IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 9, 2018, accepted April 4, 2018, date of publication April 19, 2018, date of current version May 9, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2828032

Building a Spatially-Embedded Network of Tourism Hotspots From Geotagged Social Media Data

XINYU WU[1,2,3], ZHOU HUANG[1,3], XIA PENG[4], YIRAN CHEN[1,3], AND YU LIU[1,3]

[1]Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing 100871, China
[2]Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources, Shenzhen 518034, China
[3]Beijing Key Laboratory of Spatial Information Integration and Its Applications, Peking University, Beijing 100871, China
[4]Collaborative Innovation Center of Tourism, Tourism College, Beijing Union University, Beijing 100101, China

Corresponding author: Zhou Huang (huangzhou@pku.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0503602, in part by the National Natural Science Foundation of China under Grant 41771425, Grant 41625003, and Grant 41501162, in part by the Beijing Philosophy and Social Science Foundation under Grant 17JDGLB002, and in part by the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources, under Grant KF-2016-02-023.

ABSTRACT The rapid development of social media and location-based service has generated a myriad of spatial data tagged with geo-information. Constructing a network of tourism hotspots using these geotagged data would improve our understanding of tourism activities. Thus, using Flickr data, we built a spatially-embedded tourism hotspot network for Beijing and applied complex network analysis to study the network characteristics. The results indicate that the tourism hotspot network in Beijing is scale-free and small-world. In the hotspot network, the interconnected triplets have a tendency to be formed by the edges with greater weight values, and a high-weighted edge is often connected by two high-degree vertices. Moreover, the statistics of the network provides insights for additional travel bus routes in Beijing. Finally, this paper provides a guide for building spatially-embedded hotspot networks based on geotagged social media data, which helps to understand the laws of travel and provides decision support for the development of tourism resources.

INDEX TERMS Tourism hotspot network, complex network, geotagged data, social media, big data.

I. INTRODUCTION

The mobile Internet and social media have developed rapidly in recent years. When travelling, tourists typically upload photos, text, videos and other data to the Internet, recording their travel behaviors thereby. In addition to being rich in text- and image-based information, social media data are also rich in geo-information. Both tourism hotspots and travel trajectories of individuals could be extracted from geotagged social media data [1]–[3]. Geotagged social media data enable a new environment to observe travel behaviors (e.g., popular attractions and routes) from a large number of travelers.

Complex network theory, which is widely used in geographical studies [4], [5], provides a new perspective to investigate human mobility patterns based on social media data. Numerous trajectories extracted from social media data provide a basis to construct a spatially-embedded network of tourism hotspots. Therefore, applying network theory to large amounts of social media data containing geo-information represents a powerful method to examine the characteristics of tourism networks, which helps us better understand travel behaviors.

Many complex real-world systems can be described in the form of networks, including the World Wide Web [6]–[8], the Internet [9], [10], and social networks [11]–[13]. A network, which is also called a graph, consists of vertices or nodes and edges or links. Given that tourism has become one of the most significant forces for change in the world [14], the application of complex network theory to tourism geography has the potential to reveal the complex characteristics of tourism hotspot networks and to realize the multi-view deep perception of places, routes, and networks. We are capable of exploring the space-time distribution patterns and laws of tourism hotspots. Furthermore, it is expected to provide insights for the recommendation of attractions, route prediction and other studies.

Beijing is the political, cultural and scientific center of China and contains more than 200 tourist attractions. Flickr is an image- and video-hosting service used by travel enthusiasts from all around the world. Whereas Flickr is not used extensively by Chinese users, a large number of foreign users who travel to China post photos and videos on Flickr, making it possible to use Flickr data in Beijing to study travel behaviors especially for foreigners.

In this study, we extracted tourist attractions from geotagged Flickr data in Beijing and utilized the travel trajectories of users to construct a spatially-embedded tourism hotspot network and then evaluated its characteristics. The main contributions of this paper are: 1) a method for constructing a spatially-embedded tourism hotspot network based geotagged social media data; and 2) an analysis of the characteristics of the spatially-embedded tourism hotspot network. The results are expected to provide insights for the identification and development of attractions, routes prediction and travel bus route design.

The remainder of this paper is organized as follows: Section II illustrates related work; Section III elaborates on the method used to construct the tourism hotspot network; Section IV details the characteristics of the tourism hotspot network; and Section V describes conclusions and future work.

II. RELATED WORK

With the rise in the popularity of social media and the emergence of big data, many scholars have exploited social media data to build complex network models and uncover the characteristics of complex networks [15]–[18]. For instance, Centola [16] studied how social networks affect the spread of behavior. Mislove *et al.* [18] demonstrated the power-law, small-world and scale-free characteristics of online social networks by retrieving data from more than 11.3 million users and 328 million links of Flickr, YouTube, LiveJournal and Orkut. Kumar *et al.* [19] presented a model of network growth for the Flickr and Yahoo! 360 online social networks communities.

Complex networks have been successfully applied in tourism research as well. Miguéns and Mendes [20] discussed the importance of weights on the network connections by analyzing the global travel network. Baggio *et al.* [21] summarized the application of network science in tourism research and concluded that network science methods are highly valuable for enhancing our understanding of tourism systems. Baggio and Cooper [22] demonstrated the utility of network analysis in helping deliver tourism destinations competitiveness.

Combining complex network science with tourism can help people clearly understand changes in tourism activities and the relationships among tourism elements. It can also help people establish a cognitive system for tourism economics, sociology and geography. Finally, applying complex network theory to tourism contributes to identifying and designing tourism hotspots and providing insights related to recommended travel routes, the development and protection of tourism resources and the construction of tourism facilities.

As a conclusion, tourism research based on complex network theory primarily involves the collaborative patterns of tourism researchers, the construction and investigation of tourism destination networks based on public data, and the exploration of tourism research methodology. In addition, some scholars have constructed user relationships in complex networks based on social media data and studied the social network characteristics, growth model and propagation model.

On the other hand, increasing numbers of scholars have begun to study the geo-information contained in social media data. Researchers applied these data to the identification of urban centers [23], geopolitics [24], public safety [25], the identification of photo locations [26] and other fields. However, slight studies have examined tourists travel behaviors by using complex network methods to analyze geotagged social media data. In fact, the rich geographic information contained in geotagged social media data has given people a great opportunity to study the establishment of spatially-embedded tourism hotspot network, explore travel laws and provide novel services such as travel recommendation and travel route planning.

III. CONSTRUCTION OF THE NETWORK MODEL

Network construction involved three steps: (1) preprocessing the data and removing redundant data; (2) clustering; and (3) constructing the topological relationship among hotspots and building the hotspot network.

