IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 28, 2018, accepted May 21, 2018, date of publication May 28, 2018, date of current version June 20, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2841321

Comparing Community Detection Algorithms in Transport Networks via Points of Interest

LIPING HUANG[1], (Member, IEEE), YONGJIAN YANG[1], HEPENG GAO[2], XUEHUA ZHAO[3], AND ZHANWEI DU[1]

[1]College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2]College of Software, Jilin University, Changchun 130012, China
[3]School of Digital Media, Shenzhen Institute of Information Technology, Shenzhen 518172, China

Corresponding authors: Yongjian Yang (yyj@jlu.edu.cn) and Zhanwei Du (huangliping5727@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772230 and 61702215, in part by the Jilin Province Science and Technology Development Program under Grant 20160204021GX, in part by the Special Fund for Industrial Innovation Project of Jilin Province under Grant 2017C031-1, in part by the Special Innovation Project of Guangdong Education Department under Grant 2017GKTSCX063, and in part by the Guangdong Natural Science Foundation under Grant 2016A030310072.

ABSTRACT Passengers travel in transport networks with diverse interests represented by linked points of interest (POIs) and drive urban regions to group into network communities. Previous studies focused on applying community detection methods (CDMs) to discover spatial mobility patterns or using POIs to explain the decision making of human mobility, without comparing the effectiveness of CDMs for detecting network communities. In this paper, we analyze the relationship between POIs and network communities of human mobility over diverse CDMs. Taking the taxi systems of Shanghai and Beijing as case studies, we construct transport networks with urban regions as nodes and the connections between them as links weighted by mobility flows. The spatial communities are identified based on the movement strength among regions. POIs are mapped into nodes in the network and are considered as independent variables for classifying the spatial community categories. Our study suggests that communities detected with two specific CMDs (namely, the Combo algorithm and the Walktrap algorithm) correlate to POIs, and the correlation of the Combo is the best ($R^2 = 0.3$ for Shanghai and $R^2 = 0.48$ for Beijing). In this regard, this paper can provide valuable insight into understanding the formation of spatial communities and assist in selecting reasonable CDMs.

INDEX TERMS Community detection, logistic regression, mobility flow, points of interest.

I. INTRODUCTION

People move around in a city, generating population mobility flows in urban transport networks. Knowledge of the spatial pattern of citizens' travel in a city is particularly beneficial for the convergence of applications, such as selecting locations for retail stores to allow more customers to shop around, and advertisement casting to capture as many consumers as possible [1], [2].

To analyze the spatial variability of urban mobility flows, we construct a transport network with partitioned urban regions as nodes and the connections between them as links weighted by the aggregated strengths of inter-region movements [1], [3], [4]. The community in the transport network is applied for further analysis of the spatial variability of mobility flows as it offers a visual representation of the spatial cluster features of mobility flows, where a spatial community is a set of nodes with more connections among themselves than with the remaining nodes [5]. Traditional CDMs based on Newman's modularity optimization, combined with the representative CDMs of *LPA* [6], *walktrap* [7] and a high-quality CDM called *combo* [8], are applied to detect spatial communities of mobility flows. A brief description of each CDM used in this research is provided in the third section.

Actually, each trip between urban regions connects specific POIs. For example, commuting trips connect a citizen's home and workplace. This means urban mobility flows are rooted in people's traveling activities (e.g., work or entertainment) [9], [10], reflected by specific POIs. Therefore, researching the inherent consistency between spatial mobility communities and POIs provides new insight for understanding the underlining mechanism of urban movements. The main contribution of this study consists of three points:

(1) We construct transport networks with segmented regions of the studied area as nodes and connections between regions as links weighted by the volume of mobility flows. Then, CDMs are applied to identify the spatial community pattern of mobility flows.
(2) Further, POIs are mapped into nodes to characterize the driving factors for generating spatial communities. We consider this a multi-class classification problem with the community categories as classification labels and the POIs in a node as feature vectors, which is solved by adopting a stepwise logistic regression.
(3) We evaluate the consistency between spatial communities and POIs using large-scale and real-world datasets, containing POI datasets and taxi GPS trajectory datasets of Beijing and Shanghai, China. Experiment results show that *combo* is the best CDM suite for acquiring POI motivated spatial communities of the transport networks.

According to the research problems of our work presented above, the rest of this paper is organized as follows: Section 2 presents related works. Methods used in this paper are shown in Section 3, including the construction of transport networks, the description of representative CDMs, and our proposed consistency estimation model. Experimental datasets and result analysis are reported in Section 4. Finally, the paper is concluded in Section 5.

II. RELATED WORKS

Technological advances allow for precise measurement of mobility flows in large datasets, including taxi trajectories [11]–[13], mobile phone trajectories [4], [14], and transport smart cards [9]. Retrospective studies of mobility flow focus on modeling mobility flow from one place to another, such as the universal model, called radiation model [15], which is proposed and applied to predict human mobility volume [16]. Although the model is parameter-free and only requires population distribution as input, the spatial cluster features of mobility flows are disregarded, meaning most people travel in a specific range of regions instead of the entire city and some citizens share a similar regional scope.

Combined with network techniques, applications based on mobility flow are widely developed in the field of urban computing [17], [18]. For example, the centrality metrics of a network are used to estimate the importance of road segments [11]. To determine the interaction between regions, the studied area is segmented into disjointed regions, and mobility flow between geographical regions is used to discover the connectivity between regions and reveal new latent links, thus determining the inadequacy of the existing road network [19]. Using taxi trajectory data from Shanghai, Liu *et al.* [13] built spatial networks to model intra-city spatial interactions, revealing the hierarchical and polycentric structure of Shanghai.

Studies mentioned above provide insights into using emerging data sources to reveal mobility patterns and the urban structure. However, the underlying mechanisms that generate spatial patterns and urban structures from the land-use aspects have not been researched. Complementary, mobility flows in subway systems are combined with POIs to research activity patterns and model the dynamic decision-making process that shape individuals' movements [9]. This research constructs a transport network with subway stations as nodes and mobility flow from one station to another as weights on the directed edge. When it comes to researching the overall urban movements, the city area is always segmented into regions. Segmented regions of the city carry socio-economic functions because people live in the regions and POIs exist in regions, and regions as the origin and destination of a trip cause mobility flows [20]. The studies above indicate mobility flows are related to POIs distributed among urban regions.

