Optimal Trajectory Planning of Drones for 3D Mobile Sensing

Anonymous

***Abstract*—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. In this paper, we 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, we establish a three-layer 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. By experiments over large-scale OSN datasets, we 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.**

I. Introduction

Unmanned aerial vehicle (UAV), commonly known as drone, is an aircraft without a human pilot aboard, which is commonly used in measurement and sampling. Compared to manned aircraft, drones are more suitable for data collections and mobile sensing applications that capture different dimensions of signals in the environment that are beyond our sensing capability, such as aerial photography, 3D wireless signal survey, air quality index (AQI) measurement.

However, civilian drones are still not popular these days. Furthermore, a lot of drone companies were broken down. It could be a quite confusing problem if you have never come into attach with a drone. If you've actually tried using them, you could find that civilian drones do not really apply to daily life due to:

- Low battery available time.
- Great noise during flight.
- Wing rock and more battery drain caused by poor carrying capacity.

Therefore, in order to make more use of existing drones, we must consider the following problem: **How to complete measurement (or flight) in the shortest possible time? Furthermore, in the three-dimensional space?**

Similar to traditional sensor networks and mobile base station, we consider data collection in mobile environment. So total time consumption consists of two parts: **flight time** and **measure time**. While we also have the following difference:

- We consider optimal algorithm in two-dimensional space.
- We use the routing algorithm based on graph theory apart from traditional greedy algorithms.

In this paper, we consider mobile sensing in three-dimensional space. We divide three-dimensional space into a network of observation locations (OLs) and select critical observation locations (COLs) from OLs to cover measurement space, which could be formulated as a constraint set coverage problem in graph theory. Specifically, we consider the following two special cases:

1) *Consider flight time only*: Under this condition, we assume measurement time negligible. Then we could formulate this problem as a minimum dominating set problem in lattice, which has been studied for a long time.
2) *Consider measurement time only*: Under this condition, we assume flight time negligible. Then we could formulate problem as a minimum dominating path problem in lattice, which has not been solved before. In this paper, we solved this problem in grid and give an expand in three-dimensional space.

Because of algorithms we use is based on graph theory, We could solve two problems above optimally in O(1)time. We use drones to verify our simulation in multiple scenarios. We find out that the flight time in algorithm is less than ordinary approach. And even though we take less observe point than the ordinary approach, we could have a better result.

II. Related Work

A. 3D mobile sensing

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.

Predicting mobility patterns using OSN 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 [6]. Cho et. al. 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 [2], 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.

Data dissemination in a geographical perspective. Wang et. al. 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 [10].

Online and offline social behaviors. Zheng et al. propose Location-Based Social Network (LBSN) [11], 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. experimentalized on a dataset with 74 college students as volunteers [5], 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.

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 [4]. Considering the co-occurrence in time and space [3], Crandall et al. 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.

B. Route planning in conventional wireless sensor networks

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

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

There are two parameters:

- 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, while an extremely large α makes the samples be more consistent with the base measure.
- 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 (LDA) 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.

Dirichlet Process (DP) is a class of Bayesian nonparametric models, and Dirichlet process generalizes Dirichlet distribution [7]. 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 Dirichlet process to model the process that OSN users are distributed into geographical regions.