A. DATA PREPROCESSING

Flickr provides a free application programming interface that allows developers to access data. This study used metadata from 213,938 geotagged photos taken in Beijing, China, from January 1, 2005 to January 1, 2016 from 22,354 users worldwide. After removing the distorted and redundant photos, data from 185,531 photos remained. Figure 1 illustrates the spatial distribution of Flickr photos in Beijing.

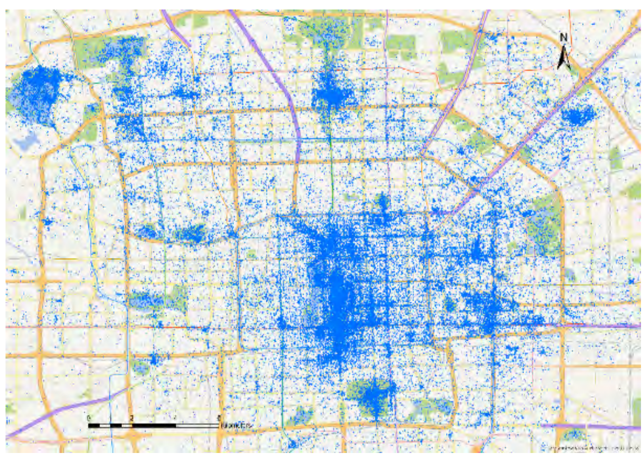


FIGURE 1. Spatial distribution of Flickr photos.

B. CLUSTERING METHOD AND RESULTS

There are several approaches for clustering big sensor data [27]–[29]. To achieve spatially clustering the geotagged Flickr photos, we used a novel clustering algorithm named Clustering by Fast Search and Find of Density Peaks (CFSFDP) [30], which is based on two assumptions: 1) cluster centers are surrounded by neighbors with lower local density; and 2) centers are at a relatively large distance from any points with higher local density. The clustering process in the CFSFDP algorithm proceeds as follows.

1) COMPUTING LOCAL DENSITY AND DISTANCE

The local density ρ_i of each point and its distance δ_i from points of higher density are computed. Both these quantities depend only on the distances d_{ij} between data points, which are assumed to satisfy the triangular inequality. The local density ρ_i of data point i is defined as

$$\rho_i = \Sigma j \chi \left(d_{ij} - d_c\right) \tag{3.1}$$

where $\chi(x) = 1$ if $x < 0$, and $\chi(x) = 0$ otherwise, and d_c is a cutoff distance. Basically, ρ_i is equal to the number of points that are closer than d_c to point i. The algorithm is sensitive only to the relative magnitude of ρ_i at different points; this implies that for large data sets, the results of the analysis are robust with respect to the choice of d_c. Then, δ_i is determined by computing the minimum distance between point i and any other point with higher density:

$$\delta_i = \min_{j:\rho_j > \rho_i} d_{ij} \tag{3.2}$$

In this step, the cutoff distance d_c has a great effect for clustering results, and is determined by the prior knowledge. If the d_c value is set too high, the final cluster will be too large; otherwise, the cluster will be too small. Hence, a moderate value is appropriate. We recommend that d_c be set to 10-50 meters, and the specific value should be adjusted according to the total density of the point set to be classified. If the overall point density is high, this value can be set lower; otherwise, it will be set higher.

2) NOISE FILTERING

The region density threshold ρ_{thr} is determined in this step. If the region density value ρ_i of one point is smaller than the threshold value ρ_{thr}, this point is considered noise and is not considered when determining the cluster center.

3) NORMALIZED PROCESSING TO OBTAIN DECISION VALUES

The region density ρ_i and distance δ_i are calculated for each point. These values are then used to obtain the normalized region density ρ_i^* and distance δ_i^*. The decision values ω_i are calculated as $\omega_i = \rho_i^* \cdot \delta_i^*$. As mentioned before, the larger the decision value ω of point has after removing noise points, the more suitable it is to be selected as the cluster center.

4) GENERATING CLUSTERS

After the cluster centers are determined by ω_i, the remaining points are classified: an unclassified point p_i belongs to the category of the point whose distance from p_i is δ_i; after recursion any point would be assigned to the extracted clusters.

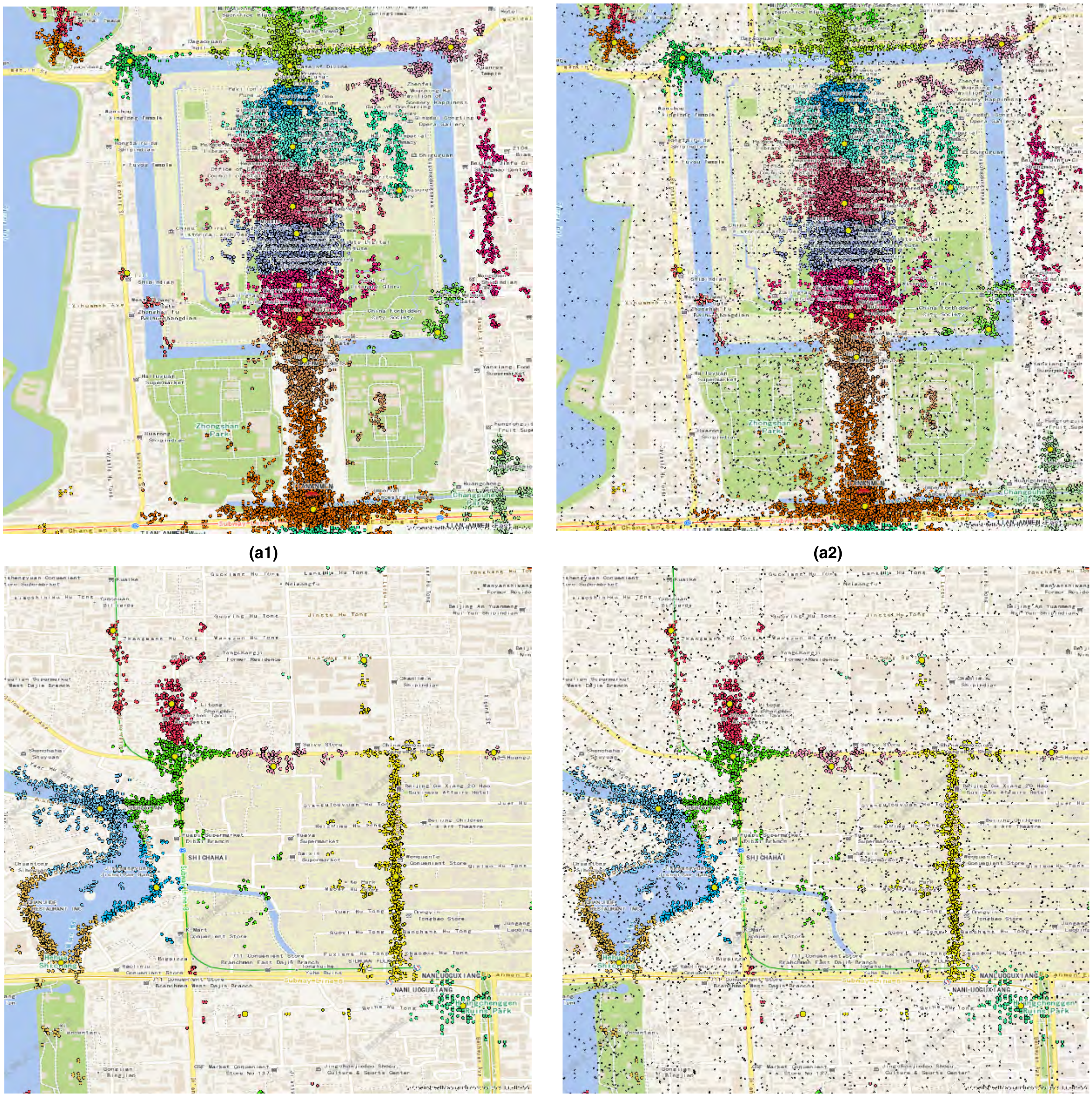
Compared with the traditional spatial clustering method such as DBSCAN, this clustering approach is significantly higher in classification accuracy, enables distinguishing adjacent high-density areas, and has better adaptability in the case of uneven density distribution [31]. Using this approach, 243 clusters i.e. 'natural' hotspots in Beijing were retrieved. In addition to tourist attractions, we also retain some of the hotspots closely related to travel, such as airports, hotels, shopping malls and so on. Thus, there were 221 hotspots as the data basis for building a spatially-embedded tourism hotspot network. Two parts of the Beijing's clustering results are shown in Figure 2.