However, there is no research specializing in the consistency between spatial communities and driving factors for urban mobility flows. Existed CDMs are usually adopted to mine the spatial mobility pattern, but determining the appropriate CDM has not been researched. Both problems are researched in this paper. Our work is different from the research mentioned in the following aspects. First, we add the POI feature to nodes in the spatial networks to characterize the socio-economic factors that motivate mobility flows. Moreover, based on the multi-class classification method of stepwise logistic regression, we estimate the consistency between spatial communities and POIs, further to determine the appropriate CDMs suitable for detecting spatial communities in keeping with the distribution of POIs.

III. METHODS

To estimate the consistency between spatial communities of mobility flows and POIs, we construct a transport network for the study area and detect the spatial communities. Then, POIs are mapped into corresponding nodes in the network to characterize the driving factors of urban movements. Spatial communities are used to classify the nodes in the network. We adopt multi-class classification methods to classify the nodes using the community as the classification label and the POI feature of each node as the independent variable. First, we depict the construction of the transport network, and present the CDMs used to identify spatial communities. Then, we depict the consistency estimation model, consisting of the map matching and the classification method of stepwise logistic regression.

A. NETWORKS AND CDMs

Mobility used in this paper is represented as a 2-tuple $< (x_o, y_o), (x_d, y_d) >$. Both (x_o, y_o) and (x_d, y_d) are geo-spatial positions, denoting the origin and destination of a trip, respectively. In detail, the origin and destination (OD) pair represents a trip starting at location (x_o, y_o) and arriving at location (x_d, y_d).

To construct the network in this research, the study area is segmented into disjoint grids, and each grid g_i is set as a node n_i, as illustrated in Fig. 1. Trips between two nodes

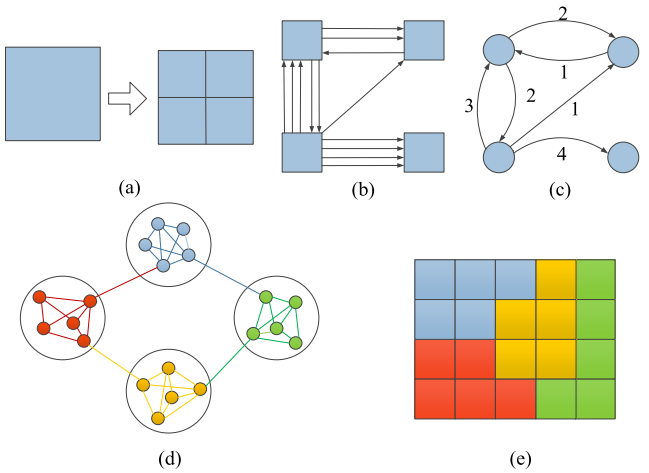


FIGURE 1. To construct a network based on mobility flows, the study area is divided into small regions (a) each small region corresponds to a node in the network. A directed edge or linkage exists between two nodes if there are mobility flows between nodes. The weight of an edge equals the volume of mobility flows represented in (b, c). Graphic (d) provides an illustration of the communities detected from a network, which is divided into four parts (depicted by four circles) in which the sub-networks have relatively dense connections. The community detection result corresponds to closely connected sub-regions (e).

indicate the existence of an edge or a linkage. After extracting mobility flows from the travel trajectory datasets, the volume of mobility flows originating from g_i and ending at g_j is set as the weight w_{ij} from n_i to n_j. Thus, the network is constructed.

As shown in Fig. 1, some nodes have much stronger connections among them than with others. By dividing the network into densely connected sub-networks, the city area is divided into intensely interactive sub-regions. In network science, community detection methods can partition an entire network into tightly connected sub-networks, called communities, and reveal the network clustering characteristics. A community, also called a cluster or a module, is normally considered as a group of nodes which probably share common properties or have similar roles within the network.

Considering the adaption to large-scale transport networks, six representative algorithms of community detection that are adapt for directed-weighted networks are utilized to acquire the spatial community pattern in our constructed networks. The metric of modularity is commonly adopted to measure the performance of network community detection. When applied to weighted and directed networks, the modularity, denoted as Q, is defined as [21]

$$Q = \sum_{i=1}^{m} \frac{w_{ij}}{w} - \frac{w_i^{in} w_i^{out}}{w} \tag{1}$$

Here, w_{ij} is the total weight of links starting and ending in community i, w_i^{in} and w_i^{out} are the total in- and out-weight of links in module i, and w is the total weight of all the links in the network.

To optimize Q, the vast majority of search strategies use one of the following steps to evolve starting partitions: merging two communities, splitting a community into two, moving nodes between distinct communities.

The *fast greedy* [22] algorithm only considers the merging strategy, beginning with each node as the sole member of a community. It only updates the j-th row and column and removes the i-th row and column altogether. The updating process is

$$\begin{cases} \Delta Q'_{jk} = \Delta Q_{ik} + \Delta Q_{jk}, & if\ k \text{ is connected to } i \text{ and } j \\ \Delta Q'_{jk} = \Delta Q_{ik} - 2a_j a_k, & if\ k \text{ is connected to } i \text{ not } j \\ \Delta Q'_{jk} = \Delta Q_{ik} - 2a_i a_k, & if\ k \text{ is connected to } j \text{ not } i \end{cases} \tag{2}$$

where $a_i = d_i/2m$, d_i is the degree of node i, and m is the weight on the edge.

Fast unfolding [23] adopts both strategies of moving nodes and merging communities. The modularity updating process is

$$\Delta Q = \left(\frac{\sum_{in,C} w + w_{i,in}}{2m} - \left(\frac{\sum_{tot} w + w_i}{2m}\right)^2\right) - \left(\frac{\sum_{in,C} w}{2m} - \left(\frac{\sum_{tot} w}{2m}\right)^2 - \left(\frac{w_i}{2m}\right)^2\right) \tag{3}$$

where $\sum_{in,C} w$ is the sum weight of the links inside C, $\sum_{tot} w$ is the sum weight of the links incident to nodes in C, w_i is the sum weight of the links incident to node i, $w_{i,in}$ is the sum weight of the links from i to nodes in C, and m is the sum weight of all links in the network.

Combo [8] involves all three possibilities of optimizing modularity, which is justified as an upper bound to the execution time of $O(N^2 \log(C))$, where N is the number of nodes, and C is the number of communities in the network.

