A Homophily-Free Community Detection Framework for Trajectories with Delayed Responses

Extended Abstract

Chung-Kyun Han
School of Information Systems
Singapore Management University
Singapore
ckhan.2015@phdis.smu.edu.sg

Shih-Fen Cheng
School of Information Systems
Singapore Management University
Singapore
sfcheng@smu.edu.sg

Pradeep Varakantham
School of Information Systems
Singapore Management University
Singapore
pradeepv@smu.edu.sg

ABSTRACT

Community detection has been widely studied in the areas of social network analysis and recommendation system. However, most existing research focus on cases where relationships are explicit or depend on simultaneous appearance. In this paper, we propose to study the community detection problem where the relationships are not based on simultaneous appearance, but time-delayed appearances. In other words, we aim to capture the relationship where one individual physically follows another individual. In our attempt to capture such relationships, the major challenge is the presence of spatial homophily, i.e., individuals are attracted to locations due to their popularities and not because of communications.

In tackling the community detection problem with spatial homophily and delayed responses, we make the following key contributions: (1) We introduce a four-phase framework, which by way of using quantified impacts excludes homophily. (2) To validate the framework, we generate a synthetic dataset based on a known community structure and then infer that community structure. (3) Finally, we execute this framework on a real-world dataset with more than 6,000 taxis in Singapore. Our results are also compared to those of a baseline approach without homophily-elimination.

KEYWORDS

Spatial homophily; Community detection; Trajectory simulation; Hotspot detection

ACM Reference Format:
Chung-Kyun Han, Shih-Fen Cheng, and Pradeep Varakantham. 2019. A Homophily-Free Community Detection Framework for Trajectories with Delayed Responses. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Community detection has been widely studied in the areas of social network analysis and recommendation system. Most early works in these areas focus on community detection with definite relationships among individuals, usually represented in the form of a connected graph; thus the methods developed are mostly related to graph partitioning [3]. However, these methods are mostly infeasible for use in cases where relationships are not explicitly observable. One such example is on detecting communities using only human trajectories obtained using sensors such as GPS, Wifi, or mobile phone communication logs [2, 4–6].

Our work differs from the link-based and trajectory-based community detection literature in the following aspects: (1) relationships among individuals are not directly observable, but have to be *inferred* from location traces, and (2) the relationships inferred from location traces are not based on simultaneous appearance, but sequential (time-delayed) appearances. In other words, we aim to capture the relationship that would induce one individual to physically follow another individual. However, a major challenge in our effort is the existence of *spatial homophily*, which refers to individuals appearing at specific places in sequence due to factors such as similar preferences or location popularity and not due to the actual relationship.

To address these challenges, we create a framework that can identify homophily-free relationships from sparse location traces, deduce communities, and derive interaction hotspots for members belonging to the same community.

2 THE COMMUNITY DETECTION FRAMEWORK

To demonstrate the practicality of our framework, we choose to study a mobility trace dataset from a large taxi fleet operator in Singapore. We assume that mobility traces of subjects can be clearly labeled into episodes with origins and destinations. We also assume that we can identify the time the subject spent around the origin before departing and moving towards the destination (this idling time can be viewed as the search cost associated with the episode). There are four interconnecting components in our community detection framework, which are described below.

Trajectory Analytics: For each episode, we process the raw traces to find out: 1) the zone and time period this episode originates from, 2) the amount of idling time or dwell time (DT) this subject spent in the zone before starting the episode, and 3) a list of other subjects who have episodes originating from the same zone earlier and might play the role of an influencer.

Graph Construction: The goal of this component is to infer the strength of the relationship between any pair of subjects and construct a graph representing the inferred social network among subjects. To estimate subject relationships without the interference of the spatial homophily effect, we adopt a statistical approach: treat all of a subject's dwell times as the response variable, and estimate the impact of other subject's presence in the same zone.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

This statistical estimation can be accomplished by executing the following regression model:

$$DT_i = \alpha_i + \beta_{ji} X_{ji} + \sum_{z \in Z} \lambda_z K_z + \epsilon_i, \forall j \in \mathbf{D}_i, \quad (1)$$

where DT_i is subject i's DT, α_i is the constant term, β_{ji} is how strongly j affects i's DT, λ_z is the zone-specific impact on DT, ϵ_i is the Gaussian noise. $\mathbf{D}_i$ is the set of all potential influencers who might have relationships with subject i and are derived from the Trajectory Analytics phase. For any pair (i, j), the corresponding X_{ij} is a linearly normalized real value between 0 and 1, where 1 denotes episodes' origins are overlapping. The vector of K_z is a list of indicator variables, where K_z is set to 1 if this particular DT is from the zone z. Finally, when constructing the relationship graph, we denote subjects as nodes, and create a directed edge (j, i) with weight β_{ji} only if it is statistically significant to the level of 1%. Note that the weight of the edge (j, i) is set to $-\beta_{ji}$, since a negative coefficient actually represents a positive relationship.

