum_{r\in R}\alpha_r\right)}
\end{aligned}
$$

According to the Bayesian theory, for user $u \notin V$, the predictive distribution becomes:

$$
\begin{aligned}
P\left(r_u=r|\{r_v\}_{v\in V}\right)&=\frac{P\left(r_u=r,\{r_v\}_{v\in V}\right)}{P\left(\{r_v\}_{v\in V}\right)}\\
&=\frac{P\left(\{n_r+1\}\cup\{n_{r'}\}_{r'\in R,r'\neq r}\right)}{P\left(\{n_{r'}\}_{r'\in R}\right)}\\
&=\frac{\Gamma\left(|V|+\sum_{r\in R}\alpha_r\right)}{\Gamma\left(|V|+1+\sum_{r\in R}\alpha_r\right)}\frac{\Gamma(\alpha_r+n_r+1)}{\Gamma(\alpha_r+n_r)}\\
&=\frac{n_r+\alpha_r}{|V|+\sum_{r\in R}\alpha_r}
\end{aligned}
$$

4.2 A special case of Chinese restaurant process

The process of distributing users over multiple regions is a special case of Chinese restaurant process (Aldous, 1993), given that $|R|$ is finite. Whenever a new user joins the OSN, he/she needs to choose a region to stay, by considering the distribution of his/her friends in the given regions:

- When the OSN has a strong geo-homophily over the regions, people prefer to communicate and stay with their friends in the same region.
- When the OSN owns a weak geo-homophily, users may communicate with online friends in a region but stay with offline acquaintances in another different region.

4.2.1 Parameters in the view of stick-breaking representation. Although n_r's are statistic variables that can be obtained directly, the scale parameters are not easy to compute. To avoid manual assignment of α_r, we change our view of the problem to be an equivalent one, i.e the stick-breaking representation.

The posterior distribution of u over $\tilde{\theta}$ is deduced as:

$$\begin{aligned} & P\left(\mathcal{U}|\{\tilde{\theta}_r\}_{r\in R}\right) \\ & \propto P\left(\{\tilde{\theta}_r\}_{r\in R}|\mathcal{U}\right)P(\mathcal{U}) \\ & = \mathcal{U}P(\mathcal{U}) \end{aligned}$$

So we have:

$$\mathcal{U}|\{\tilde{\theta}_r\}_{r\in R} \sim \mathrm{UPM}\left(\alpha+1, \frac{\alpha\mathcal{U}_0+\delta_{\tilde{\theta}}}{\alpha+1}\right)$$

where $\delta\tilde{\theta}$ is a probability measure concentrated at $\tilde{\theta}$.

Consider a partition $(\theta', U\backslash\theta')$, we have:

$$\begin{aligned} & \left(\mathcal{U}(\theta'), \mathcal{U}\left(U\backslash\theta'\right)\right) \\ & \sim \mathrm{Dir}\left((\alpha+1)\frac{\alpha\mathcal{U}_0+\delta_{\tilde{\theta}'}}{\alpha+1}(\theta'), (\alpha+1)\frac{\alpha\mathcal{U}_0+\delta_{\tilde{\theta}'}}{\alpha+1}\left(U\backslash\theta'\right)\right) \\ & = \mathrm{Beta}(1,\alpha) \end{aligned}$$

Serialize each region from 1 to $|R|$, and the stick-breaking procedure is then deduced by:

$$\begin{aligned} & \mathcal{U} \sim \mathrm{UPM}(\alpha, \mathcal{U}_0) \\ & = \beta_1\delta_{\tilde{\theta}_1} + (1-\beta_1)\mathcal{U}_1 \\ & = \ldots \\ & = \sum_{i=1}^{|R|}\pi_i\delta_{\tilde{\theta}_i} \end{aligned}$$

where $\beta_i \sim \mathrm{Beta}(1,\alpha)$ for $i \neq |R|$ and $\beta_{|R|} = 1$, whereas:

$$\pi_i = \left(1 - \sum_{j=1}^{i-1}\pi_j\right)\beta_i$$

The posterior distribution of β_i satisfies

$$p(\beta_r|\{n_{r'}\}_{r'\in R},\alpha) \propto p(n_r,|V||\beta_r)p(\beta_r|\alpha)$$

5. Floating population inference

In the physical world, people may move across regions periodically or temporally, thereby greatly influencing the geo-homophily of the OSN they use. In general, there are two important regions for every person, that is, the *home* region denoted as $\mathscr{H}$, and the *remote* region denoted as ? (e.g. the work place). According to the previous study (Cho *et al.*, 2011), most of the message diffusions usually occur in or between these two regions (e.g. an OSN user in the remote region contacts his families at home region or his colleagues at the same remote region). With these observations, we leverage the geo-attributes of message diffusions between the sender and receiver to infer the distribution of FP.