Label propagation algorithm [6], or LPA, is based solely on network structure and does not require optimization of a predefined objective function or prior information about the communities. LPA updates the label of each node according to the labels of its neighbors. Finally, each node is located in the community to which the most neighbors belong. The main idea behind the label propagation algorithm is the following: Suppose that a node x has neighbor nodes $x_1, x_2, \ldots, x_n$ and that each neighbor node has a label denoting the community to which it belongs. Then, each node in the network chooses to join the community to which the maximum number of neighboring nodes belong, and each node is initialized with an unique label and the labels propagate through the network. As the labels propagate at every step, each node updates its label based on the labels of the neighboring nodes. The asynchronous updating is:

$$C_x(t) = f(C_{x_1}(t), \ldots, C_{x_{im}}(t), C_{x_{i(m+1)}}(t-1), \ldots C_{x_{in}}(t-1)) \tag{4}$$

where $x_{i(m+1)}, \ldots, x_{in}$ are neighbor nodes that have not been updated in the current iteration.

Considering a discrete random walk process on the network, the *walktrap* algorithm [7] adopts the flow distance to merge communities. The distance between two

communities is defined as the distance difference from two communities to all other nodes:

$$r_{C_1C_2} = \sqrt{\sum_{i=1}^{n} (P^t_{C_{1i}} - P^t_{C_{2i}})^2 / d(i)} \tag{5}$$

where p^t_{ij} denotes the distance from i to j of t steps. Based on the flow distance definition, the problem of finding communities is a clustering problem, which can be solved using an efficient hierarchical clustering algorithm.

Another random walk based algorithm is the *infomap* [24]. It allocates a binary signature to each node and the Huffman code is adopted to enumerate a succession of locations visited by a random walker. The objective function is

$$\begin{cases} L(M) = q_{\curvearrowright} H(Q) + \sum_i^k p^i H(p^i) \\ q_{\curvearrowright} = \sum_i^k q_{i\curvearrowright} \\ p^i_{\circlearrowright} = q_{i\curvearrowright} + \sum_{\alpha \notin C_i} p_\alpha \end{cases} \tag{6}$$

where $q_{\curvearrowright}$ is the probability of travelling from community i to another community. p_α denotes the probability of visiting nodes in community C_i. $H(Q)$ and $H(P^i)$ denote the entropy of the community code book and the entropy of nodes in the i-th community.

B. CONSISTENCY ESTIMATION

To explain the formation of the spatial communities, the ultimate proof of the hidden reason is to match the spatial communities to POIs distributed among the regions. As shown in Fig. 2, POIs in the studied area are matched with nodes, in accordance to the geolocation using the map matching process. Region r_i is located using the longitude and latitude range (bottom, top)-(left, right). The origin and destination of a movement, as well as a specific POI position, are located using the longitude and latitude. The map matching process determines which region the point is located. After mapping the origins and destinations with partitioned regions, the weights on each directed edge in the constructed network are calculated.

By mapping each POI to the corresponding region, the POI features of each node in the spatial network are obtained. POI features are denoted as $x_i = (x^{(1)}, x^{(2)}, \ldots, x^{(\mathrm{M})})$, where M is the POI category number, and $x^{(\mathrm{j})}$ is the number of the j-th POI category in node i. After applying a CDM to the constructed network, the nodes are partitioned into disjoint sets (communities). Nodes in the same community have the same classification label value Y. Each node in the network is characterized by the POI feature vector. Then, the multi-class classification problem is solved using the stepwise logistic regression method, where the community label Y is set as the dependent variable, and the POI feature is set as the independent variable. Suppose that the value set of Y is $\{1, 2, \ldots, \mathrm{K}\}$, then, the multinomial logistic regression is defined as

$$P(Y=k|x) = \frac{\exp(w_k \cdot x + b)}{1 + \sum_{k=1}^{K-1} \exp(w_k \cdot x + b)}, \quad k = 1, 2, \ldots, K-1 \tag{7}$$

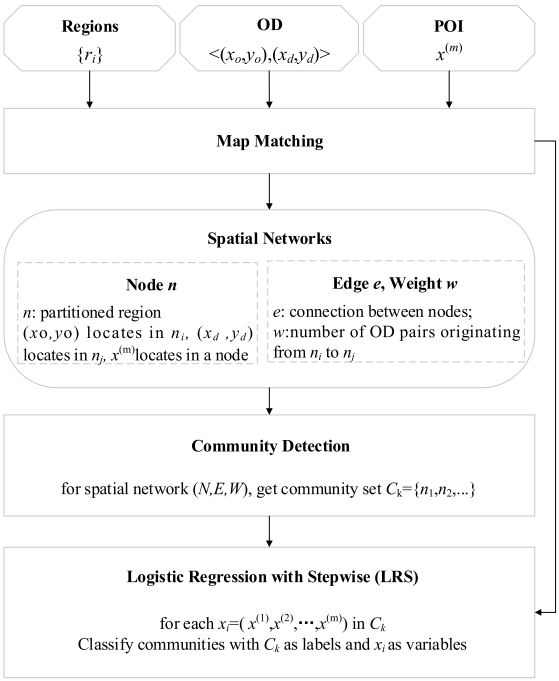


FIGURE 2. The propose onsistency estimation model first sets the partitioned regions as nodes in the network and the connection between nodes as edges weighted by the number of OD pairs. POIs are matched with nodes in the network. Then, a CDM is implemented on the network to obtain the spatial communities. By applying the multi-class classification method of stepwise logistic regression, the POI features are set as independent variables and the spatial community categories are set as classification labels. Regression fitness is adopted to estimate the consistency between spatial communities and POI feature, thus determining the most effective CDM that generates POI driven communities.

$$P(Y = K|x) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(w_k \cdot x)} \tag{8}$$

where $w = w_1, w_2, \ldots, w_M$ and b are model parameters. Given the testing set $D = \{(x_1,y_1),(x_2,y_2), \ldots, (x_\mathrm{N}, y_\mathrm{N})\}$, let D_k denote the samples labeled with k, and $\theta = (w, b)$. Then, the multi-class classification method of the stepwise logistic regression is adopted and the MLE (maximum likelihood estimation) is applied to calculate the parameters:

$$l(\theta_k) = \log P(D_k|\theta_k) = \sum_{x \in D_k} \log P(x|\theta_k) \tag{9}$$

$$\hat{\theta}_k = \arg\max_{\theta_k} l(\theta_k) \tag{10}$$

IV. EXPERIMENT

A. DATASETS

Taking the spatial networks of Beijing and Shanghai as case studies, datasets of taxi GPS trajectories in both cities are collected. The Baidu APIs[1] is used to acquire the datasets of POIs in two metropolises, both containing seventeen categories of POIs. The studied area of Beijing is of longitude and latitude (116.0, 116.8)-(39.65, 40.25), and Shanghai is

[1] http://lbsyun.baidu.com/index.php?title=jspopular.

of longitude and latitude (120.4507, 122.1024)-(30.0022-31.9270). We partition the studied areas into squared grids, each with 1 km^2 in size, using the open street map (OSM)[2].

