IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 18, 2022, accepted February 1, 2022, date of publication February 7, 2022, date of current version February 15, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3149640

Extracting Network Patterns of Tourist Flows in an Urban Agglomeration Through Digital Footprints: The Case of Greater Bay Area

JIARUI MOU
College of Resources, Shandong University of Science and Technology, Tai'an 271019, China
e-mail: moujiarui@sdust.edu.cn

ABSTRACT Network patterns of tourist flows can reveal differences in tourism resources among destinations from the perspective of network science, providing valuable suggestions for tourism managers and policymakers to promote the balanced and sustainable development of tourism. This paper focuses on urban agglomerations, a highly developed spatial form of integrated cities, and proposes a research framework to extract the network patterns of tourist flows through digital footprints. Based on an illustrative case study using geo-located travel blog data from Qunar.com, we built a tourist flow network for the Guangdong-Hong Kong-Macao Greater Bay Area (GBA), China. The analysis shows: (1) GBA's tourist flow network is obviously heterogeneous, showing a pattern of ''four cores and three poles''; (2) the strong ''administrative barrier effect,'' revealed by community detection within the network, is the main obstacle to integrating regional tourism; (3) strengthening the infrastructure of tourism mediation cities such as Guangzhou, Zhuhai, and Shenzhen, so as to avoid the ''structural hole'' caused by the tourist flows between cities, is an urgent issue that the GBA government needs to address. To summarize, the research framework can provide a theoretical basis and concrete suggestions for the planning and management of the tourism industry in urban agglomerations.

INDEX TERMS Tourist flow, digital footprint, network science, urban agglomeration, Guangdong-Hong Kong-Macao Greater Bay Area.

I. INTRODUCTION

Urban agglomerations are highly developed spatial forms of integrated cities, and are the spatial carrier of regional economic and industrial development [1]. Tourism, as an important driving factor for economic growth in cities, is an external manifestation of the integrated development process of urban agglomerations [2], [3]. Evaluating tourism development in urban agglomerations not only provides an important reference for the spatial restructuring of industry, but also has important significance for promoting industrial collaboration and resource exchange within and outside the urban agglomeration, thereby contributing to building an international trade platform [4].

Patterns of human mobility can provide insights into social phenomena in cities [5], [6]. Tourist flow, generated by the movement of tourists between destinations, is therefore an important metric to evaluate tourism development [7]. Current tourist flow studies mainly focus on theoretical frameworks [8], influencing factors [9], flow patterns [10], and prediction models [11], among which analysis of the spatial characteristics of tourist flow has been a research hotspot in recent years. However, survey-based data sources used in traditional tourism research often cannot accurately and comprehensively reflect the movement of tourists, which creates difficulties when researching spatial patterns from a quantitative perspective. In recent years, scholars have begun to introduce the concept of ''digital footprints'' [12], which facilitate detailed modeling of tourist movement.

Tourist flow can depict the spatial distribution of tourists from a dynamic perspective. Scholars usually use Markov chains [13], spatial clustering [14], kernel density estimation [15], exploratory spatial data analysis [16], and other methods to detect the spatial patterns of tourist flows. In recent years, the introduction of network science has provided a new way to analyze tourist flow [17]. Through the construction of tourist flow networks, tourist flows and destinations can be integrated into an efficient spatial

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Mei.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

system, and the role relationship among tourist destinations can be analyzed mathematically and conceptually from a sociological perspective.

In view of this approach, this paper proposes a research framework for extracting the network patterns of tourist flows in urban agglomerations through digital footprints. Taking the Guangdong-Hong Kong-Macao Greater Bay Area (GBA) in China as a case study, tourists' digital footprints (collected from travel blogs on Qunar.com) are used to verify the proposed research framework. The structure of this paper is as follows: the second section reviews the most relevant previous literature on tourists' digital footprints and tourism network research. The third section describes the research design, including the research framework and case study data. In the fourth and fifth section, the analysis results of the case study are introduced and discussed. Finally, the last section makes a summary conclusion of the practical implications of the research framework.

II. LITERATURE REVIEW

A. TOURISTS' DIGITAL FOOTPRINTS

Tourists' digital footprints are the electronic traces left by tourists during their trips [12]; they include many data sources, such as GPS trajectory data, check-in data, mobile phone signaling data, travel blogs, geotagged photos, etc. As more and more tourists are willing to share their travel experiences on social platforms, the ability of digital footprints to depict tourists' behavior has increased [18]. At present, many scholars have utilized different digital footprints to study tourist behavior in cities. For instance, Roura *et al.* [19], Li *et al.* [20], and Jing *et al.* [21] have utilized, respectively, travel blogs, GPS trajectory data, and geotagged photos.

Many scholars have proposed research frameworks based on tourists' digital footprints. For example, Walden-Schreine *et al.* [22], Mou *et al.* [23], and Martí *et al.* [24] have proposed a research framework for, respectively, exploring the spatial relationship between tourists and infrastructure in nature reserves, detecting the spatial patterns of tourist flows in the cities, and identifying urban tourism hotspots. However, these proposed research frameworks are mostly limited to a single city or scenic area, while the agglomeration of tourist destinations is rarely involved. The mechanism of tourism in urban agglomerations is more complex, making it worth exploring to obtain more valuable implications for tourism management. Therefore, to meet this need, in this study we proposed a research framework for large-scale urban agglomerations based on digital footprint data.

B. NETWORK SCIENCE IN TOURIST FLOW RESEARCH

Network science, derived from graph theory, can describe the structure of relations between given tourism entities in the form of nodes and links [25], providing a new perspective for tourist flow research [26]. The relevant research can be divided into two types according to its analysis approach. One explores the network characteristics among tourist destinations at the level of administrative units such as cities or countries. For instance, Qin *et al.* [27], Xu *et al.* [28], and Zeng [29] have studied the network structure of inbound tourist flow of cities in South Korea, China, and Japan, respectively. Provenzano *et al.* [30] and Chung *et al.* [31] both discussed the dynamic characteristics of tourist flows among European countries. Another common research stream is to explore the network characteristics of tourist flows in inner cities at the attraction level. For instance, Liu *et al.* [32] and Jin *et al.* [33] have built tourist flow networks at the attraction level and explored, respectively, the networks' underlying mechanisms and temporal heterogeneity.

At the theoretical level, many scholars have introduced social network theory in their research to explore the sociological relationship between tourist destinations. For example, Liu *et al.* [7] utilized centrality analysis and automorphic equivalence analysis to reveal the ''roles'' played by destinations in the Tourism Region of South Anhui; Mou *et al.* [23] explored the ''power distribution'' of attractions in Qingdao by detecting structural holes in the network. In addition, as the number of network nodes increases, the complexity of tourist flow networks has also attracted the attention of scholars. For example, Miguéns *et al.* [34], Wu *et al.* [35], and Gao *et al.* [36] have studied, respectively, the scale-free characteristic, small-world effect, and community structure of tourist flow networks.

In view of this, the data source and theoretical basis of the research framework proposed in this paper are digital footprints and network science, respectively. In addition, the research framework divides the tourist flow network into two levels: attractions and cities. With the help of various network indicators, the network patterns of tourist flow in urban agglomerations can be extracted in detail.