5.1 Distribution of message diffusions

We use a tetrad $S = (\mathbb{C},\mathbb{Q},\lambda,\chi)$ to represent the state of the message diffusion graph. Consider a state when the population distribution is captured as $\mathbb{C} = \{c_i\}_{i\in R}$, where c_i represents the proportion of the population of region i.

Denote the real population distribution as $\mathbb{Q} = \{q_{ij}\}_{i,j\in R}$, where q_{ij} means the proportion of people whose remote region is region j, whereas their home region is region i. We have $c_i = \sum_{j\in R} q_{ji}$... Let σ_{ij} be the proportion of users in j with home region i, i.e. $\sigma_{ij} = \frac{q_{ij}}{c_j}$. Similar to UDM, σ_{ij}'s in a specific region can also be generated from a DP.

Given a sender region, the amount of region-to-region communication is proportional to the population of the receiver region. Then for every receiver region r, we have:

$$P(r|r_{\mathcal{H}},r_{\mathcal{W}}) = \begin{cases} \lambda + (1-\lambda-\chi)c_r & r = r_{\mathcal{H}}, r_{\mathcal{H}} \neq r_{\mathcal{W}} \\ \chi + (1-\lambda-\chi)c_r & r = r_{\mathcal{W}}, r_{\mathcal{H}} \neq r_{\mathcal{W}} \\ \lambda + \chi + (1-\lambda-\chi)c_r & r = r_{\mathcal{H}} = r_{\mathcal{W}} \\ (1-\lambda-\chi)c_r & otherwise \end{cases}$$

where λ is the proportion of communications with the home region, and χ is the proportion of communications with the remote region.

State difference. Define a baseline state $S' = (\mathbb{C}',\mathbb{Q}',\lambda,\chi)$, $\mathbb{C}' = \{c'_i\}_{i\in R}$, where all people stay at their home regions, i.e. $\forall i,j \in R, i \neq j$, the corresponding $\sigma'_{ij} = 0$. Consider the difference between an arbitrary state S and the baseline state, named as *state difference* ΔS.

P2. The state difference follows a superposition of a uniform distribution and a Dirichlet distribution. Proof. The proportion of messages from $r_?$ to r_τ should be:

$$
\begin{aligned}
&P_S(r_{\mathcal{T}} = j | r_{\mathcal{O}} = i) \\
&= \sum_{r_{\mathcal{H}} \in R} P(j | r_{\mathcal{H}}, r_{\mathcal{W}} = i) \boldsymbol{\sigma}_{r_{\mathcal{H}} i} \\
&= (1 - \lambda - \chi) c_j + \lambda \boldsymbol{\sigma}_{ji} + [j = i] \chi \boldsymbol{\sigma}_{ii}
\end{aligned}
$$

Therefore, we can deduce that:

$$
\begin{aligned}
&P_{\Delta S}(r_{\mathcal{T}} = j | r_{\mathcal{O}} = i) \\
&= P_S(r_{\mathcal{T}} = j | r_{\mathcal{O}} = i) - P_{S'}(r_{\mathcal{T}} = j | r_{\mathcal{O}} = i) \\
&= \begin{cases} (1 - \lambda - \chi)(c_j - c_{j'}) + \lambda \boldsymbol{\sigma}_{ji} & i \neq j \\ (1 - \lambda - \chi)\left(c_j - c'_j\right) + (\lambda + \chi)(\boldsymbol{\sigma}_{ii} - 1) & i = j \end{cases}
\end{aligned} \tag{7}
$$

which is a constant plus a variable generated from DP. It indicates that the state difference follows a superposition of a uniform distribution and a Dirichlet distribution.

This proposition enlightens us to infer FP by methods of divide and conquer. The state difference reduces the weight of the uniform distribution component.