C. BUILDING THE NETWORK MODEL

We retrieved 221 hotspots, which are called vertices in the network. Through clustering, the mapping between the user's historical check-in and the tourism hotspot was established. In accordance with the chronological order, we generated each user's trajectory such as {Nanluoguxiang → Tiananmen Square → Palace Museum → Lama Temple → Summer Palace}. Next, we considered two hotspots where one user travels consecutively as one link between them. These two hotspots are considered to be a hotspot pair and have only one undirected connection between them. When a user accesses a hotspot pair, tourist frequency (regardless of direction) on the edge between the two hotspots increases by 1. Then, we assigned tourist frequency as weights of edges in tourism hotspot network. Thus, based on the extracted hotspots and the topological links between them, a weighted and non-directed network with 221 vertices and 3135 edges was constructed.

The degree of network vertices ranges from 1 to 147, and the built-up tourism hotspot network is visualized as Figure 3. The spatially-embedded tourism hotspot network is explicitly overlaid on the geographic map, with most edge weights less than or equal to 10.

From the point of view of computational complexity, clustering is the most time-consuming sub-process in network construction. As for the clustering algorithm, the main calculation steps are "computing local density and distance" and "generating clusters". When computing local density and distance, there is a need to calculate the distance from data point i to all other points, so the time complexity of this step is $O(n^2)$. When generating clusters, it needs to iterate through all the points, so the time complexity of this step is $O(n)$. In addition, as for the sub-process of preprocessing and the sub-process of constructing the topological relationship, all points need to be accessed, so the time complexity is $O(n)$ as well.



(a1) (a2) (b1) (b2)

FIGURE 2. **Two parts of the clustering results. (a1) The Palace Museum without noise. (a2) The Palace Museum with noise. (b1) Drum and Bell Tower-Shichahai without noise. (b2) Drum and Bell Tower-Shichahai with noise.**

IV. NETWORK CHARACTERISTICS

A. SCALE-FREE CHARACTERISTICS

To describe the characteristics of the tourism hotspot network, the following statistics are calculated.

The degree k_i of vertex i is defined as the number of edges connected to the vertex.

The degrees of vertices in weighted networks indicate the topology of the network, that is, the most intuitive topological measure of centrality [13]. The average degree of all vertices in the network is the arithmetic mean of all vertex degrees, defined as

$$\langle k \rangle = \frac{1}{N} \Sigma i = 1^N k_i \tag{4.1}$$

where N is the number of vertices. The average vertex degree reveals the universal state in which all vertices are connected to others in a network.

The statistical results indicate an extremely uneven distribution of tourism hotspots in Beijing; the average

TABLE 1. **Vertices ranking in top 10 in degree, strength and pressure.**

Rank	*Attraction_ Strength*	*Strength*	*Attraction_ Degree*	*Degree*	*Attraction_Pressure*	*Pressure*
1	***Palace Museum***	*6037*	*Sanlitun Swire*	*143*	***Palace Museum***	*50.31*
2	***Tiananmen Square***	*5756*	***Palace Museum***	*120*	***Tiananmen Square***	*48.37*
3	***Temple of Heaven***	*2491*	*Nanluoguxiang*	*119*	***Temple of Heaven***	*24.91*
4	***Summer Palace***	*2031*	***Tiananmen Square***	*119*	***Summer Palace***	*19.34*
5	***Shichahai***	*1678*	***Shichahai***	*117*	*Jingshan Park*	*18.84*
6	*Jingshan Park*	*1564*	***Beijing Olympic Park***	*112*	***Shichahai***	*14.34*
7	***Beijing Olympic Park***	*1403*	***Summer Palace***	*105*	*Beihai Park*	*14.06*
8	*Sanlitun Swire*	*1395*	*Lama Temple*	*103*	*Jiaolou of the Forbidden City*	*12.9*
9	*Beihai Park*	*1181*	*Beijing 798 Art Zone*	*101*	*Drum Tower*	*12.56*
10	*Qian Men*	*1044*	*Capital International Airport*	*100*	***Beijing Olympic Park***	*12.53*
			Temple of Heaven	*100*		

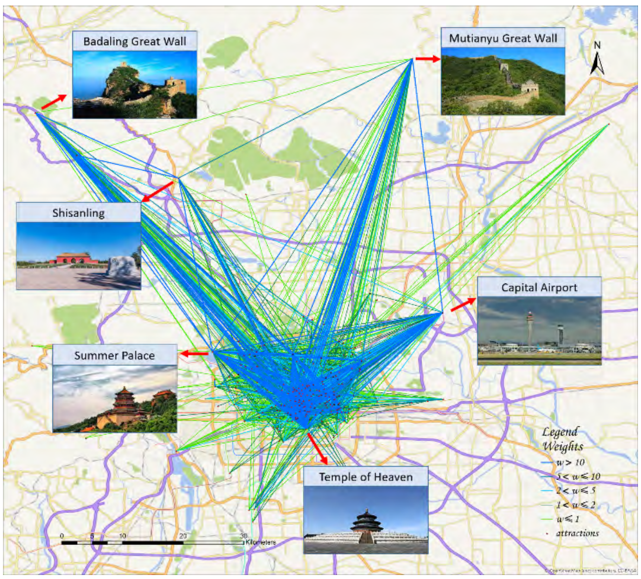


FIGURE 3. **Tourism hotspot networks in Beijing.**

degree $\langle k \rangle$ is 28.37, while the maximum vertex degree (Sanlitun Swire) is 143. The degrees of the top 20 hotspots in the network are all greater than 80, and the vertices of these high-degree hotspots provide important connections within the spatially-embedded tourism network. These high-degree vertices contain many famous attractions (e.g., Beijing Nanluoguxiang, Tiananmen Square, Lama Temple, Palace Museum and Summer Palace), entry-exit transportation hubs (e.g., Beijing Capital International Airport), Sanlitun Swire, Yintai Center, Wangfujing, Beijing 798 Art Zone, National Center for the Performing Arts and other areas that integrate shopping, art, business facilities and hotels. These vertices are visited frequently by tourists and are closely related to the other vertices in the network, making it easier to connect with other vertices.