Community Detection: Given a social network with relationships as weighted edges, the community detection problem is well-defined and well-studied. What we implement in our framework is based on a software library utilizing a standard modularity-maximizing algorithm by [1].

Interaction Hotspot Detection: For each identified community, this component is designed to identify a collection of zones where intense interactions happened. This is achieved by the following regression model:

$$DT_z^c = \alpha_z^c + \beta_z^c X_z^c + \epsilon_z^c, \quad (2)$$

where DT_z^c represents the DTs of all trips occurred in zone z for community c, α_z^c is the constant term for z, X_z^c is a binary variable indicating whether there are any pair of subjects from c appearing in z within the stipulated time window L, and ϵ_z^c is the Gaussian noise. For zones with significant negative coefficients (we again use the significance level of 1%), β_z^c, we conclude that the presence of other subjects from the same community offers significant help in reducing DT.

3 VALIDATION USING SYNTHETIC DATASET

To validate our framework, we use a synthetic dataset that is generated using a simulation. This simulation produces movement traces with a known community structure and location topology as inputs. The performance of our platform is measured using two metrics: 1) false negatives, i.e., the number of missed relationships (#MR), and 2) false positives, i.e., the number of wrongly identified relationships (#WR). Each test scenario is characterized by two parameters: the number of zones and the demand profiles. Without loss of generality, we use networks with 5 and 10 zones, and denote them as $Z(S)$ and $Z(M)$ respectively. For the demand profiles, we have low, middle, and high levels of demands, and they are denoted as $D(L)$, $D(M)$ and $D(H)$. The results along two defined dimensions are plotted in Figure 1. When looking at the results, we can see that our framework can identify all communities except for the low-demand scenarios. Our framework performs best when the level of demands is non-trivial and the network size is large. For the real-world dataset that we are about to test our framework on, these two features are both prominent, we thus expect equally good performance although we cannot directly verify the identified communities with ground truth (none existed).

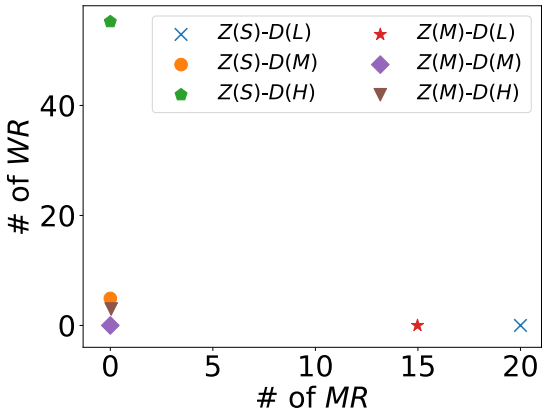

Figure 1: Number of false negatives and positives.

Table 1: Comparison against the baseline approach.

	Nodes	Edges	Density	# Com	# Mem
Baseline	6,070	164,262	0.4%	8	758.75
Ours	1,921	4,710	0.1%	104	18.47

4 A REAL-WORLD CASE STUDY

To realistically test our approach, we apply our framework to a real-world dataset collected from a large taxi fleet in 2009, in which 6,120 taxi drivers are considered. We present the results obtained both from our approach and a baseline alternative not considering spatial homophily elimination. Table 1 shows the summary statistics. From the summary we can clearly see that the baseline approach produces the much larger network, identifies much fewer communities (denoted as # Com), while having a large number of members (denoted as # Mem), which is unrealistic. The baseline outcomes clearly demonstrate the strong impact of spatial homophily, and the consequence of ignoring it. Recognizing that this is in the year of 2009, during which free mass messaging Apps were still not available in Singapore, we doubt that a community can have size beyond ten members. On examining the community structures, we identify that the larger communities are indeed mostly composed of several cliques, but connected through some key members. We also discover that members of communities mostly interact at a small number of strategic locations. The extreme cases are communities that only interact at hotspots such as the airport or the central business district (however, these interactions are not due to spatial homophily, as established via our regression model).

ACKNOWLEDGMENTS

This research is funded by the National Research Foundation Singapore under its Corp. Lab @ University scheme and Fujitsu Limited as part of the A*STAR-Fujitsu-SMU Urban Computing and Engineering Centre of Excellence.

REFERENCES

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008 (2008), P10008.

[2] Nathan Eagle, Alex Sandy Pentland, and David Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* 106, 36 (2009), 15274–15278.

[3] Santo Fortunato. 2010. Community detection in graphs. *Physics reports* 486, 3-5 (2010), 75–174.

[4] Gueorgi Kossinets and Duncan J Watts. 2006. Empirical analysis of an evolving social network. *Science* 311, 5757 (2006), 88–90.

[5] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Life in the network: the coming age of computational social science. *Science* 323, 5915 (2009), 721.

[6] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. 2007. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences* 104, 18 (2007), 7332–7336.