5.2 Export message pattern

Similar to UDM, we can extract the distribution of messages diffused to remote regions, and we use a Hierarchy DP to find the distribution, which is named as the export message pattern (EMP). For every region i, denote $\boldsymbol{\sigma}_{\mathbf{i}}$ as $\{\boldsymbol{\sigma}_{ji}\}_{i \neq j}$, following:

$$
\begin{aligned}
B_0 &\sim \mathrm{DP}\left(\tau', B'\right) \\
B_i &\sim \mathrm{DP}(\tau_i, B_0) \\
\eta_i &\sim B_i \\
\boldsymbol{\sigma}_{\mathbf{i}} &\sim F(\eta_i)
\end{aligned}
$$

where η_i is the hyper-parameters, τ_i and τ' is the corresponding scale parameter and B' is the base distribution. Consider the differential export message:

$$
\mathbf{d_i} = \{d_{ij} = P_{\Delta S}(j|i) - (1 - \lambda - \chi)\left(c_j - c'_j\right)\}_{i \neq j}
$$

which satisfies that:

$$
\mathbf{d_i}/\lambda \sim F(\eta_i)
$$

Given $\mathbf{d_i}$, Gibbs Sampling can be used to decide what $\boldsymbol{\sigma}_{\mathbf{i}}$ should be.

5.3 Self message pattern

The DPM can also explain the distribution of messages diffused inside each region, which is named as self message pattern (SMP). According to equation (7), it is not wise to gather $\boldsymbol{\sigma}_{ii}\ \forall i \in R$. Instead, we should concern $\{\boldsymbol{\sigma}_{0i} = 1 - \boldsymbol{\sigma}_{ii}\}_{i \in \mathrm{R}}$ and denote it as $\boldsymbol{\sigma}_0$. We are able to find a scale parameter τ_0 and base distributionI_0 such that:

$$I \sim \mathrm{DP}(\tau_0, I_0)$$
$$\eta_0 \sim I$$
$$\boldsymbol{\sigma_0} \sim F(\eta_0)$$

Because we have access to:

$$\mathbf{d_0} = \{d_{0i} = (1 - \lambda - \chi)(c_i - c_i') - P_{\Delta S}(i|i)\}$$

following $\mathbf{d_0}/(\lambda + \chi) \sim F(\eta_0)$, the model can be solved by Gibbs sampling according to the posterior distribution and the restriction holding $\sum_j \sigma_{ji} = 1$.

5.4 Floating population inference model

Finally, we combine UDM, EMP and SMP as a floating population inference model (FPIM). The UDM provides the population distribution across regions, whereas EMP and SMP compute the specific allocation inside each region. The model structure is shown in Figure 2.

6. Evaluation

In this section, we validate the geo-homophily over two real-world OSN data sets that have geo attributes of users and evaluate the performance of proposed UDM and FPIM models.

6.1 Data sets

We use data sets of two OSNs: Gowalla data set and WeChat Moment data set. The former one covers most of western countries, whereas the latter covers the China mainland (where Internet censorship is enforced and people have restricted access to popular OSN sites/apps like Facebook).

"Gowalla" (Cho *et al.*, 2011) is a typical LBSN where users share their locations by "checking-in". The information regarding friend relationship was collected using their public API, which consists of 196,591 nodes and 950,327 edges. The edges can be seen as undirected. This Gowalla data set collects a total of 6,442,890 check-ins of these users over the time period from February 2009 to October 2010.

"WeChat Moments (WM)" (Schiavenza, 2013) is the social network of a mobile messaging app (Wechat) popular in China, where the contents shared over WM are HTML5 pages

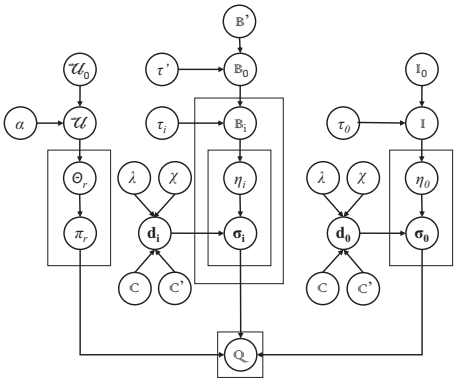

Figure 2. The FPIM has three parts, namely, UDM, EMP, SMP from left to right

(Zhang *et al.*, 2016). This WM data set contains 137,509,889 users with 1,671,692,424 retweeting/forwarding records of 329,465 pages from January 14, 2016 to February 27, 2016, telling us when, where, from whom a page is re-tweeted; how many pages a user reads; and whether one has re-tweeted a page. WM can only be used on mobile devices, and the user location can be inferred from the IP address. The period of data covers Spring Festival, a traditional festival in China when most of Chinese people migrate back to their home province from the work place.

Note that although the number of users of each data set is much less than the population of a country, it is sufficiently large to derive the proportion of OSN user distribution, as well as the population distribution over geographical regions, which helps us determine how a new OSN user is distributed or how FP varies across regions.