III. MATERIALS AND METHODS

A. RESEARCH FRAMEWORK

We propose a novel research framework, as shown in Figure 1. In the framework, digital footprints are used to build the tourist flow network at the level of attractions and cities. By dividing the tourist flow network into two levels, a multi-scale analysis can be performed, which has been proved to provide a more systematic and coherent knowledge of tourism [37], [38]. In the network, attractions or cities are abstracted as nodes, and the tourist flows between nodes are abstracted as weighted edges. The weights are based on the number of tourist movements between two nodes in the network. On this basis, four types of network indicators are selected to analyze the network characteristics of tourist flows.

1) NODE DEGREE, WEIGHTED DEGREE, AND FLOW BETWEENNESS CENTRALITY

The three indicators (node degree, weighted degree and flow betweenness centrality) reflect importance of particular nodes to tourist flows. Node degree is commonly utilized to plot the degree distribution to analyze the heterogeneity of a complex network with a large number of nodes. Therefore,

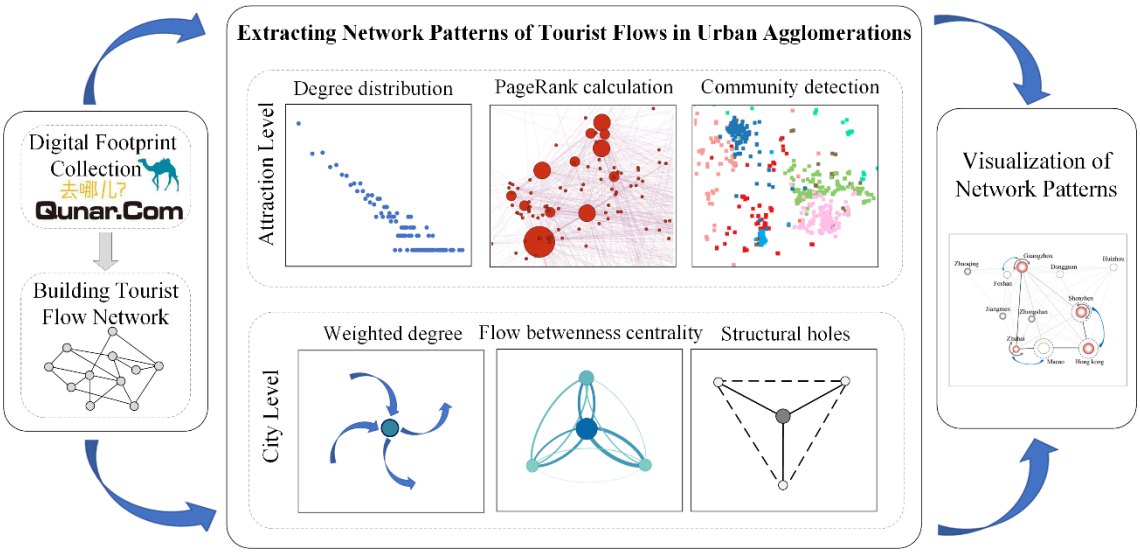

FIGURE 1. Research framework for extracting network patterns of tourist flows in urban agglomerations.

this indicator is applied to the attraction level of the research framework. Weighted degree and flow betweenness centrality are applied to the city level of the research framework since these two indicators are more focused on a fine-grained analysis of each node's importance in the network. The three indicators are described below.

Node degree and weighted degree reflect the link strength of the node in the network. The main difference between the two indicators is that weighted degree takes into account the influence of the weight of the edge (i.e., the volume of tourist flow) on the link strength of the node. The two indicators are defined as follows:

$$ND_i = \sum_{j \in N_i} l_{ij} \tag{1}$$

$$WD_i = \sum_{j \in N_i} w_{ij} \tag{2}$$

where ND_i is the node degree of node v_i and WD_i is the weighted degree of node vi, l_{ij} is the number of directed edges between node v_i and node v_j, w_{ij} is the sum of the weights of directed edges between node v_i and node v_j (which can be regarded as the volume of the inflows and outflows of the target destination), and N_i is a collection of adjacent nodes of node v_i.

Flow betweenness centrality measures the degree of control of the target node over other nodes. Our research framework utilized the form of flow betweenness centrality in directed-weighted networks [39]:

$$BC_i = \sum_{j}^{n} \sum_{k}^{n} \frac{m_{jk}(i)}{m_{jk}}, \quad j < k,\ j \neq k \neq i \tag{3}$$

where BC_i is the flow betweenness centrality of node v_i, m_{jk} is the maximum flow from node v_j to node v_k (i.e., the weight value of the path with the largest weight from node v_j to node v_k), node $m_{jk}(i)$ is the maximum flow from node vj to node vk through node vi, and n is the total number of nodes.

2) STRUCTURAL HOLES MEASUREMENT

Structural holes are an indicator of connection fractures between network nodes. The calculation of structural holes enables the identification of bottleneck problems in regional tourist flows. The nodes with structural hole advantages generally have strong regional competitive advantages [40]. Therefore, our research framework utilizes structural holes at the city level to evaluate the competitive advantage of each city as a tourist destination. Effective size and constraint are utilized to measure the positive and negative aspects of structural holes, as described below.

Effective size measures the non-redundant part of the target node connected to all other nodes. The higher the effective size, the more obvious the competitive advantage of the target node. It is calculated as follows:

$$ES_i = \sum_{j}^{n} \left(1 - \sum_{q}^{n} p_{iq} m_{jq}\right), \quad q \neq i, j, \quad \text{where}$$

$$p_{iq} = \frac{(z_{iq} + z_{qi})}{\sum_{j}^{n} (z_{ij} + z_{ji})}, \quad i \neq j$$

$$m_{jq} = \frac{(z_{jq} + z_{qj})}{\max (z_{jk} + z_{kj})}, \quad j \neq k \tag{4}$$

where z_{iq} is the number of connections from node v_i to node v_q, p_{iq} is the proportional relationship between the tourist node v_i and node v_q (i.e., the number of connections between node v_i and node v_q divided by the number of all the connections of nodes v_i), m_{jq} is the marginal strength between nodes v_j and node v_q (i.e., the number of connections between nodes v_j and node v_q divided by the maximum number of

connections between node v_j and other nodes), and n is the number of nodes in the network.

Constraint reflects the degree of the target node's dependence on other nodes. The smaller the constraint, the higher the status and therefore the competitive advantage of the target node in the region. On the contrary, the greater the constraint, the greater the impact of other nodes on the target node, signifying lower competitiveness. It is calculated as follows:

$$CT_i = \sum_{j}^{n} \left(p_{ij} + \sum_{q}^{n} p_{iq} p_{qj} \right)^2, \quad q \neq i, j \tag{5}$$

where p_{ij}, p_{iq}, and p_{qj} are the proportional relationships between nodes v_i and v_j, nodes v_i and v_q, and nodes v_q and v_j, respectively, and n is the number of nodes in the network.