III. 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 [8] 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.

A. A Three-layer Framework

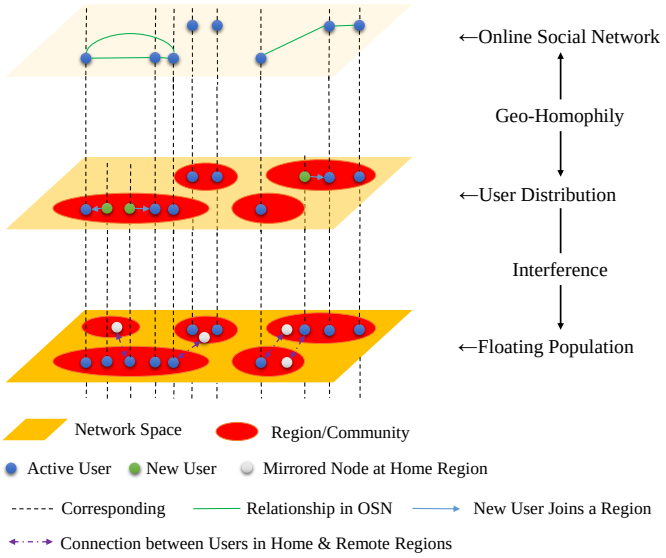


Fig. 1: A three-layer analytical framework that defines the geo-homophily of OSNs to map the OSN message diffusion (Layer 1) to the offline user distribution (Layer 2), and infers the floating population (Layer 3) based on the derived distribution (Layer 2).

In Figure 1, we show a three-layer 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 floating population 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 floating population across regions can be further inferred based on the derived distribution.

B. Geo-homophily of An OSN 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} representing the number of views from node u to the content of node v during time period T. The e_{uv} exists if the number of views from node u to the content of node 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 $\mathcal{E}$, where $\forall \epsilon_{ij} \in \mathcal{E}$ has value

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

which represents the 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

$$Q_{RT} = \sum_{i \in R} (p_{iiT} - \sum_{j \in R} p_{ijT} \sum_{j \in R} p_{jiT}).$$

Apparently, Q_{RT} ranges in $[-1, 1]$. 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}.

C. 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 since 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. 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:

$$H(G) = -\sum_{i \in R} \sum_{j \in R} p_{ijT} \times \log p_{ijT}. \tag{1}$$

As the redistribution of message diffusion inside each region do not significantly affect the modularity, 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:

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

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 $\overline{\mathbf{L}}_{|R| \times |R|}$ 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

$$\mathcal{G}(G, \overline{\mathbf{L}}) = \frac{(\omega_{I_T} + l_I)^{(\omega_{I_T} + l_I)} \prod_{i,j \in R, i \neq j} (\omega_{ijT} + l_{ij})^{(\omega_{ijT} + l_{ij})}}{\omega_{I_T}^{\omega_{I_T}} \prod_{i,j \in R, i \neq j} \omega_{ijT}{}^{\omega_{ijT}}}. \tag{3}$$

According to (2), the social-entropy becomes:

$$\begin{aligned} H(G') = &-\frac{\omega_{I_T} + l_I}{\mathcal{L} + \Omega_T} \log \frac{\omega_{I_T} + l_I}{\mathcal{L} + \Omega_T} \\ &- \sum_{i \in R} \sum_{j \in R, i \neq j} \frac{\omega_{ijT} + l_{ij}}{\mathcal{L} + \Omega_T} \log \frac{\omega_{ijT} + l_{ij}}{\mathcal{L} + \Omega_T}. \end{aligned} \tag{4}$$

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.

Proposition 1. *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:

$$\Delta H = \frac{1}{\Omega_T}(\log\Omega_T^{\mathcal{L}} - \log\mathcal{G}(G,\overline{\mathbf{L}})). \tag{5}$$

Then we can substitute $\mathcal{G}(G,\overline{\mathbf{L}}) \geq \Omega_T^{\mathcal{L}}$ into (5) and we could conclude with this proposition. ∎

IV. 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 Dirichlet process), 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 Dirichlet process, 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.

A. 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 $\mathcal{U}$. A UDM has two parameters: base distribution $\mathcal{U}_0$ 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:

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

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

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

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 $\widetilde{\theta}_r, r \in R$ be the non-redundant hyper-parameters.

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

$$\mathcal{U} \sim \text{Dir}(\{\alpha_r\}_{r\in R}).$$

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:

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

Thus, the marginal probability would be

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

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

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

B. A Special Case of Chinese Restaurant Process

The process of distributing users over multiple regions is a special case of Chinese Restaurant Process [1], 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.