As shown in Fig. 3, taxi trajectories are used to extract mobility flows. A taxi trajectory is a sequence of GPS points pertaining to the sampling location of the taxi over time. Each point consists of a tuple $< (x, y), f >$ with location (x, y), and the taxi's occupancy status f, where (x,y) is a pair of spatial coordinates representing latitude and longitude. $f = 1$ means the taxi is occupied by passengers, otherwise $f = 0$. The flag f bound to each trajectory position is essential for determining the taxi occupation state, which is utilized to extract the origin and destination of a trip. All other GPS points between a pair of (x_o, y_o) and (x_d, y_d) own the same occupation state, $f = 1$.

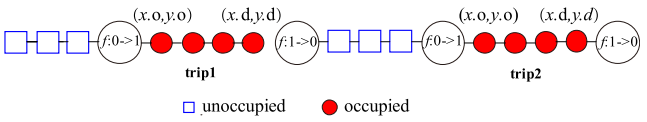


FIGURE 3. Mobility extraction from taxi trajectories with occupation state variation.

The extracted mobility flows consist of 186,2799 OD pairs for Beijing and 38,0640 OD pairs for Shanghai. The mobility volume between any pair of nodes is acquired by matching origins and destinations to the geographical grids using the OSM. Disregarding grids with no OD pairs, 2926 grids remain for Shanghai and 3995 grids remain for Beijing. These grids are set as nodes in the transport networks, and the mobility flow volume originating from grid i to grid j is set as the weight on the directed edge.

The seventeen dimensions of POI features are set as the independent variables X for the logistic regression, and each dimension is set as a component $X^{(i)}$ of the independent variable. The spatial communities are set as classification labels during the multi-class classification process with the stepwise logistic regression. The dataset description of POIs is shown in Tab. 1.

B. RESULTS AND ANALYSIS

The community snapshot is affected by the travel distance. Thus, a distance threshold (DT) is added to the community detecting process. Here in this paper, we focus on the spatial features between nodes in the transport networks, thus the distance threshold refers to the Euclidean distance between grid cells instead of the trip length, which are regarded as nodes in the transport networks. As shown in Fig. 4, for the spatial network of Shanghai, the edge number and mobility flow reach 90% as the distance threshold gradually increases to 20 km and 14 km, respectively. This is similar for the spatial network of Beijing, where the critical distances are 25 km and 9 km.

As the metric of modularity is commonly used for the optimization of CDMs used in this paper, we compare the modularity metric results of different CDMs to directly estimate the community detection result. Besides, as we concentrate on finding the CDMs that are consistent to the POI distribution in a city, the regression fitness measure of R-Square is adopted to measure the consistency between communities and POI features. The modularity of community detection results for the two cities is respectively shown in Fig. 5 and Fig. 6, along with the regression fitness degree metric R-Square. The modularity decreases as the distance threshold increases (except for the *walktrap* and *LPA*, which will be explained later based on the visualization of the detected communities). Larger distance thresholds mean that more edges and more mobility flows are added to the spatial networks, resulting in a smaller modularity value. When the mobility flow proportion is approximately 1, the modularity tends to be convergent as a low number of edges and flows are added to the network.

Combined with the modularity metric, the regression fitness metric, R-Square, is ranked to determine suitable CDMs for generating communities motivated by POIs. Note that the node number in the Beijing network is 1.37 times greater than the node number in the Shanghai network, and the gross mobility flows of Beijing are 4.87 times greater than that of Shanghai. This means that the flow density in the Beijing network is 3.6 times greater than that in the Shanghai network. As shown in Fig. 7, the network scale and flow density affect the value of modularity and R-Square, but does not affect the relative rank measured by both modularity and R-Square (except for *LPA* and *infomap*, which will be explained with the illustration of community detection results). It shows that *combo* has the largest R-Square for both networks. The median value is 0.3 for the Shanghai network and 0.48 for the Beijing network. This indicates that the spatial community is correlated with POI features. The *walktrap* has the lowest modularity for both networks, but the regression fitness, R-Square, of the *walktrap* is just smaller than that of *combo* for both networks. Next, we further analyze the community detection results and the logistic regression fitness combined with the spatial communities' visualization.

As shown in Fig. 5, *infomap* has larger modularity in the Shanghai network than other algorithms. This is explained with the visualization of the community detection results. As shown in Fig. 8(a and b), *infomap* has the largest community in the city center of Shanghai (Fig. 8(a)), which is not the same for Beijing (Fig. 8(b)). As shown in Fig. 8(c), LPA has the most communities in the city center of Shanghai, which is similar to the Beijing network (Fig. 8(d)). According to the algorithm theory of LPA, when a network has sparse edges, it always has a community much larger than the other communities. From the community visualization, we can find that the large proportion of nodes in the largest community leads to increased modularity, and the spatially separated communities lead to the poor regression fitness measured by R-Square, as shown in Fig. 5 and Fig. 6. The mobility flow density of the Beijing network is 466, while it is just

[2]http://www.openstreetmap.org/copyright

TABLE 1. POI categories.

No.	POI Category	Variable	Beijing	Shanghai
1	Beauty	$X^{(1)}$	55,696	51,689
2	Traffic Facility	$X^{(2)}$	96,356	123,124
3	Entertain	$X^{(3)}$	140,445	183,402
4	Enterprise	$X^{(4)}$	128,188	178,562
5	Hospital	$X^{(5)}$	12,924	9,680
6	Real Estate	$X^{(6)}$	300,214	233,777
7	Government	$X^{(7)}$	25,556	20,660
8	Education	$X^{(8)}$	40,381	39,343
9	Culture	$X^{(9)}$	3,971	3,723
10	Scenic Spot	$X^{(10)}$	56,996	48,463
11	Auto Service	$X^{(11)}$	50,898	55,479
12	Living Service	$X^{(12)}$	158,121	149,576
13	Food	$X^{(13)}$	86,301	82,021
14	Shopping	$X^{(14)}$	208,245	230,715
15	Spots	$X^{(15)}$	11,026	9,561
16	Hotel	$X^{(16)}$	8,501	3,704
17	Finance	$X^{(17)}$	22,139	23,386