6.2 Geo-homophily of online social networks

As mentioned in Section 3, we can divide users of an OSN into a division of regions, according to users' geo attributes.

6.2.1 Geo-homophily of WeChat

6.2.1.1 Message diffusion in China. The WM data set records the page re-tweeting in 34 provinces in China, and we use these provinces as the geographical regions in this experiment. Every user in WM should have viewed a collection of pages, and each page view's IP address corresponds to a province, among which the most frequently recorded one is set as the province where the user is located. We analyze the message diffusion process in two time periods:

(1) Before Spring Festival, we monitor the message diffusion from January 14 to January 31, 2016, which are pre-holiday working and weekend days.

(2) On the Spring Festival day, most people stay at home, and hence the structure of the message diffusion graph would be different.

6.2.1.2 Results for pre-holidays. The modularity in the pre-holidays is approximately 0.49. Figure 3(a) shows the volume of message diffusion inside each province and that between

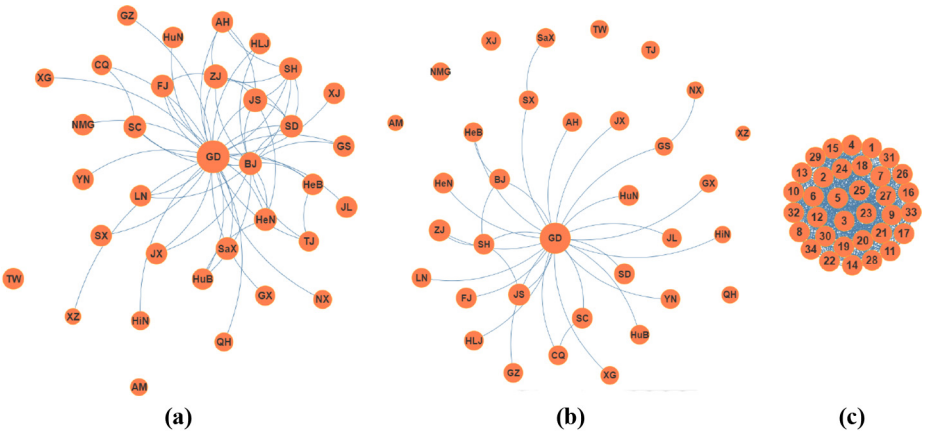

Figure 3.
Graph representations of message diffusion inside and across provinces in WM data set

Notes: (a) Before Spring Festival; (b) on the Spring Festival day; (c) baseline network

every pair of provinces of China, where the amount of message diffusion inside a province is proportional to the size of the corresponding circle, and the amount between provinces are represented by the length of arcs:

- The larger size an orange circle has, more friend relationships between two people exist inside the province labeled in the circle.
- The shorter a blue arc is, more friend relationships exist between people in two different provinces whose corresponding circles are connected by the arc.
- For a province whose corresponding circle have no arc connected, only a very small number of friend relationships exist between this province and another different one, and the resulting arc could be very long in proportion to the length of other arcs, and thus we skip plotting such very-long arcs in the figure.

The results indicate that most of the diffusions occur inside provinces, so the arcs are relatively sparse. In particular, there lies no arc between some pairs of circles in this figure, which does not mean that there is no message diffusion between the corresponding two provinces but implies that the message diffusion between them is much weaker than that between those pairs of circles having arcs. For example, there were only hundreds of message diffusions between Tibet and Taiwan in the data set; in contrast, several millions of message diffusions occur between Beijing and Guangdong. This can be explained by the fact that those provinces are at distant locations, or they have little communication with most provinces in China mainland.

6.2.1.3 Results on holiday. On the Spring Festival holiday, most people stay with their families in their home province. The graph structure changes, as most of the messages are sent for appointments and greetings, and these diffusions mainly took place between friends in the same vicinity. Thus, the proportion of the diffusion inside regions increases, leading to a modularity of 0.53. The graph structure is illustrated in Figure 3(b), where some inter-province arcs disappear.

6.2.1.4 Results for the baseline graph. However, results become quite different if we use the baseline graph where edges of diffusion graph during pre-holidays were placed at random (Newman, 2004). Then we obtain a chaotic segregation with the modularity of −0.05, which can hardly be said to have any geo-homophily. The amount of message diffusions inside and across these regions are shown in Figure 3(c). Compared to Figure 3(a), under the same scale of plotting, the distribution of circles representing regions are very dense and most of circles' sizes are similarly small.