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 $\mathcal{U}$ over $\widetilde{\theta}$ is deduced as:

$$\begin{aligned} &P(\mathcal{U}|\{\widetilde{\theta_r}\}_{r\in R}) \\ &\propto P(\{\widetilde{\theta_r}\}_{r\in R}|\mathcal{U})P(\mathcal{U}) \\ &= \mathcal{U}P(\mathcal{U}). \end{aligned}$$

So we have

$$\mathcal{U}|\{\widetilde{\theta_r}\}_{r\in R} \sim \text{UPM}(\alpha+1, \frac{\alpha\mathcal{U}_0 + \delta_{\widetilde{\theta}}}{\alpha+1})$$

where $\delta_{\widetilde{\theta}}$ is a probability measure concentrated at $\widetilde{\theta}$.

Consider a partition $(\theta', \mathbb{U}\backslash\theta')$, we have

$$\begin{aligned} &(\mathcal{U}(\theta'), \mathcal{U}(\mathbb{U}\backslash\theta')) \\ &\sim \text{Dir}((\alpha+1)\frac{\alpha\mathcal{U}_0 + \delta_{\widetilde{\theta'}}}{\alpha+1}(\theta'), (\alpha+1)\frac{\alpha\mathcal{U}_0 + \delta_{\widetilde{\theta'}}}{\alpha+1}(\mathbb{U}\backslash\theta')) \\ &= \text{Beta}(1, \alpha). \end{aligned}$$

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

$$\begin{aligned} \mathcal{U} &\sim \text{UPM}(\alpha, \mathcal{U}_0) \\ &= \beta_1\delta_{\widetilde{\theta_1}} + (1-\beta_1)\mathcal{U}_1 \\ &= \dots \\ &= \sum_{i=1}^{|R|} \pi_i \delta_{\widetilde{\theta_i}} \end{aligned}$$

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

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

The posterior distribution of β_i satisfies

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

V. Optimal Trajectory Planning Algorithms

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 $\mathcal{H}$, and the *remote* region denoted as $\mathcal{W}$ (e.g., the work place). According to the previous study [2], 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 floating population.

A. Distribution of Message Diffusions

We use a tetrad $\mathbb{S} = (\mathbb{C}, \mathbb{Q}, \lambda, \chi)$ 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 while 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 Dirichlet Process.

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

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

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 $\mathbb{S}' = (\mathbb{C}', \mathbb{Q}', \lambda, \chi)$, $\mathbb{C}' = \{c'_i\}_{i\in R}$, 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 $\mathbb{S}$ and the baseline state, named as *State Difference* $\Delta\mathbb{S}$.

Proposition 2. *The State Difference follows a superposition of a uniform distribution and a Dirichlet distribution.*

Proof: The proportion of messages from $r_{\mathcal{O}}$ to $r_{\mathcal{T}}$ should be

$$\begin{aligned} &P_{\mathbb{S}}(r_{\mathcal{T}} = j|r_{\mathcal{O}} = i) \\ &= \sum_{r_{\mathcal{H}}\in R} P(j|r_{\mathcal{H}}, r_{\mathcal{W}} = i)\sigma_{r_{\mathcal{H}}i} \\ &= (1-\lambda-\chi)c_j + \lambda\sigma_{ji} + [j=i]\chi\sigma_{ii}. \end{aligned}$$

Therefore, we can deduce that

$$\begin{aligned} &P_{\Delta\mathbb{S}}(r_{\mathcal{T}} = j|r_{\mathcal{O}} = i) \\ &= P_{\mathbb{S}}(r_{\mathcal{T}} = j|r_{\mathcal{O}} = i) - P_{\mathbb{S}'}(r_{\mathcal{T}} = j|r_{\mathcal{O}} = i) \\ &= \begin{cases} (1-\lambda-\chi)(c_j - c_{j'}) + \lambda\sigma_{ji} & i \neq j \\ (1-\lambda-\chi)(c_j - c'_j) + (\lambda+\chi)(\sigma_{ii} - 1) & i = j \end{cases} \end{aligned} \tag{6}$$

which is a constant plus a variable generated from Dirichlet process. It indicates that the State Difference follows a superposition of a uniform distribution and a Dirichlet distribution. ■

This proposition enlightens us to infer floating population by methods of divide and conquer. The State Difference reduces the weight of the uniform distribution component.