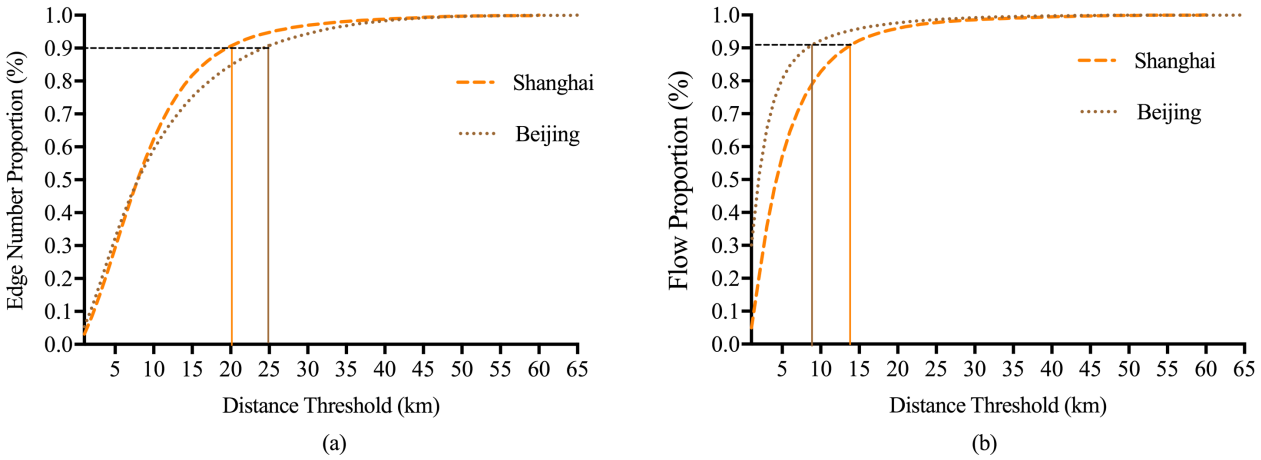


FIGURE 4. Variation of edge number (a) and low (b) with distance threshold changing in the Shanghai and Beijing networks.

130 for the spatial network of Shanghai. Thus, the community snapshots acquired by *LPA* and *infomap* are not stable, which is strongly affected by the edge density or the mobility flow density.

To explain the regression results of *walktrap* and *combo*, we visualize the community detection results of both algorithms. According to the community detection theory of *walktrap*, it uses flow distance as the measure to merge communities for optimizing the modularity metric. The spatial communities with DT=16 and DT=17 identified by *walktrap* are shown in Fig. 9(a) and Fig. 9(b), respectively. This shows that the algorithm merges community pair C_1 and C_2, C_3 and C_4, C_5 and C_6, C_7 and C_8 to $C_{1+2}, C_{3+4}, C_{5+6,}$ and C_{7+8}. Small communities are spatially scattered around the suburban area. While communities detected using *combo* are quite different, meaning that even nodes in the suburban area are connected to the spatially close communities.

Theoretically, to optimize modularity during the community detection process, *combo* adopts the strategies of merging, splitting and moving nodes between existing communities, meaning the algorithm considers each node in every iteration step. Thus, the community patterns found by *combo* are spatially connected, and the regression results are always optimal. With the worst modularity rank for both networks, the R-Square value of *walktrap* ranks only behind *combo*. As shown in Section 3, *walktrap* adopts flow distance to merge sub-communities, meaning that the

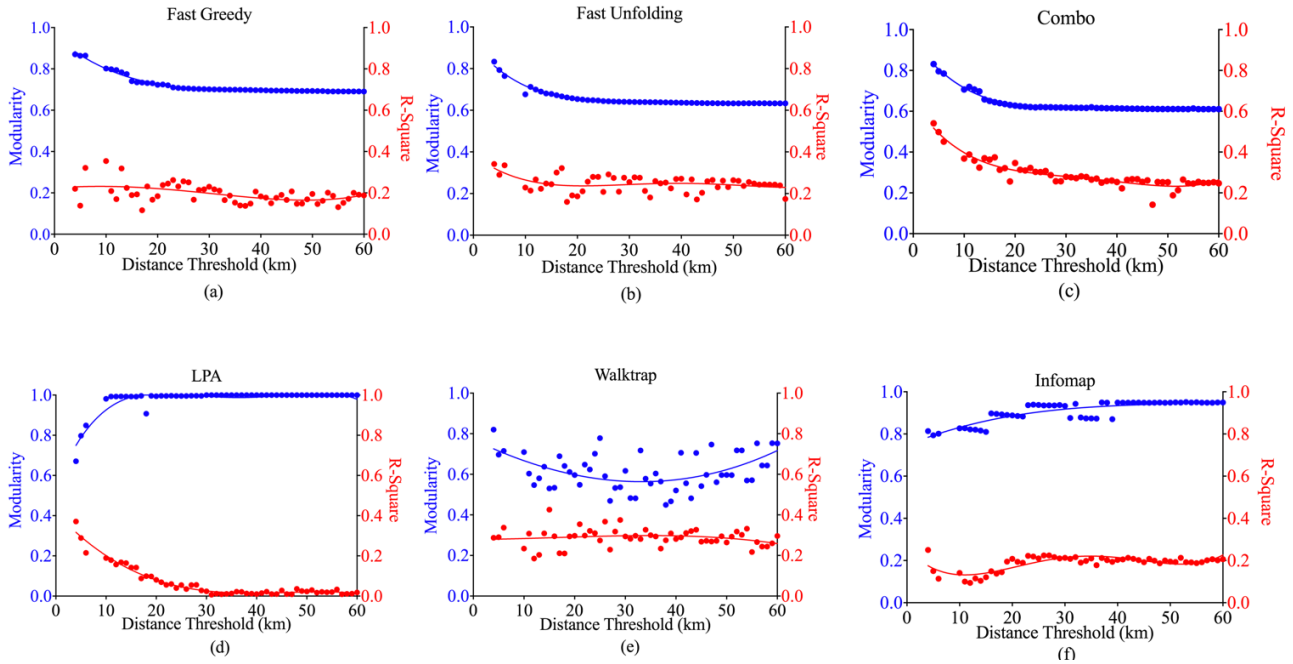


FIGURE 5. Modularity an R-Square acquired by the algorithms: fast greedy (a), fast unfolding (b), combo (c), LPA (d), walktrap (e), and infomap (f) for the networks of Shanghai with distance threshold variation.