A more significant measure of the weighted network is the vertex strength s_i [13], defined as

$$s_i = \Sigma j \in N_i \omega_{ij} \tag{4.2}$$

where N_i is the set of vertices connected to vertex i, and ω_{ij} is the weight of edge e_{ij} that connects vertices i and j together.

The vertex strength, which is a localized synthetic measure of the vertex, indicates the topology of vertices, as well as the characteristics of edges. It also measures the popularity of attractions in the tourism hotspot network.

The vertex pressure is calculated as the ratio of strength s_i and degree k_i, defined as

$$p_i = \frac{s_i}{k_i} \tag{4.3}$$

The vertex pressure is the average weight of the edges, which indicates the general popularity of the edges connected to the vertex.

As shown in Table 1, the vertices which rank in top 10 in degree, strength and pressure are listed, respectively. The attractions whose name in bold font (i.e., Tiananmen Square, the Palace Museum, Temple of Heaven, the Summer Palace, Shichahai and Beijing Olympic Park) rank in top 10 in all three statistical indicators. Then, all popular attractions are spatially exhibited in Figure 4, making readers aware of their geographical distribution. It's observed that most popular attractions in Beijing are located in the downtown area within the second ring road.

The log-log plots of cumulative frequency versus vertex degree, vertex strength, vertex pressure and edge weight are shown in Figure 5, respectively. If the distribution is a power-law distribution, the curve fitted in the log-log plot should be a straight line. It's observed that the distributions were fitted with four linear regression equations with adjusted coefficients of determination of $R^2_{(a)} = 0.7256$, $R^2_{(b)} = 0.9159$, $R^2_{(c)} = 0.9865$, $R^2_{(d)} = 0.9870$. The plots

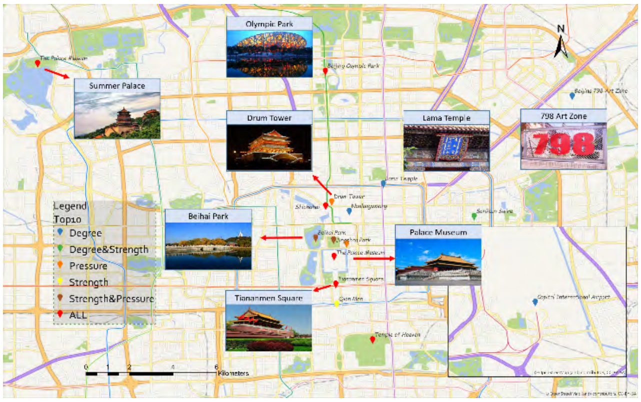


FIGURE 4. Thematic map of attractions ranking in top 10 in indicators of degree, strength and pressure.

present power-law distributions, especially with a downward bend in the tail in 5(a) and 5(b). Amaral *et al.* [32] found similar downward bends in the degree distribution of power station networks in California. The explanation is that distance has a large impact on the connections in the spatially-embedded network. Vertices (i.e., attractions or power stations) are spatially scattered, and the vertex tend to be connected with nearby vertices, so the greater the degree value, the less the vertices.

In geography, the interaction between objects decreases as the distance increases [33]; in the tourism hotspot network, a vertex is more likely to be connected to a proximal vertex (i.e., two places are visited consecutively). In other words, if the next destination is near the current location, the probability of visiting that destination is high. In contrast, if the next destination is far away from the current location, the probability of access is low. Although in modern society the convenience of traffic reduces the cost of travelling between attractions, the "downward-bend" fact indicates that the distance effect is still an important factor in the formation of spatially-embedded network topology.

The above results indicate that the distribution of tourism hotspots and links of them in Beijing generally follows a power-law distribution, and the tourism hotspot network has obvious scale-free characteristics.

B. CLUSTERING COEFFICIENT

The clustering coefficient is divided into a local clustering coefficient and a global clustering coefficient.

The clustering coefficient in the unweighted networks characterizes both the local and global topological properties of the network. Suppose that the degree of a vertex i in an unweighted network is k_i; that is, there are k_i vertices connected to the vertex, and these k_i vertices are called the adjacent vertices of vertex i. Then, $C_{k_i}^2$ is theoretically the largest number of edges between the k_i vertices. The actual number of edges between the k_i vertices is E_i.

The local clustering coefficient of unweighted networks is the ratio of E_i and $C_{k_i}^2$ [12], defined as

$$C_i^u = \frac{E_i}{C_{k_i}^2} = \frac{2E_i}{k_i\,(k_i-1)} \tag{4.4}$$

The global clustering coefficient is measured based on the vertex triplets. Triplet is divided into the closed triplet and the open triplet.

The global clustering coefficient of unweighted networks [34] is defined as

$$C^u = \frac{N_{ct}}{N_t} = \frac{N_{ct}}{N_{ct}+N_{ot}} \tag{4.5}$$

where N_{ct} is the number of the closed triplets in the network, N_{ot} represents the number of the open triplets, and N_t is the sum of the two.

Considering the influence of the edge weights on the formation of network topology, the clustering coefficient in weighted networks are extended as follows.

The local clustering coefficient of weighted networks [13] is defined as

$$C_i^w = \frac{1}{s_i\,(k_i-1)}\Sigma h, j\frac{(w_{hi}+w_{ij})}{2}a_{hi}a_{ij}a_{jh} \tag{4.6}$$

where $a_{ij} \in$ A, A is the adjacency matrix of the network; w_{ij} indicates the weight of edge connected to vertex i and vertex j.

Obviously, only vertex j and vertex h, which constitute a closed triplet with vertex i, and weights of edges e_{hi} and e_{ij} are involved with computation of the local clustering coefficient. Thus, the value of C_i^w is between 0 and 1. The local clustering coefficient in unweighted networks implies the probability of that "friends" of vertex i are also "friends" with each other. In weighted networks, the local clustering coefficient covers not just the number of closed triplets in the neighborhood of the vertex, but also their total weight relative to the strength of the vertex [13].