B. Export Message Pattern (EMP)

Similar to UDM, we can extract the distribution of messages diffused to remote regions, and we use a Hierarchy Dirichlet

Process to find the distribution, which is named as the Export Message Pattern (EMP). For every region i, denote $\sigma_{\mathbf{i}}$ as $\{\sigma_{ji}\}_{i\neq j}$, following

$$\begin{aligned}\mathbb{B}_0 &\sim \text{DP}(\tau', \mathbb{B}')\\ \mathbb{B}_i &\sim \text{DP}(\tau_i, \mathbb{B}_0)\\ \eta_i &\sim \mathbb{B}_i\\ \sigma_{\mathbf{i}} &\sim F(\eta_i)\end{aligned}$$

where η_i is the hyper-parameters, τ_i and τ' is the corresponding scale parameter and $\mathbb{B}'$ is the base distribution. Consider the differential export message

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

which satisfies that

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

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

C. Self Message Pattern (SMP)

The Dirichlet Process Mixture can also explain the distribution of messages diffused inside each region, which is named as Self Message Pattern (SMP). According to (6), it is not wise to gather $\sigma_{ii}\ \forall i \in R$. Instead, we should concern $\{\sigma_{0i} = 1 - \sigma_{ii}\}_{i\in R}$ and denote it as $\sigma_{\mathbf{0}}$. We are able to find a scale parameter τ_0 and base distribution $\mathbb{I}_0$ such that

$$\begin{aligned}\mathbb{I} &\sim \text{DP}(\tau_0, \mathbb{I}_0)\\ \eta_0 &\sim \mathbb{I}\\ \sigma_{\mathbf{0}} &\sim F(\eta_0).\end{aligned}$$

Since we have access to

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

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 $\sum_j \sigma_{ji} = 1$.

D. 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, while EMP and SMP compute the specific allocation inside each region. The model structure is shown in Figure 2.

VI. Evaluation

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

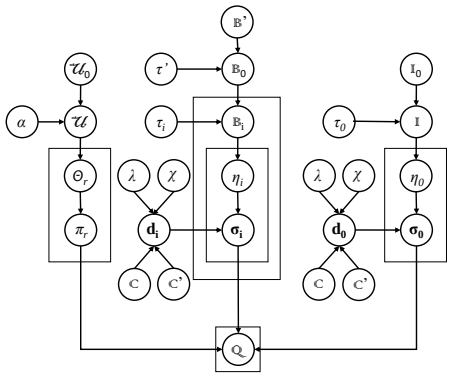


Fig. 2: The Floating Population Inference model has three parts, namely UDM, EMP, SMP from left to right.

A. Datasets

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

"Gowalla" [2] is a typical location-based social network (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 dataset collects a total of 6,442,890 check-in's of these users over the time period from Feb. 2009 to Oct. 2010.

"WeChat Moments (WM)" [9] is the social network of a mobile messaging app (Wechat) popular in China, where the contents shared over WM are HTML5 pages. This WM dataset contains 137,509,889 users with 1,671,692,424 retweeting/forwarding records of 329,465 pages from Jan. 14 2016 to Feb. 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. WeChat Moments 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.

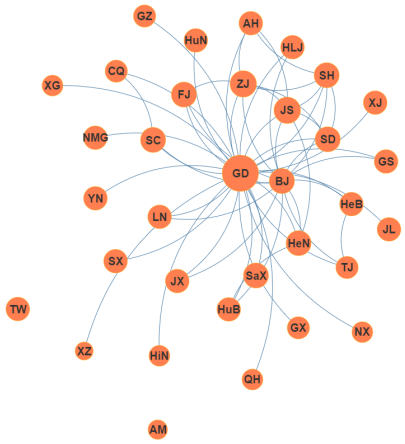
Note that: although the number users of each dataset is much less than the population of a country, it is sufficiently large to derive the proportion of OSN user distribution, as well as the population distribution over geographical regions, which helps us determine how a new OSN user is distributed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