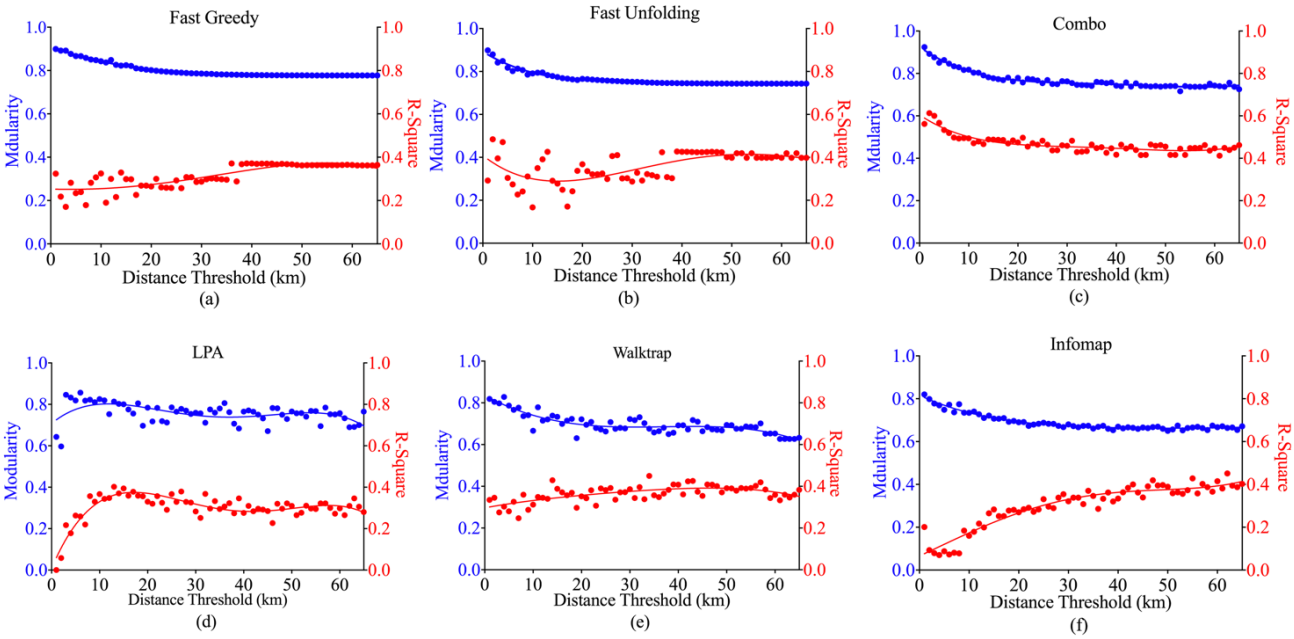


FIGURE 6. Modularity and R-Square acquired by the algorithms: fast greedy (a), fast unfolding (b), combo (c), LPA (d), walktrap (e), and infomap (f) for the networks of Beijing with distance threshold variation.

similarity between two nodes can be described the difference between these two nodes to others. Similarly, the correlation between two urban regions, featured by mobility flows and POIs, can also be measured using flow distance. For example, an urban region functions similarly to another one if two regions connect other regions with a similar amount of mobility flows. Meanwhile, regions covered by the comparable category and number of POIs are of similar functions in the urban daily life, and then they connect to other regions with similar mobility pattern. This may be the reason why *walktrap* has better regression fitness despite having the worst modularity.

For the transport network of Shanghai, with the increased distance threshold, all the community detection results show

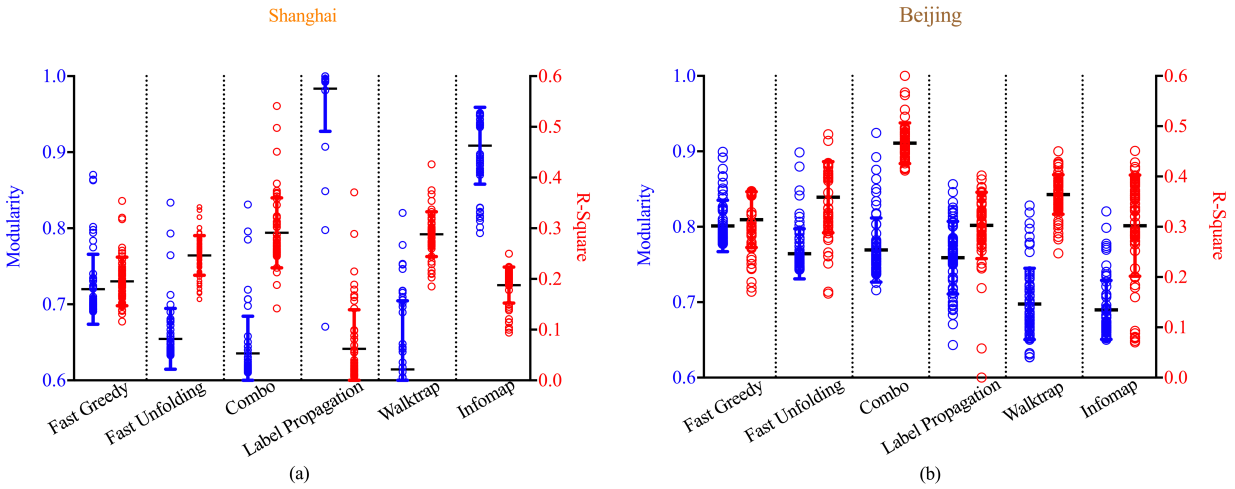


FIGURE 7. Rank of modularity of each algorithm combined with R-Square for the Shanghai network (a) and Beijing network (b).

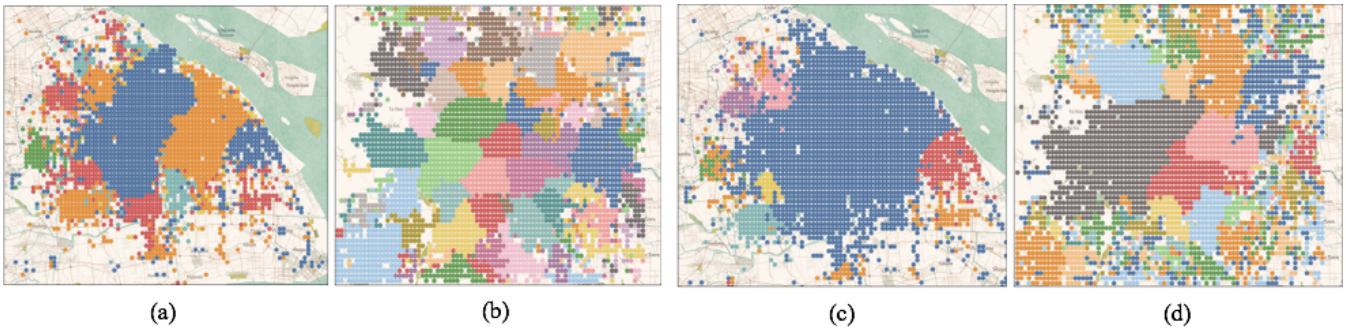


FIGURE 8. Communities obtained using Infomap for the Shanghai network (a) contains the largest community in the city center, which is not contained in the Beijing network (b). LPA also acquires the largest community in the Shanghai network (c), and the largest community in the Beijing network (d).