6.2.2 Geo-homophily of Gowalla. The Gowalla data set provides latitudes and longitudes of check-ins, involving the users as well as the friend relationship among them. Over 80 per cent of the users are Americans or Europeans. Here, we use 50 states and a federal district of USA as the geographical communities. Because we find that the distribution of check-ins within the USA is approximately proportional to each state's population[1], we conclude that there lies a certain degree of geo-homophily in Gowalla data set, and the corresponding modularity is 0.34.

6.3 Stability of divided regions

When considering the diffusion of a single page, we find that it will be reposted many times in the home region of the sender, whereas it may be sent to only a few non-home regions – those diffusions across regions only take up a small part. For example, we illustrate the distribution of views to a popular page with approximately one million views in Figure 4, where the page is originally sent from the region of Beijing.

Recall that we propose Information Increment to measure the change of geo-homophily between two time points in Section 3.2. To test its impact, we simulate eight instances of message diffusions and add them sequentially to the message diffusion graph at the end of Jan.

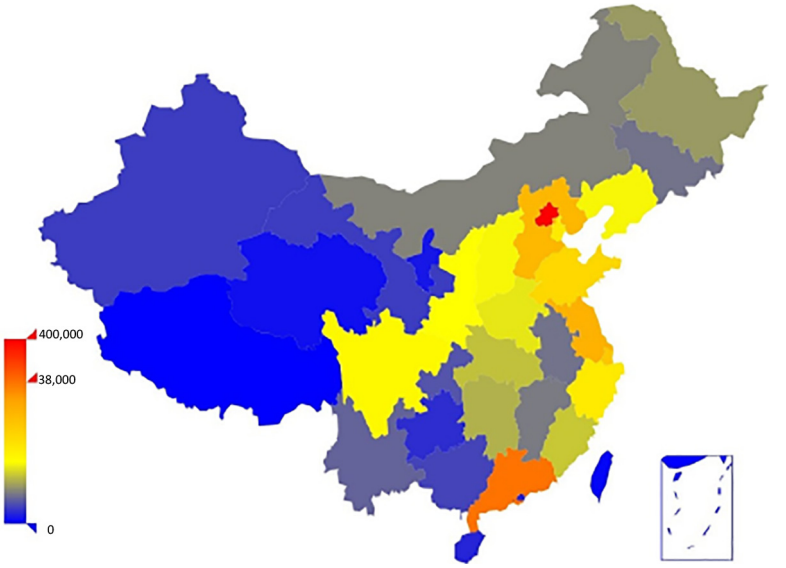

Figure 4.
The viewing distribution of a hot message originated from Beijing

31 of the WM data set. For each simulated message diffusion, first we construct a Dirichlet distribution according to the previously formed message distribution by randomly selecting a province among 34 provinces as the sender region, from which we can obtain a multinomial distribution. One million retransmissions are then sampled through the multinomial distribution. The experimental results are shown in Figure 5, where the geo-homophily will not decrease when $\mathcal{G}(G, \overline{\mathbf{L}}) \geq \Omega_T^{\mathcal{L}}$ and vice versa. This implies that the stability will not decrease when the previously mentioned condition is satisfied, which is consistent with Proposition 1.

6.4 Performance of user distribution model

Given the order of users' joining the OSN and their home regions, we are able to train the UDM. We evaluate the performance of UDM on WM and Gowalla data sets, respectively.

On WM data set, we monitor the order of 30 million users' joining the OSN and then predict the distribution of the next 10 million users, which are tested by 10 experiment runs (each run contains one million users).

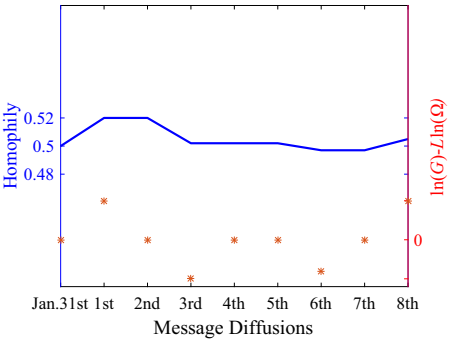

Figure 5.
The change of modularity toward information increment

As for Gowalla, we choose the group of first 100,000 users that have mostly checked in USA to be the training set, and use a group of 28,000 users as the testing set, which are tested by ten experiment runs as well.

Because we know the exact proportion of each region in the data set, we use histogram intersection (HI) to measure the prediction accuracy, which ranges between 0 and 1. Besides, we compare UDM to the baseline method (that naively predicts the future population of each region as proportional to the previously observed population of each region). The result is shown in Figure 6, which clearly indicates that UDM has better performance both on WM and Gowalla than the baseline method. We also observe that UDM provides a greater HI on the data set with a stronger geo-homophily (i.e. WM data set).