The global clustering coefficient of weighted networks [35] is similar to that of unweighted networks, defined as

$$C^w = \frac{W_{ct}}{W_t} = \frac{W_{ct}}{W_{ct}+W_{ot}} \tag{4.7}$$

where w_{ct} is the total weight of the closed triplets in the network, w_{ot} represents the total weight of the open triplets, and w_t is the sum of the two.

Then, $C^w(k)$ is defined as the average of the weighted local clustering coefficient over all vertices with degree k, and $C^u(k)$ is defined as the average of the unweighted local clustering coefficient over all vertices with degree k. The measure of $C^w(k)$ provides global information on the correlation between topology and weights. As shown in Figure 6, it's observed that almost 73% of $C^w(k)$ values are greater than that in the unweighted network, and only 15% of $C^w(k)$ values are less than the $C^u(k)$ value. In addition, global clustering coefficients C^w and C^u were calculated for the tourism hotspot network, with values of 0.6643 and 0.4469,

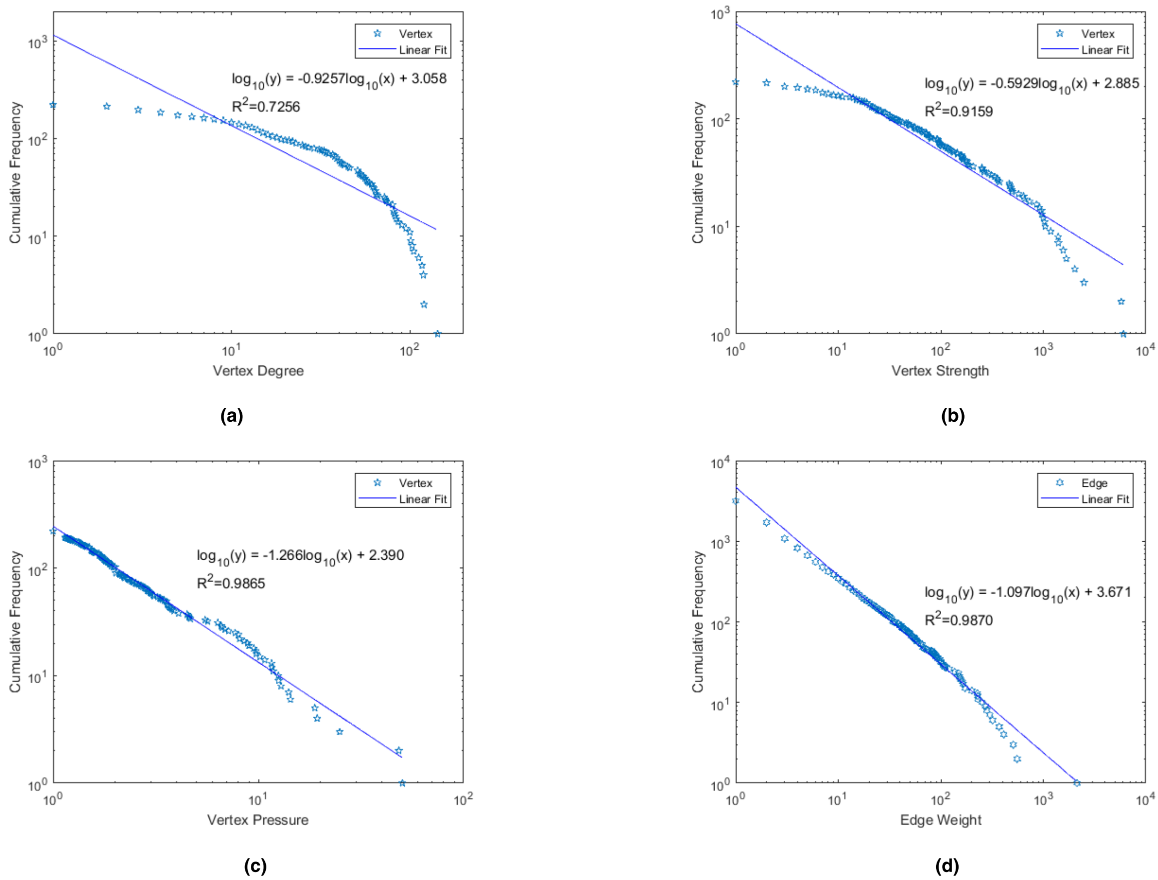


FIGURE 5. **Log-log scatter plots of cumulative frequency versus vertex degree, vertex strength, vertex pressure and edge weight. (a) Degree-frequency. (b) Strength-frequency. (c) Pressure-frequency. (d) Weight-frequency.**

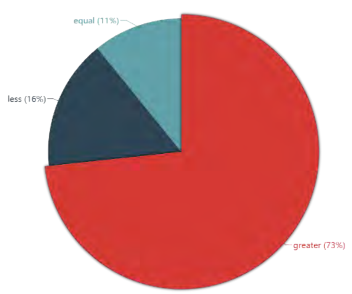


FIGURE 6. **$C^w(k)$ versus $C^u(k)$ for the tourism hotspot network.**

respectively (see Table 2). In most cases, $C^w(k)$ is greater than $C^u(k)$, and C^w is greater than C^u, that is, the interconnected triplets are more likely formed by the edges with greater weights.

C. SMALL WORLD CHARACTERISTICS

In the graph theory, there is an illustrious conjecture named ''six degree of separation'', that is, the small-world theory. The small-world characteristic is one of the most important features of complex networks. In general, quantities such as the clustering coefficient and the average path length are utilized to illustrate small-world characteristics of unweighted networks. However, the weight of edges should not be neglected when portraying the small-world characteristics of weighted networks. Thus, after defining the shortest path length between vertices, a quantity, i.e. efficiency [36], is introduced to determine whether a weighted network is small-world.

TABLE 2. **The global efficiency and local efficiency of unweighted and weighted tourism hotspot networks.**

	E_{glob}	E_{loc}
Tourism Network (unweighted)	0.5352	0.7777
Tourism Network (weighted)	0.6049	0.9523

1) SHORTEST PATH LENGTH

There are several approaches to identify the shortest path in weighted networks [37], [38]. Dijkstra [37] proposed an algorithm to find the path of least resistance, and the edge weight represents the cost of transmitting [39]. However, in the spatially-embedded tourism hotspot network, the edge weight, i.e. the frequency of visits, should not be resistance. Conversely, the greater the weight of the edge, the less resistance to travel between the two vertices. Therefore, numerical conversion should be performed to explore the shortest path length. One of the most prevailing method is to invert the tie weights [40], [41]. In addition, Opsahl *et al.* [39] extended a shortest path algorithm by taking into account the number of intermediary vertices, which is adopted to compute the shortest path length in this study.