3) PAGERANK

PageRank was originally designed to measure the importance of specific webpages relative to other webpages in a search engine. Later, it was introduced into complex network analysis to measure the importance of nodes [41]. Compared with centrality indicators such as node degree and weighted degree, PageRank can achieve a more comprehensive consideration of a target node's network status. Our research framework utilizes this indicator at the attraction level to observe the spatial distribution of high-importance attractions. The indicator is defined as follows:

$$PR_i = \frac{1-\alpha}{n} + \alpha \sum_{j \in N_i} \frac{w\left(v_j, v_i\right)}{\sum_{k \in N_j} w\left(v_j, v_k\right)} \cdot PR_j \tag{6}$$

where PR_i is the PageRank value of node v_i, N_i and N_j are the collections of adjacent nodes of nodes v_i and v_j, respectively, $w(v_j, v_k)$ is the weight value of the directed edge from node v_j to v_k, and α ($0 < \alpha < 1$) is the damping coefficient. The larger α is, the easier it is to distinguish nodes' PageRank values. Generally, α takes the empirical value 0.85. Through several iterations, the PageRank value of each node will eventually stabilize, and the sum of the PageRank values of all nodes will approach 1.

4) COMMUNITY DETECTION

Communities are dense subnets in the whole complex network. Therefore, the community detection algorithm is applied to the attraction level of our research framework. The structural strength of communities can be measured by their modularity, which is defined as follows:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i c_j) \tag{7}$$

where m is the number of edges in the network, A_{ij} is the adjacency matrix representing the weight of edges between nodes v_i and v_j, k_i is the weight of all edges connected with node v_i, and c_i is the community number of node v_i. If nodes v_i and v_j are in the same community, the return value of $\delta(c_i c_j) = 1$, otherwise $\delta(c_i c_j) = 0$.

Communities can be detected by finding the approximate optimal partition with larger modularity among all possible community partitions. Traag *et al.* [42] recently proposed the Leiden algorithm, which is considered to be the best community detection method based on modularity optimization in operating performance and effectiveness. Therefore, our research framework utilized this method to effectively detect community structures in the network.

B. STUDY AREA

We chose GBA, a world-class urban agglomeration in China, as the study area for the validation of our framework. GBA, also known as the "upgraded version of the Pearl River Delta Urban Agglomeration," is a large urban agglomeration composed of 11 cities (Guangzhou, Zhuhai, Shenzhen, Foshan, Dongguan, Huizhou, Zhongshan, Jiangmen, Zhaoqing, Hong Kong and Macao), with a total land area of about 56,000 square kilometers. By the end of 2017, the permanent residential population of GBA had reached nearly 70 million, and its GDP had exceeded 10 trillion yuan (US$ 1.5 trillion), making it one of the most economically dynamic bay areas in the world. The aim of Chinese government planning is that GBA should not only be built as a world-class urban agglomeration with high-quality development, but also become a high-quality region suitable for living, business, and tourism. GBA has been identified as the fastest growing area in China's tourism industry, and received more than 400 million visitors from home and abroad in 2016 [43].

According to the statistical data of 2017, Guangzhou, Shenzhen, Hong Kong and Macau attracted 204 million, 130 million, 58 million, and 33 million visitors respectively, and their tourism revenue accounted for 73 percent of the whole GBA [44]. Therefore, these four cities are generally regarded as the leading cities of the tourism in GBA. There are also numerous well-known attractions in GBA, such as the Window of the World in Shenzhen, the Canton Tower in Guangzhou, the Victoria Harbor in Hong Kong, and the Ruins of St. Paul's in Macao.

C. DIGITAL FOOTPRINT DATA

Since different data sources for digital footprints portray different aspects of tourists' behavior, their selection has proved controversial. Tourists' location data as recorded by sensor-based data sources (such as geotagged photos, check-in data, GPS trajectory data, etc.) are affected by regulatory issues (such as ethics, areas where photography is prohibited, signal shielding of position sensors, etc.) [45] and often report redundant locational information. Geo-located travel blog data, directly edited by tourists, is therefore chosen as the digital footprint data source within our framework.

The two most popular travel websites in China are Ctrip (www.ctrip.com/) and Qunar (https://www.qunar.com/) [46]. Despite the fact that Ctrip has a slightly larger user base than Qunar, the website's travel blog data lacks standardized

geotags, making it unable to suit our geospatial research demands. Therefore, we chose Qunar as the case's primary data source. Qunar.com provides a smart travel blog editing platform; when users write travel blogs on the website, they can insert point of interest (POI) tags for specific content. When the travel blog is published, these POI tags are recorded in the source code of the blog's webpage and can be utilized to visualize the user's travel route. Kaufmann *et al.* [47] and Zheng *et al.* [48] called this type of data ''geo-located travel blog data.''

We collected 8786 travel blogs from Qunar.com dating from 2010 to 2019, which were voluntarily shared by tourists, as the initial data. The collected data contained user ID, blog ID, departure date, and the sequence of POIs visited (Table 1). The travel blog data often have information errors and other problems, so we designed rules to clean the data, as shown in Table 2. After data cleaning, a total of 26843 visits to 1419 attractions in 3765 blogs were retained. The spatial distribution of the data is shown in Figure 2.

IV. RESULTS

A. ATTRACTION LEVEL

1) DEGREE DISTRIBUTION

We built a GBA tourist flow network at the attraction level. In the tourist flow network, attractions were abstracted as nodes and tourist flows between attractions were abstracted as weighted edges. Figure 3 shows the distribution of the node degrees in the network (based on Equation 1) and its power fitting curve; it can be seen that the node degree distribution of the tourist flow network at the attraction level basically meets the power law distribution, and the R^2 value (i.e., the degree of power fitting) reaches 0.886, which indicates that the network has obvious scale-free characteristics. Tourist flow networks with scale-free characteristics are usually heterogeneous [49], that is, there are fewer nodes with larger degree values and more nodes with smaller degree values, indicating that most of the tourist flows in GBA are mainly concentrated in a few attractions.

2) NODE IMPORTANCE

To comprehensively evaluate the status of network nodes, the PageRank algorithm (based on Equation 6) was utilized to calculate the importance of nodes in the network at the attraction level. The spatial distribution of the calculation result of each node (attraction) is shown in Figure 4.

As shown in Figure 4, most of the important nodes in the GBA tourist flow network are located in Guangzhou, Hong Kong, Macao, and Shenzhen. These four cities are therefore the most popular tourist destinations in GBA. In addition, the attractions in the cities near the three most popular cities (Guangzhou, Hongkong and Macao), such as those in eastern Foshan, southern Shenzhen, and eastern Zhuhai, also have higher PageRank values, showing the effect of geographical proximity. Therefore, the results highlight the ''four cores and three poles'' distribution of the tourism industry in GBA: There are four core cities of Hong Kong, Macao, Guangzhou and Shenzhen, and three growth poles of Hong Kong-Shenzhen, Macao-Zhuhai, and Guangzhou-Foshan.

3) COMMUNITY DETECTION

The Leiden algorithm was utilized to detect communities in the tourist flow network at the attraction level; a total of 32 communities were detected. The spatial distribution of the results is shown in Figure 5. Since the number of nodes in the 10th to 32nd communities is less than 1% of the total, and most of these communities are located on the geographic edge of GBA, they are unified as one independent community in the map.