B. Geo-homophily of OSNs

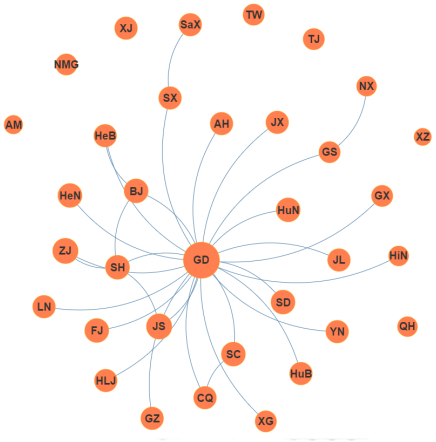
As mentioned in Section III, we can divide users of an OSN into a division of regions, according to users' geo attributes.

1) Geo-homophily of WeChat:

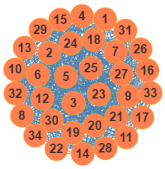
Message diffusion in China. The WM dataset records the page re-tweeting in 34 provinces in China, and we use these provinces as the geographical regions in this experiment. Every user in WM should have viewed a collection of pages, and



(a) Before Spring Festival.



(b) On the Spring Festival day.



(c) Baseline network.

Fig. 3: Graph representations of message diffusion inside and across provinces in Wechat Moments dataset.

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 Jan. 14 to Jan. 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.

Results for pre-holidays. The modularity in the pre-holidays is approximately 0.49. Figure 3a shows the volume of message diffusion inside each province and that between 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: (1) the larger size an orange circle has, more friend relationships between two people exist inside the province labeled in the circle; (2) the shorter a blue arc is, more friend relationships exist between people in two different provinces whose corresponding circles are connected by the arc; (3) 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 dataset; 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.

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 3b, where some inter-province arcs disappear.

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 [8]. 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 3c. Compared to Figure 3a, under the same scale of plotting, the distribution of circles representing regions are very dense; and most of circles' sizes are similarly small.

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

[1]http://www.census.gov/popest/data/datasets.html

C. 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, while 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 III-C. 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. 31 of the WM dataset. 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 Multi-nomial distribution. One million retransmissions are then sampled through the Multi-nomial 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.

D. Performance of UDM

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 datasets respectively.

On WM dataset, 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).

As for Gowalla, we choose the group of first 100 thousand users that have mostly checked in USA to be the training set, and use a group of 28 thousand users as the testing set, which are tested by 10 experiment runs as well.

Since we know the exact proportion of each region in the dataset, 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 dataset with a stronger geo-homophily (i.e. WM dataset).

E. Performance of FPIM

In this subsection, we evaluate the performance of FPIM on WMeChat dataset, and compare it against the results of the latest national population census in China [2], which provides us the statistics of floating population in China. Here, the floating population (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).

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 $\mathbb{S}'$. We collect the statistics on Feb. 8, the Spring Festival day.

Intuitively, one may think that the proportion of floating population 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 7a, 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 $P_{\mathbb{S}}$. Since $P_{\mathbb{S}}$ holds the same superposition with $P_{\Delta\mathbb{S}}$, we then evaluate the performance of FPIM based on $P_{\mathbb{S}}$, as shown in Figure 7b. The correlation coefficient is about 0.40, which is due to the uniform part of $P_{\mathbb{S}}$, leading to a biased sampling on DP.

In contrast to this, FPIM based on State Difference works better, illustrated in Figure 7c, 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 dataset that the mean of distribution difference $\Delta\mathbb{C} = \mathbb{C} - \mathbb{C}'$ is approximately 0, while the variance is about 1.4×10^{-4}. In other words, FPIM has significantly reduced the impact of the uniform part on $P_{\Delta\mathbb{S}}$.

By comparing the ticks over the x-axis and y-axis of Figure 7c, we observe that FPIM predicts a floating population (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 floating population do not view the WM pages or may not even use Wechat Moments. As mentioned earlier, although there exist people not covered by the WM dataset, the number of users in the dataset is sufficiently large to derive the distributions using the proposed models.

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

For every province r, FPIM calculates the proportion of floating population, 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 floating population data of the national population census can also produce two rankings of provinces on immigrants and emigrants.