The shortest path length [39] is defined as

$$d^{w\alpha}(i,j) = \min_{\substack{n \in p_{ij} \\ p_{ij} \in P_{ij}}} \left(\frac{1}{(w_{ih})^{\alpha}} + \cdots + \frac{1}{(w_{hj})^{\alpha}} \right) \tag{4.8}$$

where P_{ij} is the set of all reachable path between vertex i and vertex j, p_{ij} represents one of the paths, h is an intermediary vertex, and α is a tuning parameter (assumed at 0.1 in order to smooth the difference between the edge weights).

2) EFFICIENCY

A small world network has a high global efficiency and a high local efficiency, which indicates that it is efficient in both global and local communication [36]. The global efficiency and the local efficiency of weighted networks are defined as follows.

The global efficiency [36] of the weighted network is defined as

$$E_{glob} = E(\mathbf{G}) = \frac{1}{N(N-1)} \Sigma i{\neq}j{\in}\mathbf{G} \frac{1}{d_{ij}} \tag{4.9}$$

where $\mathbf{G}$ indicates the whole network, N represents the number of vertices in $\mathbf{G}$, and d_{ij} is calculated by Equation 4.8.

Suppose the local subgraph $\mathbf{G}_i$ is formed by the neighbor vertices of vertex i, the local efficiency E_{loc} is defined as the average efficiency of the local subgraphs [36]

$$E_{loc} = \frac{1}{N} \Sigma i \in GE(\mathbf{G}_i) \tag{4.10}$$

Even if considering the tourism hotspot network as an unweighted one, that is, the weight of all edges is 1, the network is still high efficient at global and local levels, with global efficiency value 0.5352 and local efficiency value 0.7777. When taking into account the actual weight, E_{glob} increases to 0.6049, and E_{loc} increases to 0.9523. The results indicate that the tourism hotspot network is highly fault-tolerant in addition to having the small-world characteristic. This also means closure of few attractions would not have a damaging impact on the overall structure of the tourism network.

D. ASSORTATIVE NETWORK

In an unweighted network, if vertices with high degrees tend to be connected with other high-degree vertices, the network is said to have positive degree-degree correlation and is called an assortative network. On the other hand, if high-degree vertices tend to be connected with low-degree vertices, the network is said to have negative degree-degree correlation and is called disassortative network.

In order to quantify the assortativity of an unweighted network, Newman [42] called degree-degree correlation as mixing pattern and presented a method to calculate the Pearson correlation coefficient, which is defined as the assortativity coefficient of the network.

The Pearson correlation coefficient (i.e. assortativity coefficient) of unweighted networks is defined as

$$r = \frac{M^{-1}\Sigma e_{ij}{\in}E k_i k_j - \left(M^{-1}\Sigma e_{ij}{\in}E\left(k_i+k_j\right)/2\right)^2}{M^{-1}\Sigma e_{ij}{\in}E\left(k_i^2+k_j^2\right)/2 - \left(M^{-1}\Sigma e_{ij}{\in}E\left(k_i+k_j\right)/2\right)^2} \tag{4.11}$$

where M is the total number of edges of the network, E is the edge set of the network, and k_i and k_j are the degrees of the two vertices v_i and v_j of the edge e_{ij}.

The degree Pearson correlation coefficient r is in the range of $-1 \leq r \leq 1$. When r is negative, the network is negatively correlated (i.e., the network is disassortative). Alternatively, when r is positive, the network is positively correlated (i.e., the network is assortative). When r is zero, the network is not correlated.

Then, the Pearson correlation coefficient can be extended and applied to the weighted network as follows.

The weighted assortativity coefficient [43] is defined as

$$r^w = \frac{H^{-1}\Sigma e_{ij}{\in}E w_{ij} k_i k_j - \left(H^{-1}\Sigma e_{ij}{\in}E w_{ij}\left(k_i+k_j\right)/2\right)^2}{H^{-1}\Sigma e_{ij}{\in}E w_{ij}\left(k_i^2+k_j^2\right)/2 - \left(H^{-1}\Sigma e_{ij}{\in}E w_{ij}\left(k_i+k_j\right)/2\right)^2} \tag{4.12}$$

where H is the total weight of all edges in the network, and w_{ij} is the weight of the edge e_{ij}. Just like r, r^w is also between -1 and 1. In fact,r^w would be reduced to r if the weights of all edges are equal.

In real-world weighted networks with a positive assortativity coefficient, high-degree vertices could be connected to small-degree vertices with less weights, while connected to high-degree vertices with greater weights [43]. A case study of world-wide airport network indicated that high-degree airports could have a great number of flight directly to high-degree airports, while have less number of flight to small-degree airports [44]. The unweighted assortative coefficient and weighted assortative coefficient of the tourism hotspot network in this study were determined to be -0.2869 (r) and 0.1087 (r^w), respectively. On the one hand, the results reveal that the unweighted assortativity coefficient is negative, which means the high-degree attractions tend to be connected to low-degree attractions topologically. On the other hand, the weighted assortativity coefficient is positive, which illustrates the tourist flows among attractions are positively correlated with the degrees of vertices.

TABLE 3. Tourist flows of long-distance hotspot pairs.

Area	a	b	c	d	e	f	g	h	i	j	k
a	\	17			**154**	**164**	**199**	10		**56**	**53**
b		\	38	13	**53**	**205**	**100**	11		26	12
c			\	**50**			10	13		45	**54**
d				\	24	**90**	24	20	34		**76**
e					\		**404**			**84**	**72**
f						\				**108**	
g							\			39	
h								\			
i									\		
j										\	
k											\

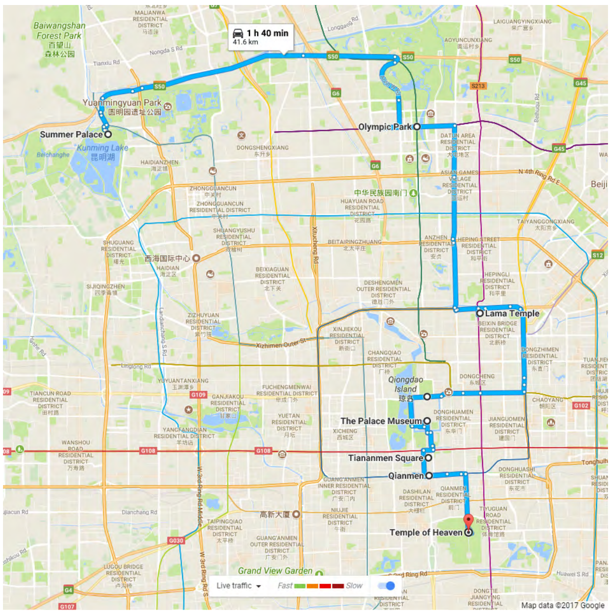


FIGURE 7. Proposed new travel bus route (1).