As shown in Figure 5, although the spatial distribution of attractions in GBA is relatively dispersed, the communities of attractions have clear geographical boundaries. The communities of attractions in Hong Kong, Macao, and Shenzhen are close to the actual administrative divisions. The community boundaries between Zhuhai and Macao, Hong Kong and Shenzhen, and Guangzhou and Foshan are also consistent with the actual administrative boundaries. In particular, the attractions of Wanshan archipelago, located to the east of Zhuhai, are not included in the communities of Hong Kong or Macao due to geographical proximity, but are in the same communities as those of Zhuhai inland. Therefore, the network structure of tourist flows in GBA has been affected by real-world administrative divisions, showing a strong ''administrative barrier effect.''

B. CITY LEVEL

1) WEIGHTED DEGREE AND FLOW BETWEENNESS CENTRALITY

We also built a tourist flow network at the city level. In this network, attractions located in the same city are merged into one node. Since the weighted degree (based on Equation 2) of a city in the tourist flow network is the volume of its bi-directional tourist flow, Figure 6 was drawn to show the flow relation between cities in GBA.

As shown in Figure 6, the tourist flows between Hong Kong and Macao, Zhuhai and Macao, Guangzhou and Zhuhai, and Hong Kong and Shenzhen have the largest volume. In terms of weighted degree, Hong Kong, Macao, Zhuhai, Guangzhou, and Shenzhen have the highest values. Among these five cities, Zhuhai, Guangzhou, and Shenzhen are the cities with the highest weighted degree values in the part of GBA that lies within mainland China. Considering that transportation between cities in mainland China is more convenient, it seems that Guangzhou, Shenzhen, and Zhuhai play a ''mediation role'' in the GBA network pattern. This interpretation is further confirmed by the calculation of flow betweenness centrality (based on Equation 3) as shown in Figure 7. It can be seen from Figure 7 that Zhuhai, Guangzhou, and Shenzhen are indeed the three mediation nodes of the tourist flow network.

2) MEASUREMENT OF STRUCTURAL HOLES

The two indicators of effective size (based on Equation 4) and constraint (based on Equation 5) were utilized to measure

TABLE 1. Sample records of geo-located travel blogs from Qunar.

User Id	Blog Id	Departure date	The sequence of the POIs visited
171***996@qunar	50***07	2012/12/29	5475514;5475515;706589…
148***892@qunar	66***27	2015/5/25	8304997;5475515
151***933@qunar	68***42	2016/11/12	702128;706245;710128…
150***473@qunar	66***39	2017/1/31	7469417;714012;9127292…
255***414@qunar	75***90	2019/7/30	714566;701866;706555…

TABLE 2. Overview of the data cleaning procedure for geo-located travel blog data.

Cleaning procedure	Remaining amount of data		
	Blog	Attraction	Visit
Initial data	8786		
Remove blogs without POI tags	5622	5105	40263
Remove the records of attractions outside GBA	5158	1511	29377
Remove continuous duplicate records of attractions	5158	1511	28236
Remove blogs that contain only one attraction	3765	1419	26843

FIGURE 2. The spatial distribution of the travel trajectories of tourists in the Guangdong-Hong Kong-Macao greater bay area.

structural holes in the GBA tourist flow network. The results for each city are shown in Table 3, and a visualization of the tourist flow network, according to the effective size and constraint values, is shown in Figure 8. The size of the nodes in the figure represents the level of the structural hole indicators.

When comparing the results of structural hole indicators in Figure 8(a) and (b), it becomes evident that the distribution of nodes' effective size and constraint in the tourist flow network possesses certain regular characteristics. The four city nodes of Guangzhou, Zhuhai, Hong Kong, and Shenzhen have higher effective size values but lower constraint values, indicating that these cities have obvious competitive advantages. Zhaoqing has the lowest effective size value and the highest constraint value. Therefore, Zhaoqing is the city with the least competitive advantage, which is consistent with Zhaoqing's marginal status in GBA. In addition to Zhaoqing, Zhongshan and Jiangmen also have competitive disadvantages due to their lower effective size value and higher constraint value. This indicates that the tourist flows to and from these

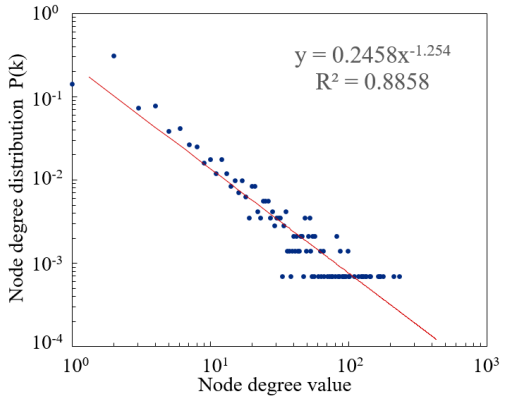

FIGURE 3. Node degree distribution and power fitting curve.

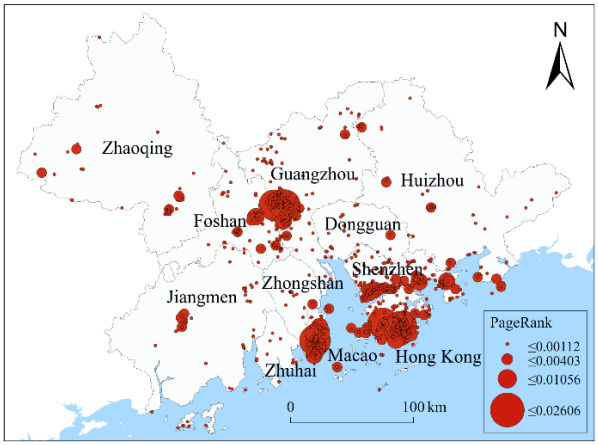

FIGURE 4. Spatial distribution of PageRank calculation results of attractions in GBA.

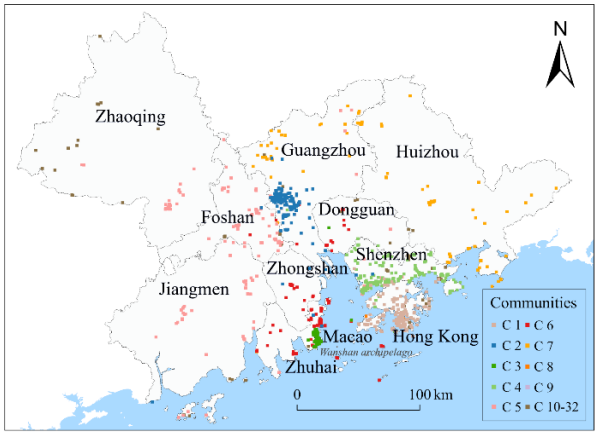

FIGURE 5. Spatial distribution of communities in the GBA tourist flow network.

cities have been limited; they are disadvantaged in the competition for tourists. In addition, the two indicators for Macao are at a medium level, indicating that Macao almost does not participate in competition for tourism; i.e., it neither significantly impacts, nor is impacted by, other cities.