[2]http://www.stats.gov.cn/tjsj/pcsj/rkpc/6rp/indexch.htm

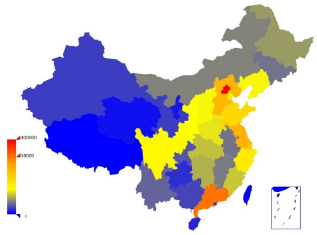


Fig. 4: The viewing distribution of a hot message originated from Beijing.

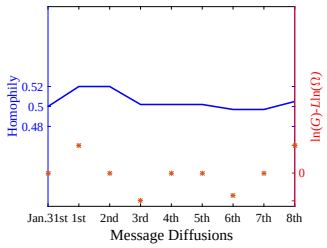


Fig. 5: The change of modularity towards Information Increment.

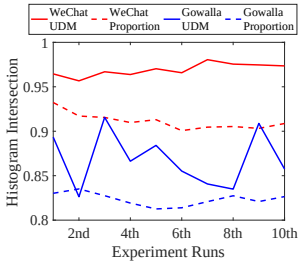
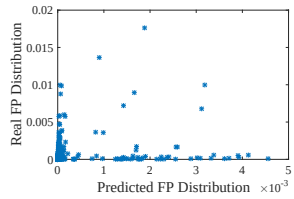
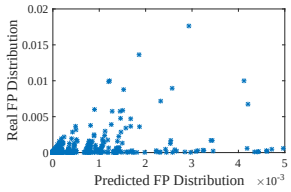


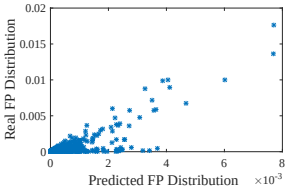
Fig. 6: Performance of UDM on WM and Gowalla datasets.



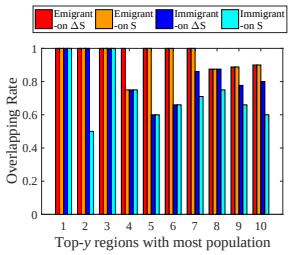
(a) Based on Message Diffusion.



(b) Based on $P_{\mathbb{S}}$.



(c) Based on $P_{\Delta\mathbb{S}}$.



(d) Proportion of correct predictions.

Fig. 7: Results of inferring the floating population (FP) that has excluded those whose home and remote regions are the same.

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 7d, 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 $\Delta\mathbb{S}$. Besides, the correctness on $\Delta\mathbb{S}$ is higher than that calculated on basis of $\mathbb{S}$; 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.

VII. Conclusions

In this paper, we propose a systematic study on the population distribution projection over offline geographical regions by analyzing the geographical attributes of online social networks (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 Dirichlet process, and present prediction models to show the process that OSN users are distributed into regions, and infer the floating population across regions. By experiments over the large scale datasets, 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 floating population across regions.

References

[1] D. J. Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII-1983*, pages 1–198. Springer, 1985.

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

[3] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 107(52):22436–22441, 2010.

[4] J. Goldenberg and M. Levy. Distance is not dead: Social interaction and geographical distance in the internet era. *arXiv preprint arXiv:0906.3202*, 2009.

[5] D. Hristova, M. Musolesi, and C. Mascolo. Keep your friends close and your facebook friends closer: A multiplex network approach to the analysis of offline and online social ties. *Eprint Arxiv*, 2014.

[6] Z. Li, B. Ding, J. Han, R. Kays, and P. Nye. Mining periodic behaviors for moving objects. *ACM SigKDD*, pages 1099–1108, 2010.

[7] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[8] M. E. Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[9] M. Schiavenza. Wechat—not weibo—is the chinese social network to watch. *The Atlantic*, 30, 2013.

[10] C. Wang, S. Tang, L. Yang, Y. Guo, F. Li, and C. Jiang. Modeling data dissemination in online social networks: a geographical perspective on bounding network traffic load. In *ACM MobiHoc*, pages 53–62, 2014.

[11] Y. Zheng. Tutorial on location-based social networks. *Microsoft Research*, 2012.