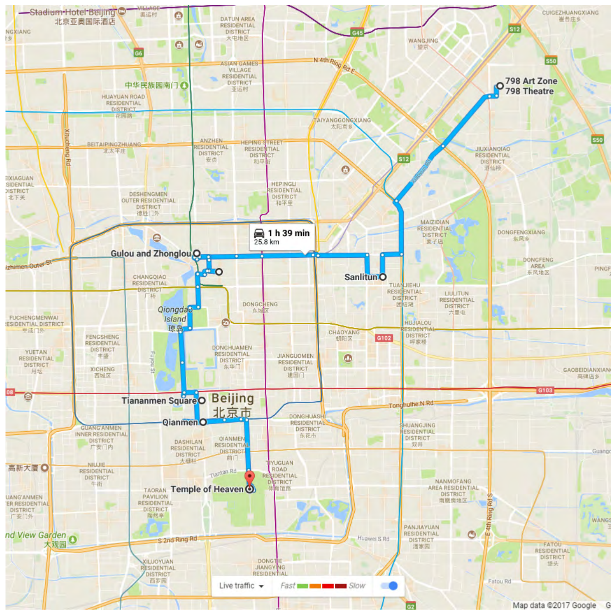


FIGURE 8. Proposed new travel bus route (2).

E. HOTSPOT PAIRS AND TRAVEL BUS ROUTE DESIGN

As mentioned in Section III-C, two hotspots visited consecutively by a user are called a hotspot pair, and the two hotspots in a pair have only one direct connection between them. When a user accesses a hotspot pair, the tourist frequency on the edge between the two hotspots increases by 1.

Table 3 illustrates the triangular matrix of tourist flows of long-distance (i.e., more than 5 kilometers) hotspot pairs. The vertices described in Table 1 are (a) Summer Palace, (b) Beijing Olympic Park, (c) Beijing 798 Art District, (d) Sanlitun Swire, (e) Temple of Heaven, (f) Tiananmen Square–Qianmen area, (g) Palace Museum, (h) Wangfujing, (i) Beijing Capital International Airport, (j) Lama Temple, and (k) Nanluoguxiang–Bell and Drum Tower area. The high-flow values (greater than or equal to 50) are in bold font, and corresponding hotspot pairs should be given priority in the design of new travel bus routes.

Although most tourist attractions in Beijing can be accessed by public transportation, visitors may have a poor travelling experience because of the multiple transit modes, numerous connections, and long transit time. In consideration of the tourist hotspot network characteristics along with the current bus routes in Beijing, we proposed two new travel bus routes to accommodate tourists: (1) the Summer Palace – Beijing Olympic Park – Lama Temple – the Palace Museum – Tiananmen Square – Qian Men – Temple of Heaven; and (2) Beijing 798 Art District – Sanlitun Swire – Nanluoguxiang – Bell and Drum Tower – Tiananmen Square – Qian Men (see Fig. 7 and Fig. 8). On the one hand, these two travel bus routes cover as far as possible the high-flow and long-distance hotspot pairs; on the other hand, these two travel bus routes cross the roads with better traffic conditions and thus complement the existing travel bus routes in Beijing.

V. CONCLUSIONS AND FUTURE WORK

This research introduced the networks science methods to build a spatially-embedded tourism hotspot network and provide insights for the identification of attractions and travel route design. The network vertices were retrieved by the clustering algorithm and the original Flickr dataset. Then, a spatially-embedded tourism hotspot network was built up and complex network analysis was performed. The results indicate that the network possesses several interesting characteristics:

1) The vertex degree, strength, pressure and edge weight are generally subject to power-law distributions, and the network has obvious scale-free characteristics.
2) $C^w > C^u$ indicates that the interconnected triplets are more likely formed by the edges with larger weights. The same happens for $C^w(k)$.
3) The network is efficient at global and local levels no matter it is weighted or not. The network has obvious small-world characteristics. The high value of local efficiency indicates that the network is highly fault-tolerant.
4) The assortativity coefficient r^w of the network is positive, indicating that the tourist flows among attractions are positively correlated with the degrees of vertices.
5) Based on tourist travel patterns and existing transit options, two new travel bus routes had been suggested.

This study constructed a spatially-embedded tourism hotspot network in Beijing using the complex network theory. The results are expected to help visitors to understand the layout of tourist attractions in Beijing and plan reasonable travel routes. The constructed network can also help travel agencies and other organizations design, operate and sell travel products, and help government departments to adjust and add travel bus routes to enhance the tourism industry in Beijing.

Given the recent advances in big data and machine learning, this work could be expanded on in the future in the following two ways:

1) Explore the artificial intelligence based growth model of the tourism hotspot network to provide more suggestions of tourism development.
2) Apply complex network theory to recommend attractions and route prediction.

We believe that researchers will increasingly utilize geotagged social media data and complex network theory in the future to expand the scope of research in tourism field.