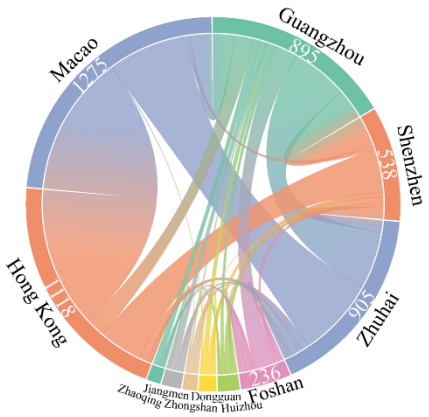

FIGURE 6. Distribution of tourist flows between cities in GBA (cities' weighted degree values are marked with white numbers).

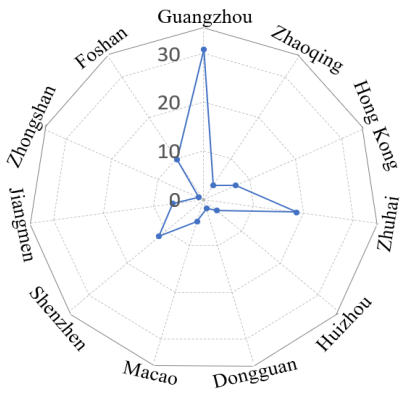

FIGURE 7. Calculation of cities' flow betweenness centrality in GBA.

TABLE 3. Structural hole indicators of cities in GBA.

City name	Effective size	Constraint
Guangzhou	7.496	0.484
Hong Kong	6.798	0.672
Zhuhai	6.302	0.636
Shenzhen	5.425	0.662
Huizhou	5.168	0.700
Foshan	4.891	0.690
Dongguan	4.553	0.737
Jiangmen	3.909	0.875
Zhongshan	3.482	0.974
Macao	3.300	0.617
Zhaoqing	1.701	1.116

3) VISUALIZATION OF NETWORK PATTERNS

Based on the above analysis, the network patterns of tourist flows in GBA can be summarized according to the proposed research framework (Figure 1). A schematic diagram of the visualization of network patterns is presented in Figure 9.

As shown in Figure 9, the tourist flows in GBA form a network pattern showing fierce competition and

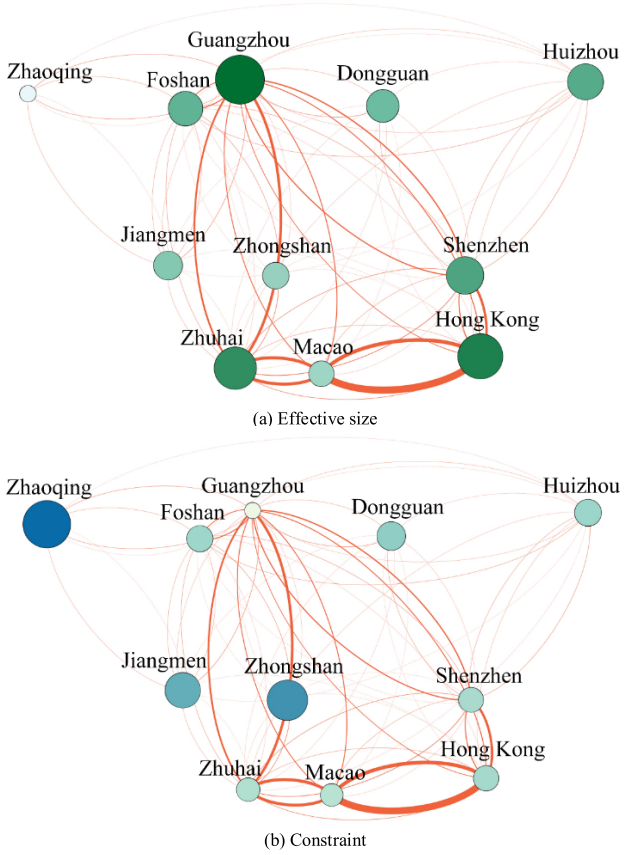

FIGURE 8. Network structure of tourist flows in GBA based on structural hole indicators.

power imbalances between cities. The main characteristics of the network patterns can be summarized as follows:

1) Guangzhou, Shenzhen, Hong Kong, and Macao are the core nodes of the network because of their high weighted degrees and their sub-nodes' (attractions) high PageRank values.
2) Due to the high volume of tourist flows between them, and the proximity of their important attractions, there are three prominent associations between Hong Kong and Shenzhen, Macao and Zhuhai, and Guangzhou and Foshan.
3) Guangzhou, Shenzhen, and Zhuhai have high weighted degree and flow betweenness centrality values and therefore function as mediation nodes of the network.
4) Guangzhou, Shenzhen, Zhuhai, and Hong Kong are the nodes with structural hole advantages in the network, while Zhaoqing, Jiangmen, and Zhongshan have a structural hole disadvantage. City nodes with structural hole advantages will take the lead in the competition for tourist flows.
5) The tourist flow network is affected by ''administrative barriers''; this is most obvious in Hong Kong, Macao, and Shenzhen.

V. DISCUSSION

Building GBA into a world-class tourist destination is an important strategic goal for the Chinese government. The analysis results verified that the proposed research framework was able to effectively extract the network patterns of tourist flows in an urban agglomeration at the attraction and city levels. The framework can provide valuable implications for tourism managers and policymakers to promote the development and integration of the tourist industry, as described below.

First, the GBA tourist flow network has been proved to be heterogeneous at the attraction level (Figure 3). GBA is a large-scale bay area, with diverse tourism resources and environments among member cities. Therefore, it seems difficult to promote balanced tourism development in GBA.

Three city groups (also the development poles), Guangzhou and Foshan, Hong Kong and Shenzhen, and Macao and Zhuhai, have many important nodes in the tourist flow network (Figure 4). Efforts to strengthen the tourism connections between these three city groups and their surrounding cities, such as by promoting tourism industry cooperation and combined marketing, will drive the development of tourism in the entire region to a certain extent.

Second, ''administrative barriers'' are the main obstacle to promoting the integration of tourism in GBA. For example, the spatial distribution of attraction communities located at the three development poles of Guangzhou-Foshan, Hong Kong-Shenzhen and Macao-Zhuhai are obviously influenced by administrative boundaries (Figure 5). GBA has a unique ''9+2'' structure, which covers nine mainland cities and two special administrative regions (Hong Kong and Macao). This structure appears to make it difficult to break the ''administrative barriers'' of the tourism industry in this region through policy measures. In order to solve this problem, the local tourism management departments need to take the ''Outline Development Plan for the Guangdong-Hong Kong-Macao Greater Bay Area'' [50], promulgated by the Chinese government, as an important reference in promoting reforming innovation and industrial cross-regional cooperation. One such effort is the construction of ''Guangfo city'' by integrating Guangdong and Foshan into one connected city. Furthermore, cities should cooperate to formulate a development plan for tourism resource integration, so as to weaken the impact of the ''administrative barriers'' on GBA tourism. Additionally, GBA's transport network needs to be optimized. At present, the Guangzhou-Hong Kong High-Speed Railway and the Hong Kong-Zhuhai-Macao Bridge have been built to enhance transport links between cities in GBA. However, due to various policy issues, especially complicated transport procedures, the benefits of these two transport infrastructures have not been maximized. In the future, it will be necessary to enhance the convenience of transportation and create high-quality tourism routes, to promote the coordinated development of tourism in GBA.