6.5 Performance of floating population inference model

In this subsection, we evaluate the performance of FPIM on WMeChat data set and compare it against the results of the latest national population census in China[2], which provides us the statistics of FP in China. Here, the FP in our experiments has excluded those whose home and remote regions are the same (e.g. those who rarely move out of the home region, as the work place and the home belong to the same region).

6.5.1 Correlation. As mentioned above, people tend to stay at the home region during Spring Festival in China. Therefore, the state of the message diffusion graph during Spring Festival can be seen as the baseline state S'. We collect the statistics on February 8, the Spring Festival day.

Intuitively, one may think that the proportion of FP would have certain correlation with the proportion of message diffusion in each region, and this leads to a naive prediction method that directly uses the latter to infer the former. We plot the correlation coefficient between the predicted FP distribution and the real FP distribution in Figure 7(a), which is as low as 0.2, indicating a poor direct correlation between FP and message diffusion. Indeed, this is a special case which ignores the uniform part of P_S. Because P_S holds the same superposition with $P_{\Delta S}$, we then evaluate the performance of FPIM based on P_S, as shown in Figure 7(b). The correlation coefficient is about 0.40, which is due to the uniform part of P_S, leading to a biased sampling on DP.

In contrast to this, FPIM based on state difference works better, illustrated in Figure 7(c), where the correlation coefficient reaches 0.8, indicating that the population prediction approximates the distribution derived from the national population census. Here, we notice in the WM data set that the mean of distribution difference $\Delta\mathbb{C} = \mathbb{C} - \mathbb{C}'$ is approximately 0, whereas the variance is about 1.4×10^{-4}. In other words, FPIM has significantly reduced the impact of the uniform part on $P_{\Delta S}$.

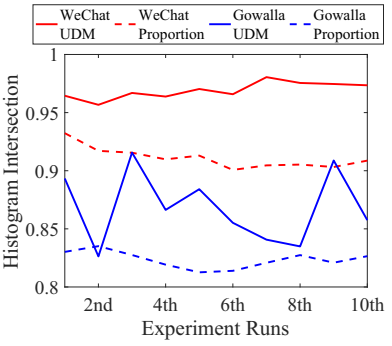

Figure 6.
Performance of UDM on WM and Gowalla data sets

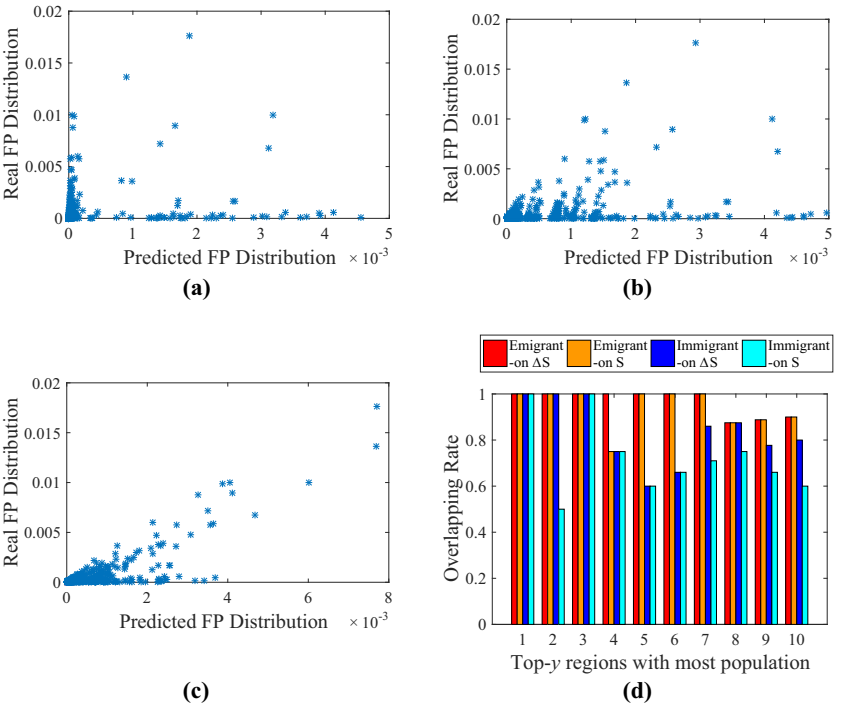

Figure 7.
Results of inferring the FP that has excluded those whose home and remote regions are the same

Notes: (a) Based on message diffusion; (b) based on P_s; (c) based on $P\Delta_s$; (d) proportion of correct predictions

By comparing the ticks over the x-axis and y-axis of Figure 7(c), we observe that FPIM predicts a FP (ticks over the x-axis) lower than that obtained in the national census (ticks over the y-axis). This can be attributed to the fact that a non-negligible proportion of FP do not view the WM pages or may not even use WM. As mentioned earlier, although there exist people not covered by the WM data set, the number of users in the data set is sufficiently large to derive the distributions using the proposed models.