REFERENCES

[1] Y. Liu, Z. Sui, C. Kang, and Y. Gao, "Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data," *PLoS One*, vol. 9, no. 1, p. e86026, 2014.
[2] M. H. Salas-Olmedo, B. Moya-Gómez, J. C. García-Palomares, and J. Gutiérrez, "Tourists' digital footprint in cities: Comparing Big Data sources" *Tourism Manag.*, vol. 66, pp. 13–25, Jun. 2018.
[3] L. Yang, L. Wu, Y. Liu, and C. Kang, "Quantifying tourist behavior patterns by travel motifs and geo-tagged photos from flickr," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 12, p. 345, 2017.
[4] J. Wang, H. Mo, F. Wang, and F. Jin, "Exploring the network structure and nodal centrality of China's air transport network: A complex network approach," *J. Transp. Geogr.*, vol. 19, no. 4, pp. 712–721, Jul. 2011.
[5] C. Zhong, S. M. Arisona, X. Huang, M. Batty, and G. Schmitt, "Detecting the dynamics of urban structure through spatial network analysis," *Int. J. Geogr. Inf. Sci.*, vol. 28, no. 11, pp. 2178–2199, 2014.
[6] R. Albert, H. Jeong, and A.-L. Barabasi, "Internet: Diameter of the world-wide Web," *Nature*, vol. 401, pp. 130–131, Sep. 1999.
[7] A.-L. Barabási, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: The topology of the world-wide Web," *Phys. A, Statist. Mech. Appl.*, vol. 281, no. 1, pp. 69–77, 2000.
[8] G. W. Flake, S. Lawrence, and C. L. Giles, "Efficient identification of Web communities," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2000, pp. 150–160.
[9] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Phys. Rep.*, vol. 424, nos. 4–5, pp. 175–308, 2006.
[10] X. F. Wang and G. Chen, "Complex networks: Small-world, scale-free and beyond," *IEEE Circuits Syst. Mag.*, vol. 116, no. 1, pp. 6–20, Sep. 2003.
[11] D. J. Watts, P. S. Dodds, and M. E. J. Newman, "Identity and search in social networks," *Science*, vol. 296, no. 5571, pp. 1302–1305, 2002.
[12] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
[13] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proc. Nat. Acad. Sci. USA.*, vol. 101, no. 11, pp. 3747–3752, 2004.
[14] S. Williams, *Tourism Geography*. London, U.K.: Routledge. 1998.
[15] J. Fogel and E. Nehmad, "Internet social network communities: Risk taking, trust, and privacy concerns," *Comput. Hum. Behav.*, vol. 25, no. 1, pp. 153–160, 2009.
[16] D. Centola, "The spread of behavior in an online social network experiment," *Science*, vol. 329, no. 5996, pp. 1194–1197, 2010.
[17] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the flickr social network," in *Proc. 1st Work. Online Soc. Netw. (WOSP)*, vol. 8. 2008, pp. 25–30.
[18] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas. (IMC)*, vol. 7. 2007, pp. 29–42.
[19] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Proc. 12th Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 611–617.
[20] J. I. L. Miguéns and J. F. F. Mendes, "Travel and tourism: Into a complex network," *Phys. A, Stat. Mech. Appl.*, vol. 387, no. 12, pp. 2963–2971, 2008.
[21] R. Baggio, N. Scott, and C. Cooper, "Network science: A review focused on tourism," *Ann. Tourism Res.*, vol. 37, no. 3, pp. 802–827, 2010.
[22] R. Baggio and C. Cooper, "Knowledge transfer in a tourism destination: The effects of a network structure," *Service Ind. J.*, vol. 30, no. 10, pp. 1757–1771, 2010.
[23] L. Hollenstein and R. Purves, "Exploring place through user-generated content: Using Flickr tags to describe city cores," *J. Spatial Inf. Sci.*, vol. 2010, no. 1, pp. 21–48, 2010.
[24] M. Wall and T. Kirdnark, "Online maps and minorities: Geotagging Thailand's Muslims," *New Media Soc.*, vol. 14, no. 4, pp. 701–716, 2012.
[25] J. W. Crampton *et al.*, "Beyond the geotag: Situating 'big data'and leveraging the potential of the geoweb," *Cartogr. Geogr. Inf. Sci.*, vol. 40, no. 2, pp. 130–139, 2013.
[26] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proc. 18th Int. Conf. World Wide Web (WWW)*, vol. 9. 2009, pp. 761–770.
[27] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "PPHOPCM: Privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing," *IEEE Trans. Big Data*, to be published. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7920374
[28] L. Yi, X. Deng, M. Wang, D. Ding, and Y. Wang, "Localized confident information coverage hole detection in Internet of things for radioactive pollution monitoring," *IEEE Access*, vol. 5, pp. 18665–18674, 2017.
[29] Q. Zhang and Z. Chen, "A distributed weighted possibilistic C-means algorithm for clustering incomplete big sensor data," *Int. J. Distrib. Sensor Netw.*, vol. 10, no. 5, pp. 430814-1–430814-8, 2014.
[30] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.

[31] X. Peng and Z. Huang, "A novel popular tourist attraction discovering approach based on geo-tagged social media big data," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 7, p. 216, 2017.
[32] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, "Classes of small-world networks," *Proc. Nat. Acad. Sci. USA*, vol. 97, no. 21, pp. 11149–11152, 2000.
[33] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Econ. Geogr.*, vol. 46, pp. 234–240, Jun. 1970.
[34] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.
[35] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Netw.*, vol. 31, no. 2, pp. 155–163, May 2009.
[36] V. Latora and M. Marchiori, "Efficient behavior of small-world networks," *Phys. Rev. Lett.*, vol. 87, no. 19, p. 198701, 2001.
[37] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numer. Math.*, vol. 1, pp. 269–271, Dec. 1959.
[38] S. Yang and D. Knoke, "Optimal connections: Strength and distance in valued graphs," *Social Netw.*, vol. 23, no. 4, pp. 285–295, 2001.
[39] T. Opsahl, F. Agneessens, and J. Skvoretzc, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Netw.*, vol. 32, no. 3, pp. 245–251, 2010.
[40] M. E. J. Newman, "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 1, p. 016132, 2001.
[41] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, 2001.
[42] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett.*, vol. 89, no. 20, p. 208701, Oct. 2002.
[43] C. C. Leung and H. F. Chau, "Weighted assortative and disassortative networks model," *Phys. A, Statist. Mech. Appl.*, vol. 378, no. 2, pp. 591–602, 2007.
[44] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, "Module identification in bipartite and directed networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 76, no. 3, p. 036102, 2007.

XINYU WU received the B.Sc. degree in geographic information science from Peking University, Beijing, China, in 2017, where he is currently pursuing the M.Sc. degree in geographic information science.

His research interests include data mining and complex network and its geospatial applications.

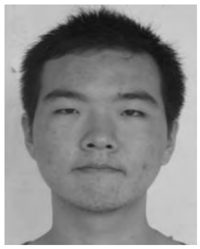
ZHOU HUANG received the B.Sc. degree in geographic information science and the Ph.D. degree in cartography and geographic information science from Peking University, China, in 2004 and 2009, respectively. He is currently an Associate Professor of geographic information science with the Institute of Remote Sensing and Geographical Information Systems, Peking University.

He has authored over 50 academic papers in international journals or conferences. His main current research interests include big geo-data, high-performance geocomputation, distributed geographic information processing, spatial data mining, and spatial database.

Dr. Huang was selected for the Youth Talent Innovation Plan in Remote Sensing Science and Technology through the Ministry of Science and Technology of China in 2015.

XIA PENG received the B.S. degree in geographical information system from the China University of Geosciences, Wuhan, China, in 2004, the M.S. degree in cartography and geographical information system from Peking University, Beijing, China, in 2007, and the Ph.D. degree in urban and rural planning from Tsinghua University, Beijing, China, in 2013.

She is currently an Associate Professor with the Tourism College, Beijing Union University, Beijing. Her major research interests include data mining, geographic information science, and tourism decision support system.



YIRAN CHEN received the B.Sc. degree in geographic information science from Peking University, Beijing, China, in 2016, where he is currently pursuing the M.Sc. degree in geographic information science.

His research interests include high-performance distributed geo-computing and big data mining.

YU LIU is currently a Professor of geographic information science with the Institute of Remote Sensing and Geographical Information Systems, Peking University. His research interests cover several theoretical aspects in geographical information science (e.g., spatial cognition and spatial reasoning) and human mobility patterns.

He has authored over 50 refereed papers in international journals. He has conducted about 10 research projects as PI or Co-PI granted by the National Science Foundation of China and the Ministry of Science and Technology of China.

He is currently an Associate Editor of *Computers, Environment and Urban Systems* and an Editorial Board Member of the *Journal of Spatial Information Sciences*.