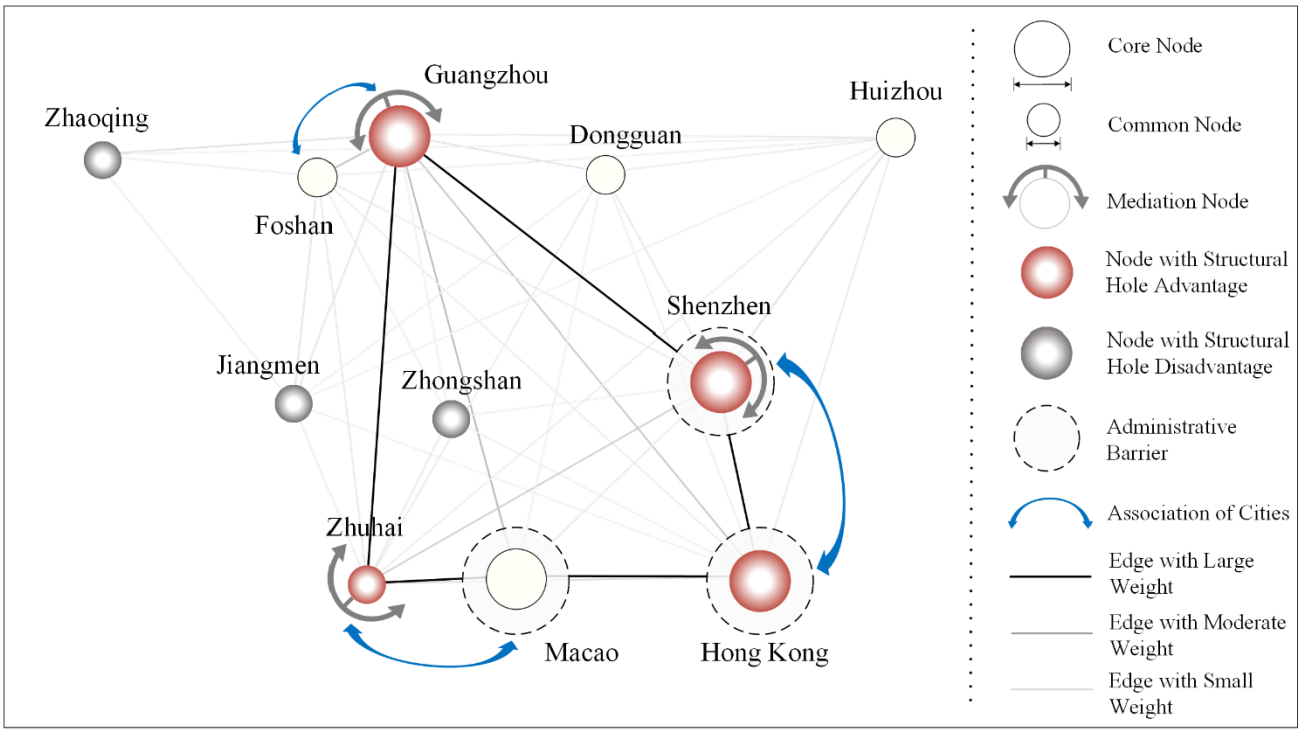

FIGURE 9. Network patterns of tourist flows in Guangdong-Hong Kong-Macao greater bay area.

Finally, more attention should be paid to the ''interaction mechanism'' of tourist flows among cities in GBA. Guangzhou, Zhuhai, and Shenzhen have high flow betweenness centrality values in the tourist flow network. These three cities are, thus, identified as the important mediation cities of tourist flows, which is consistent with their ''hub'' role in the GBA transportation system. Strengthening the construction of expressways, ports, airports, and other transportation infrastructures in these three hub cities will promote the efficient connection of tourist flows among cities in GBA. In addition, the structural hole effect caused by the adjustment of tourist flows requires vigilance. For example, the calculation of structural hole indicators show that although Zhongshan and Jiangmen are located between Guangzhou-Foshan and Macao-Zhuhai, the two development poles of GBA, they lack a competitive advantage in the tourist flow network. Therefore, giving full play to the positive role of competitive cities and avoiding their negative impact on tourist flows in surrounding cities is a vital issue that tourism managers and policymakers need to address.

The analysis clarified the development status of cities in GBA. For example, Zhaoqing has the most obvious competitive disadvantage in the tourist flow network, which is consistent with the city's status as the lowest ranking city in total annual GDP and proportion of service industry output value. Zhaoqing is a city that has yet to break through the development dilemma in terms of its industrial structure. Although Zhuhai's ''static'' tourism popularity (i.e., number of visits) is not high, Zhuhai's high weighted degree and high flow betweenness centrality in the network reveal its strong ''dynamic'' role in circulation of tourist flows, which shows its tourism development potential. Macao has no obvious structural hole effect in the tourist flow network, and lacks participation in the competition of tourist flows. Macao is thus indeed a suitable place for building recreation centres in GBA.

VI. CONCLUSION

This paper proposed a novel research framework for extracting the network patterns of tourist flows in urban agglomeration through digital footprints. Geo-located travel blog data from Qunar.com were collected for the GBA case study. The conclusions can be summarized as follows:

At the attraction level, the research framework effectively evaluated the complexity of the tourist flow network. In the case of GBA, the heterogeneity of the network was proved by its significant scale-free characteristics, and the development issue of ''administrative barriers'' was revealed by the detected communities.

At the city level, the research framework clarified the interaction of tourist flows among cities from a sociological perspective. In the case of GBA, Guangzhou, Zhuhai, and Shenzhen play a meditating role in the tourist flow network. Zhongshan and Jiangmen are the most disadvantaged in competition for tourism.

The empirical results of the research framework can provide a theoretical basis and practical implications for the planning and management of tourism industry in urban

agglomerations. The new findings of empirical results are mainly reflected in three aspects. Firstly, the pattern of ''four cores and three poles'' in GBA was extracted from the perspective of tourism, which corresponds to the ''Outline Development Plan for the Guangdong-Hong Kong-Macao Greater Bay Area'' promulgated by the Chinese government [50]. Secondly, the ''administrative barrier effect'' was quantitatively confirmed to have an impact on the tourism in GBA. Finally, the status of cities in the tourism of GBA was effectively determined through network indicators. In addition, our case study of GBA also provides implications for other bay areas or urban agglomerations. For example, the ''four cores and three poles'' pattern of tourism in GBA may give guidelines for tourism development in single-core Bay areas such as Tokyo Bay Area in Japan. The structural hole effect, caused by the limits of surrounding dominant cities, has impacts on the tourism of Zhaoqing, Zhongshan, and Jiangmen in GBA. Similar phenomena may have occurred in other urban agglomerations similar to GBA, such as the Yangtze River Delta urban agglomeration, which should arouse the vigilance of tourism managers.

In summary, the framework proposed in this paper is highly feasible and can be applied to other tourists' digital footprint data sources. At the theoretical level, our research framework simultaneously introduces network science and multi-scale analysis into tourist flow research, allowing for a more comprehensive and systematic extraction of network patterns of tourist flows. At the application level, our research framework was applied to GBA, the first Bay Area Urban Agglomeration in China, and exhibited the pattern detection findings in the form of a simplified map (Figure 9), which provided intuitive implications for tourism managers and policymakers.