6.5.2 Prediction correctness. Apart from correlation, we always have a concern on the densely populated region which has the most of a large FP that may cause changes to the online and offline social networks. We use the sets of regions who have the most proportion of FPs to measure the prediction correctness of FPIM.

For every province r, FPIM calculates the proportion of FP, by two sets, i.e. the set of emigrants whose currently located region is the remote region but the home region is r, and the set of immigrants whose currently located region is their remote region r but the home region is different.

Then, we rank the provinces by the number of immigrants and emigrants and obtain a ranking of provinces on immigrants and a ranking of provinces on emigrants, respectively. Meanwhile, the FP data of the national population census can also produce two rankings of provinces on immigrants and emigrants.

We compare the corresponding rankings obtained from FPIM and the national census and calculate the *overlapping rate* between two rankings on immigrants or emigrants (which is defined as the number of regions that appear in both rankings divided by the total number of

regions in a ranking) over the top-y provinces according to their normalized proportion values. We vary y from 1 to 10 and plot the histogram in Figure 7(d), telling that FPIM works satisfactorily with a match between our prediction results and the data of the national census. The two types of rankings have a high consistency on ΔS. Besides, the correctness on ΔS is higher than that calculated on basis of S the performance of FPIM on predicting the set of top provinces on emigrants is better than that on predicting the set of top provinces on immigrants, which is a result of the fact that FPIM uses σ_{ij}'s which pay more attention on the emigrant proportion.

7. Conclusions

In this paper, we propose a systematic study on the population distribution projection over offline geographical regions by analyzing the geographical attributes of OSNs. We propose the concept of geo-homophily in OSNs to establish the correlation between online message diffusion and the stability of geographical regions where a population distribution can be drawn. We formulate the population distribution problem from the perspective of DP and present prediction models to show the process that OSN users are distributed into regions, and infer the FP across regions. By experiments over the large-scale data sets, it is shown that the online message diffusions can help evaluate the stability of geographical regions, which further facilitates the determination of population distribution over fixed regions; the proposed prediction models have a high prediction accuracy in inferring the change of FP across regions.

Notes

1. www.census.gov/popest/data/datasets.html
2. www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm

References

Aldous, D.J. (1993), "Exchangeability and related topics", *École d'Été de Probabilités de Saint-Flour*, Vol. 33, pp. 1-198.

Backstrom, L., Sun, E. and Marlow, C. (2010), "Find me if you can: improving geographical prediction with social and spatial proximity", *Proceedings of the 19th International Conference on World Wide Web*, pp. 61-70.

Chi, G. (2010), "The impacts of highway expansion on population change: an integrated spatial approach", *Rural Sociology*, Vol. 75 No. 1, pp. 58-89.

Cho, E., Myers, S.A. and Leskovec, J. (2011), "Friendship and mobility: user movement in location-based social networks", *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1082-1090.

Crandall, D.J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D. and Kleinberg, J. (2010), "Inferring social ties from geographic coincidences", *Proceedings of the National Academy of Sciences*, Vol. 107 No. 52, pp. 22436-22441.

Gao, C., Zhang, Y., Bian, K., Li, Z., Bai, Y. and Liu, X. (2017), "Holiday syndrome: a measurement study of mobile social network use during holidays", *2017 IEEE International Conference on Communications (ICC)*, pp. 1-6.

Girvan, M. and Newman, M.E. (2002), "Community structure in social and biological networks", *Proceedings of the National Academy of Sciences*, Vol. 99 No. 12, pp. 7821-7826.

Goldenberg, J. and Levy, M. (2009), "Distance is not dead: social interaction and geographical distance in the internet era", arXiv preprint arXiv: 0906.3202.

Gonzalez, M.C., Hidalgo, C.A. and Barabasi, A.L. (2008), "Understanding individual human mobility patterns", arXiv preprint arXiv: 0806.1256.