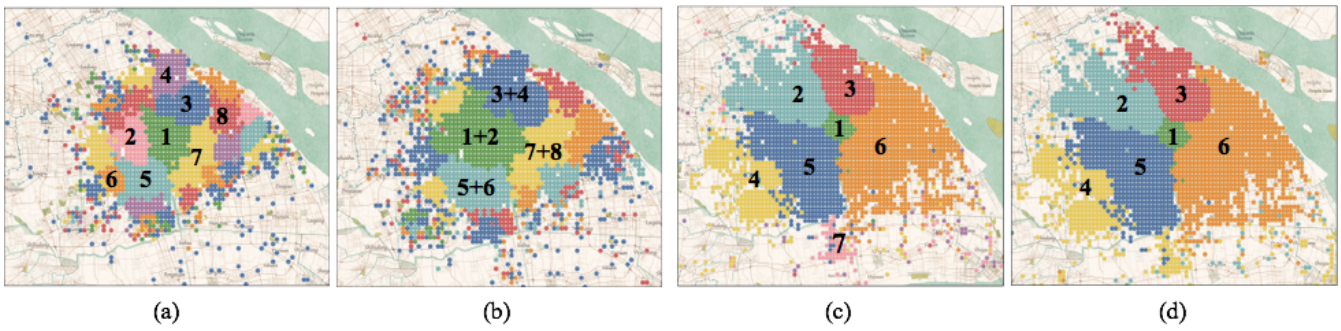


FIGURE 9. Communities etected by walktrap and combo for the spatial network of Shanghai. Communities detected by walktrap with DT=16 and DT=17 are shown in (a) and (b). Communities detected by combo with DT=16 and DT=17 are shown in (c) and (d).

communities spread from the city center to suburban areas. However, communities in Beijing are scattered spatially with similar size. We find that Shanghai and Beijing are polycentric, communities in the spatial networks of Shanghai circle around the city center, and the communities in Beijing are decentralized.

As shown in Fig. 10, we further studied the correlation between modularity and the regression fitness metric R-Square. R-Square presents a positive linear correlation with the modularity of the algorithm *combo*. The median value of R-Square is 0.3 for the Shanghai network using combo and 0.48 for the Beijing network. This further certifies that the community patterns can be explained from the perspective of POIs and communities are correlated with POIs according to the regression results. It can also be found that spatial communities in Beijing are better matched to the POI feature. As the mobility flow density of the Beijing network is much more than that of Shanghai, the better matching result means

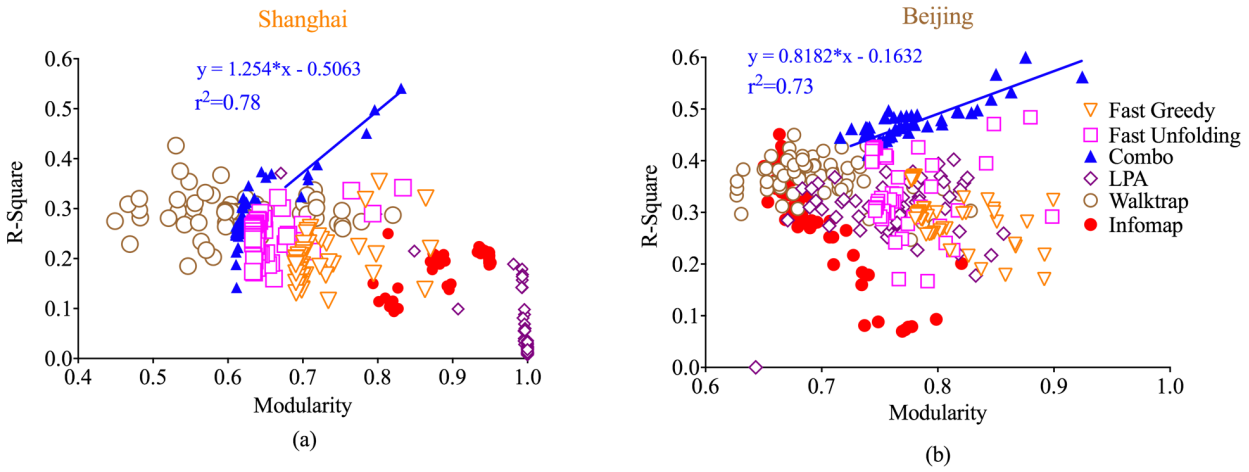


FIGURE 10. Scatter lot of the modularity and R-Square of the six algorithms for the Shanghai network (a) and the Beijing network (b).

more mobility flows can better reflect the spatial communities driven by POIs.

For both networks, the community detection results of LPA are affected by the network scale and edge density. *Combo* has better regression results with larger modularity, while *infomap* has better regression results with smaller modularity. The regression result of *walktrap* is more stable with DT variation, and has a larger R-Square value than that of *fast unfolding*. Comparing the community detection result visualization of these algorithms, it also finds that only the result got by combo is the most stable when the distance threshold varies. Comprehensively, from the perspectives of detecting spatial communities motivated by POIs and community stability in urban transport networks, *combo* is the best choice.

V. CONCLUSION

Researching the spatial communities of mobility flows in urban transport networks is helpful for understanding of urban movement and improving urban planning. Spatial communities in transport networks are rooted in the POI features distributed in the city area. This paper proposes to apply the consistency between network communities of mobility flows and urban POIs to compare the CDMs most suitable for detecting POI driven spatial communities.

Mobility flows of the studied city are collected to construct transport networks with partitioned grids as nodes and the connections between them as links weighted by the mobility volume. POIs are mapped into nodes in the network and are used to characterize each node. Representative community detection algorithms are adopted to explore mobility communities. Then, we use stepwise logistic regression to estimate the consistency between mobility communities and POIs, with the POI feature as an independent vector and the community category as a dependent classification label. Taking the taxi systems of Beijing and Shanghai as case studies, experimental results show that the CDMs, *combo* and *walktrap,* could identify mobility communities that are explained by the POI features, and *combo* is presented as the best CDM.