However, the limitations of the research and suggestions for future work are as follows: although travel blog data has advantages in describing tourists' behavior compared with other digital footprint data sources (see section 3.3), its sample size is usually small, and research is thus often based on only hundreds of blogs [29], [48], [51]. Therefore, the use of travel blog data is often accompanied by user representativeness and data bias. The users of Qunar are mainly born between the years 1980 and 1999 and have high educational levels and income [52]. Furthermore, although Qunar has been operating in Hong Kong and Macao for nearly ten years, it is still less popular in these two regions than in mainland China. Therefore, the case study in this paper may be affected by this ''skewness.'' As a typical social media data source, an ideal bias evaluation of travel blog data can be carried out from four perspectives (who, where, when, and what), but there is still a lack of solutions to completely address or quantify these perspectives [53]. In view of the above limitations, an obvious next step is to improve the data collection part of the research framework, such as by incorporating multi-source data fusion, and designing more effective methods of data cleaning and data bias evaluation. Furthermore, our research framework only utilizes the geotags of travel blogs, with no regard for the blog text. We will utilize the text of travel blogs to expand the framework since it can assist us comprehend the driving forces of tourist flows.

ACKNOWLEDGMENT

The author is grateful to Yunhao Zheng for the inspiration of research idea. He also offered valuable suggestions for data visualization and paper writing.

REFERENCES

[1] C. Fang and D. Yu, ''Urban agglomeration: An evolving concept of an emerging phenomenon,'' *Landscape Urban Planning*, vol. 162, pp. 126–136, Jun. 2017, doi: 10.1016/j.landurbplan.2017.02.014.

[2] R. Li, Q. Guo, D. Wu, H. Yin, H. Zhang, and T. Zhu, ''Spatial characteristics of development efficiency for urban tourism in eastern China: A case study of six coastal urban agglomerations,'' *J. Geographical Sci.*, vol. 24, no. 6, pp. 1175–1197, Oct. 2014, doi: 10.1007/s11442-014-1146-7.

[3] D.-G. Wang, Y. Niu, and J. Qian, ''Evolution and optimization of China's urban tourism spatial structure: A high speed rail perspective,'' *Tourism Manage.*, vol. 64, pp. 218–232, Feb. 2018, doi: 10.1016/j.tourman.2017.08.010.

[4] J. Wan, J. Yan, X. Wang, Z. Liu, H. Wang, and T. Wang, ''Spatial-temporal pattern and its influencing factors on urban tourism competitiveness in city agglomerations across the Guanzhong plain,'' *Sustainability*, vol. 11, no. 23, p. 6743, Nov. 2019, doi: 10.3390/su11236743.

[5] Y. Liu, X. Liu, S. Gao, L. Gong, C. Kang, and Y. Zhi, ''Social sensing: A new approach to understanding our socioeconomic environments,'' *Ann. Assoc. Amer. Geograph.*, vol. 105, no. 3, pp. 512–530, 2015, doi: 10.1080/00045608.2015.1018773.

[6] N. Mou, Z. Liu, Y. Zheng, T. Makkonen, T. Yang, and L. Zhang, ''Cycling in Tibet: An analysis of tourists' spatiotemporal behavior and infrastructure,'' *Tourism Manage.*, vol. 88, Feb. 2022, Art. no. 104418, doi: 10.1016/j.tourman.2021.104418.

[7] F. Liu, J. Zhang, J. Zhang, D. Chen, Z. Liu, and S. Lu, ''Roles and functions of tourism destinations in tourism region of south Anhui: A tourist flow network perspective,'' *Chin. Geographical Sci.*, vol. 22, no. 6, pp. 755–764, Jul. 2012, doi: 10.1007/s11769-012-0557-6.

[8] A. M. Caldeira and E. Kastenholz, ''Spatiotemporal tourist behaviour in urban destinations: A framework of analysis,'' *Tourism Geographies*, vol. 22, no. 1, pp. 22–50, Jan. 2020, doi: 10.1080/14616688.2019.1611909.

[9] S. Su, C. Wan, Y. Hu, and Z. Cai, ''Characterizing geographical preferences of international tourists and the local influential factors in China using geo-tagged photos on social media,'' *Appl. Geography*, vol. 73, pp. 26–37, Aug. 2016, doi: 10.1016/j.apgeog.2016.06.001.

[10] F. Batista e Silva, M. A. Marín Herrera, K. Rosina, R. R. Barranco, S. Freire, and M. Schiavina, ''Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources,'' *Tourism Manage.*, vol. 68, pp. 101–115, Oct. 2018, doi: 10.1016/j.tourman.2018.02.020.

[11] K. Li, C. Liang, W. Lu, C. Li, S. Zhao, and B. Wang, ''Forecasting of short-term daily tourist flow based on seasonal clustering method and PSO-LSSVM,'' *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 11, p. 676, Nov. 2020, doi: 10.3390/ijgi9110676.

[12] F. Girardin, F. Calabrese, F. D. Fiore, C. Ratti, and J. Blat, ''Digital footprinting: Uncovering tourists with user-generated content,'' *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 36–43, Oct. 2008, doi: 10.1109/MPRV.2008.71.

[13] J. Xia, P. Zeephongsekul, and C. Arrowsmith, ''Modelling spatio-temporal movement of tourists using finite Markov chains,'' *Math. Comput. Simul.*, vol. 79, no. 5, pp. 1544–1553, Jan. 2009, doi: 10.1016/j.matcom.2008.06.007.

[14] J. Rodríguez-Echeverría, I. Semanjski, C. Van Gheluwe, D. Ochoa, H. IJben, and S. Gautama, ''Density-based spatial clustering and ordering points approach for characterizations of tourist behaviour,'' *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 11, p. 686, Nov. 2020, doi: 10.3390/ijgi9110686.

[15] X. Su, B. Spierings, M. Dijst, and Z. Tong, ''Analysing trends in the spatio-temporal behaviour patterns of mainland Chinese tourists and residents in Hong Kong based on Weibo data,'' *Current Issues Tourism*, vol. 23, no. 12, pp. 1542–1558, Jun. 2020, doi: 10.1080/13683500.2019.1645096.