Guo, Z., Huang, J., He, J., Hei, X. and Wu, D. (2013), "Unveiling the patterns of video tweeting: a sina weibo-based measurement study", *International Conference on Passive and Active Network Measurement*, pp. 166-175.

Harris, J.R. and Todaro, M.P. (1970), "Migration, unemployment and development: a two-sector analysis", *The American Economic Review*, Vol. 60 No. 1, pp. 126-142.

Hristova, D., Musolesi, M. and Mascolo, C. (2014), "Keep your friends close and your facebook friends closer: a multiplex network approach to the analysis of offline and online social ties", *ICWSM*.

Kido, H., Yanagisawa, Y. and Satoh, T. (2005), "An anonymous communication technique using dummies for location-based services", *Proceedings: International Conference on Pervasive Services, 2005: ICPS'05*, pp. 88-97.

Li, Z., Ding, B., Han, J., Kays, R. and Nye, P. (2010), "Mining periodic behaviors for moving objects", *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1099-1108.

Li, Z., Chen, L., Bai, Y., Bian, K. and Zhou, P. (2016), "On diffusion-restricted social network: a measurement study of wechat moments", *2016 IEEE International Conference on Communications (ICC)*, pp. 1-6.

Liao, Y., Bian, K., Song, L. and Han, Z. (2015), "Full-duplex MAC protocol design and analysis", *IEEE Communications Letters*, Vol. 19 No. 7, pp. 1185-1188.

National Bureau of Statistics of China (2010), "2010 population census", available at: www.stats.gov.cn/zgrkpc/dlc/yw/t20110428_402722384.htm

Nazir, A., Raza, S. and Chuah, C.N. (2008), "Unveiling facebook: a measurement study of social network based applications", *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, pp. 43-56.

Neal, R.M. (2000), "Markov chain sampling methods for Dirichlet process mixture models", *Journal of Computational and Graphical Statistics*, Vol. 9 No. 2, pp. 249-265.

Newman, M.E. (2004), "Fast algorithm for detecting community structure in networks", *Physical Review E*, Vol. 69 No. 6, p. 066133.

Schiavenza, M. (2013), "Wechat–not weibo–is the chinese social network to watch", *The Atlantic*, p. 30.

Wang, C., Tang, S., Yang, L., Guo, Y., Li, F. and Jiang, C. (2014), "Modeling data dissemination in online social networks: a geographical perspective on bounding network traffic load", Proceedings of the 15th ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 53-62.

Zhang, Y., Bai, Y., Chen, L., Bian, K. and Li, X. (2016), "Influence maximization in messenger-based social networks", *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1-6.

Zheng, Y., Xie, X. and Ma, W.Y. (2010), "Geolife: a collaborative social networking service among user, location and trajectory", *IEEE Data Engineering and Bulletin*, Vol. 33 No. 2, pp. 32-39.

Zheng, Y. (2012), "Tutorial on location-based social networks", *Proceedings of the 21st International Conference on World Wide Web*, Vol. 12 No. 5.

About the authors

Yuanxing Zhang is a PhD student at the School of Electrical Engineering and Computer Science, Peking University, China. He received his B.S. degree in computer science from Beijing University of Technology in 2015. His current research interests include AI-assisted resource allocation and recommender system. Yuanxing Zhang can be contacted at: longo@pku.edu.cn

Zhuqi Li is a PhD student at the Computer Science Department at Princeton University, USA. He received his B.S. degree in computer science from Peking University in 2017. His current research interests include social network and mobile sensing.

Kaigui Bian received the PhD degree in computer engineering from Virginia Tech, Blacksburg, USA in 2011. He is currently an associate professor in the Institute of Network Computing and

Information Systems, School of EECS, Peking University. His research interests include mobile computing, cognitive radio networks, network security, and privacy.

Yichong Bai received the B.S. degree in computer science from Peking University in 2014. He was the COO of Fibonacci Data Consulting Services Inc., Shenzhen, Guangdong, China.

Zhi Yang received the BS degree from the Harbin Institute of Technology, in 2005, and the PhD degree in computer science from Peking University, in 2010. He is currently an associate professor at Peking University. His research interests include the areas of security and privacy, networked and distributed systems, and data intensive computing.

Xiaoming Li graduated from Stevens Institute of Technology, New Jersey, with a PhD in computer science. He is now a professor at Peking University. His research interests include search engines and web mining. He is a Fellow of the Computer Federation of China and a member of Eta Kappa Nu. He serves on the Editorial Board of Concurrency and Computation (Wiley). He received the CCF Wang Xuan Award and Outstanding Educator Award in 2013 and 2014, respectively.