As this paper only focus on comparing the CDMs from the perspective of POI feature in transport networks, the common measure of modularity is applied to estimate community detection results. In the future, we will compare additional CDMs with our proposed model with other community feature measurement, as well, we intend to employ other mobility data sources, such as the cell-tower traces, for more experimental verification.

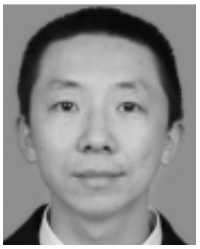
REFERENCES

[1] M. Barthélemy, "Spatial networks," *Phys. Rep.*, vol. 499, pp. 1–101, Feb. 2011.
[2] C. Zhong, S. M. Arisona, X. Huang, M. Batty, and G. Schmitt, "Detecting the dynamics of urban structure through spatial network analysis," *Int. J. Geograph. Inf. Sci.*, vol. 28, no. 11, pp. 2178–2199, 2014.
[3] I. Walde, S. Hese, C. Berger, and C. Schmullius, "From land cover-graphs to urban structure types," *Int. J. Geograph. Inf. Sci.*, vol. 28, no. 3, pp. 584–609, 2014.
[4] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin–destination trips by purpose and time of day inferred from mobile phone data," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 240–250, Sep. 2015, doi: 10.1016/j.trc.2015.02.018.
[5] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003, doi: 10.1137/S003614450342480.
[6] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 3, p. 036106, 2007.
[7] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *J. Graph Algorithms Appl.*, vol. 10, no. 2, pp. 191–218, 2005.
[8] S. Sobolevsky, R. Campari, A. Belyi, and C. Ratti, "General optimization technique for high-quality community detection in complex networks," *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.*, vol. 90, no. 1, p. 012811, 2014, doi: 10.1103/PhysRevE.90.012811.
[9] Z. Du, B. Yang, and J. Liu. (Feb. 2017). "Understanding the spatial and temporal activity patterns of subway mobility flows." [Online]. Available: https://arxiv.org/abs/1702.02456
[10] C. Zhong, E. Manley, S. M. Arisona, M. Batty, and G. Schmitt, "Measuring variability of mobility patterns from multiday smart-card data," *J. Comput. Sci.*, vol. 9, pp. 125–130, Jul. 2015.

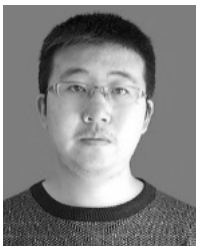
[11] X. Huang, Y. Zhao, C. Ma, J. Yang, X. Ye, and C. Zhang, "TrajGraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 160–169, Jan. 2015.
[12] L. Gong, X. Liu, L. Wu, and Y. Liu, "Inferring trip purposes and uncovering travel patterns from taxi trajectory data," *Cartogr. Geograph. Inf. Sci.*, vol. 43, no. 2, pp. 103–114, 2016.
[13] X. Liu, L. Gong, Y. Gong, and Y. Liu, "Revealing travel patterns and city structure with taxi trip data," *J. Transp. Geogr.*, vol. 43, pp. 78–90, Feb. 2015, doi: 10.1016/j.jtrangeo.2015.01.016.
[14] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti, "Towards a comparative science of cities: Using mobile traffic records in New York, London, and Hong Kong," in *Computational Approaches for Urban Environments. Geotechnologies and the Environment*, vol. 13, M. Helbich, J. J. Arsanjani, and M. Leitner, Eds. Cham, Switzerland: Springer, 2015, pp. 363–387.
[15] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, 2012.
[16] Y. Ren, M. Ercsey-Ravasz, P. Wang, M. C. González, and Z. Toroczkai, "Predicting commuter flows in spatial networks using a radiation model based on temporal ranges," *Nature Commun.*, vol. 5, no. 5, 2014, Art. no. 5347, doi: 10.1038/ncomms6347.
[17] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Sep. 2014.
[18] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in *Proc. Int. Conf. Ubiquitous Comput.*, 2011, pp. 89–98.
[19] S. Sarkar *et al.*, "Effective urban structure inference from traffic flow dynamics," *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 181–193, Jan. 2017.
[20] N. J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 186–194.
[21] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, p. 026113, Feb. 2004.
[22] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 70, no. 6, p. 066111, 2004, doi: 10.1103/PhysRevE.70.066111.
[23] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. (Jul. 2008). "Fast unfolding of communities in large networks." [Online]. Available: https://arxiv.org/abs/0803.0476
[24] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 4, pp. 1118–1123, 2007. [Online]. Available: https://arxiv.org/abs/0707.0609v2

LIPING HUANG (S'11–M'14) born in Liaoyuan, Jilin, China, in 1988. She received the master's degree from the College of Computer Science and Technology, Jilin University, Changchun, China, in 2011, where she is currently pursuing the Ph.D. degree. Her current research interests include trajectory computing, data mining, machine learning, urban computing, complex networks, and traffic data analysis.

YONGJIAN YANG received the B.E. degree in automatization from the Jilin University of Technology, Changchun, Jilin, China, in 1983, the M.E. degree in computer communication from the Beijing University of Post and Telecommunications, Beijing, China, in 1991, and the Ph.D. degree in software and theory of computer from Jilin University, Changchun, Jilin, China, in 2005. He is currently a Professor and a Ph.D. Supervisor with the College of Computer Science and Technology, Jilin University. He participated in four projects of NSFC, 863 and funded by the National Education Ministry for Doctoral Based Foundation. As the first author, he has published over 60 papers in national and foreign journals. His recent research interests include theory and software technology of network intelligence management, crowd sensing, and data mining in urban computing field.



HEPENG GAO received the B.E. degree in software engineering from the College of Software, Jilin University, Changchun, China, in 2015, where he is currently pursuing the master's degree. His research interests include complex networks, data mining, machine learning, and traffic big data analysis.



XUEHUA ZHAO received the Ph.D. degree from the College of Computer Science and Technology, Jilin University, in 2014. He is currently a Lecturer with the School of Digital Media, Shenzhen Institute of Information Technology. His main research interests are related to machine learning and data mining.

ZHANWEI DU was born in 1988. He received the Ph.D. degree from the Department of Computer Science and Technology, Jilin University, in 2015. He is currently a Post-Doctoral Fellowship with The University of Texas at Austin. He held a postdoctoral position (Hong Kong Scholar) for one year at Hong Kong Baptist University under the supervision of Prof. J. Liu. His current research interests include complex networks, smart city of traffic network, and Epidemic disease propagation.