[16] Y. Xing-zhu and W. Qun, "Exploratory space-time analysis of inbound tourism flows to China cities," *Int. J. Tourism Res.*, vol. 16, no. 3, pp. 303–312, May 2014, doi: 10.1002/jtr.1932.
[17] C. Casanueva, Á. Gallego, and M.-R. García-Sánchez, "Social network analysis in tourism," *Current Issues Tourism*, vol. 19, no. 12, pp. 1190–1209, Oct. 2016, doi: 10.1080/13683500.2014.990422.
[18] I. Önder, W. Koerbitz, and A. Hubmann-Haidvogel, "Tracing tourists by their digital footprints: The case of Austria," *J. Travel Res.*, vol. 55, no. 5, pp. 566–573, Dec. 2014, doi: 10.1177/0047287514563985.
[19] R. Roura, "Being there: Examining the behaviour of Antarctic tourists through their blogs," *Polar Res.*, vol. 31, no. 1, pp. 10905–10923, Dec. 2012, doi: 10.3402/polar.v31i0.10905.
[20] Y. Li, L. Yang, H. Shen, and Z. Wu, "Modeling intra-destination travel behavior of tourists through spatio-temporal analysis," *J. Destination Marketing Manage.*, vol. 11, pp. 260–269, Mar. 2019, doi: 10.1016/j.jdmm.2018.05.002.
[21] C. Jing, M. Dong, M. Du, Y. Zhu, and J. Fu, "Fine-grained spatiotemporal dynamics of inbound tourists based on geotagged photos: A case study in Beijing, China," *IEEE Access*, vol. 8, pp. 28735–28745, 2020, doi: 10.1109/ACCESS.2020.2972309.
[22] C. Walden-Schreiner, Y.-F. Leung, and L. Tateosian, "Digital footprints: Incorporating crowdsourced geographic information for protected area management," *Appl. Geography*, vol. 90, pp. 44–54, Jan. 2018, doi: 10.1016/j.apgeog.2017.11.004.
[23] N. Mou, Y. Zheng, T. Makkonen, T. Yang, J. J. Tang, and Y. Song, "Tourists' digital footprint: The spatial patterns of tourist flows in Qingdao, China," *Tourism Manage.*, vol. 81, Dec. 2020, Art. no. 104151, doi: 10.1016/j.tourman.2020.104151.
[24] P. Martí, C. García-Mayor, and L. Serrano-Estrada, "Taking the urban tourist activity pulse through digital footprints," *Current Issues Tourism*, vol. 24, no. 2, pp. 157–176, Jan. 2021, doi: 10.1080/13683500.2019.1706458.
[25] H.-Y. Shih, "Network characteristics of drive tourism destinations: An application of network analysis in tourism," *Tourism Manage.*, vol. 27, no. 5, pp. 1029–1039, 2006, doi: 10.1016/j.tourman.2005.08.002.
[26] R. Baggio, N. Scott, and C. Cooper, "Network science: A review focused on tourism," *Ann. Tourism Res.*, vol. 37, no. 3, pp. 802–827, Jul. 2010, doi: 10.1016/j.annals.2010.02.008.
[27] J. Qin, C. Song, M. Tang, Y. Zhang, and J. Wang, "Exploring the spatial characteristics of inbound tourist flows in China using geotagged photos," *Sustainability*, vol. 11, no. 20, p. 5822, Oct. 2019, doi: 10.3390/su11205822.
[28] Y. Xu, J. Li, A. Belyi, and S. Park, "Characterizing destination networks through mobility traces of international tourists—A case study using a nationwide mobile positioning dataset," *Tourism Manage.*, vol. 82, Feb. 2021, Art. no. 104195, doi: 10.1016/j.tourman.2020.104195.
[29] B. Zeng, "Pattern of Chinese tourist flows in Japan: A social network analysis perspective," *Tourism Geographies*, vol. 20, no. 5, pp. 810–832, Oct. 2018, doi: 10.1080/14616688.2018.1496470.
[30] D. Provenzano, B. Hawelka, and R. Baggio, "The mobility network of European tourists: A longitudinal study and a comparison with geo-located Twitter data," *Tourism Rev.*, vol. 73, no. 1, pp. 28–43, Feb. 2018, doi: 10.1108/TR-03-2017-0052.
[31] H. Chung, N. Chung, and Y. Nam, "A social network analysis of tourist movement patterns in blogs: Korean backpackers in Europe," *Sustainability*, vol. 9, no. 12, p. 2251, Dec. 2017, doi: 10.3390/su9122251.
[32] B. Liu, S. Huang, and H. Fu, "An application of network analysis on tourist attractions: The case of Xinjiang, China," *Tourism Manage.*, vol. 58, pp. 132–141, Feb. 2017, doi: 10.1016/j.tourman.2016.10.009.
[33] C. Jin, J. Cheng, and J. Xu, "Using user-generated content to explore the temporal heterogeneity in tourist mobility," *J. Travel Res.*, vol. 57, no. 6, pp. 779–791, Jul. 2018, doi: 10.1177/0047287517714906.
[34] 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, May 2008, doi: 10.1016/j.physa.2008.01.058.
[35] X. Wu, Z. Huang, X. Peng, Y. Chen, and Y. Liu, "Building a spatially-embedded network of tourism hotspots from geotagged social media data," *IEEE Access*, vol. 6, pp. 21945–21955, 2018, doi: 10.1109/ACCESS.2018.2828032.
[36] Y. Gao, C. Ye, X. Zhong, L. Wu, and Y. Liu, "Extracting spatial patterns of intercity tourist movements from online travel blogs," *Sustainability*, vol. 11, no. 13, p. 3526, Jun. 2019, doi: 10.3390/su11133526.
[37] D. G. Pearce, "An integrative framework for urban tourism research," *Ann. Tourism Res.*, vol. 28, no. 4, pp. 926–946, Aug. 2001, doi: 10.1016/S0160-7383(00)00082-7.
[38] J. Gao, P. Peng, F. Lu, and C. Claramunt, "A multi-scale comparison of tourism attraction networks across China," *Tourism Manage.*, vol. 90, Jun. 2022, Art. no. 104489, doi: 10.1016/j.tourman.2022.104489.
[39] L. C. Freeman, S. P. Borgatti, and D. R. White, "Centrality in valued graphs: A measure of betweenness based on network flow," *Social Netw.*, vol. 13, no. 2, pp. 141–154, Jun. 1991, doi: 10.1016/0378-8733(91)90017-N.
[40] R. S. Burt, "Structural holes and good ideas," *Amer. J. Sociol.*, vol. 110, no. 2, pp. 349–399, Sep. 2004, doi: 10.1086/421787.
[41] N. Mou, J. Li, S. Sun, T. Yang, L. Zhang, H. Zhang, and W. Liu, "The impact of opening the Arctic northeast passage on the global maritime transportation network pattern using AIS data," *Arabian J. Geosci.*, vol. 13, no. 11, p. 419, May 2020, doi: 10.1007/s12517-020-05432-5.
[42] V. A. Traag, L. Waltman, and N. J. van Eck, "From Louvain to leiden: Guaranteeing well-connected communities," *Sci. Rep.*, vol. 9, no. 1, p. 5233, Mar. 2019, doi: 10.1038/s41598-019-41695-z.
[43] (2019). *Secretary General of the World Tourism Organization: Guangdong-Hong Kong-Macao Greater Bay Area are the Best Examples of Global Tourism Growth*. [Online]. Available: http://kb.southcn.com/content/2019-10/16/content_189244997.htm
[44] Q. Zhu. (2018). *Tourism Industry in Guangdong-Hong Kong-Macao Greater Bay Area: Cities Have Different Tourism Characteristics*. [Online]. Available: https://www.qianzhan.com/analyst/detail/220/180727-bc2fac44.html
[45] M. H. Salas-Olmedo, B. Moya-Gómez, J. C. García-Palomares, and J. Gutiérrez, "Tourists' digital footprint in cities: Comparing big data sources," *Tourism Manage.*, vol. 66, pp. 13–25, Jun. 2018, doi: 10.1016/j.tourman.2017.11.001.
[46] Bala Ranking. (2020). *Ranking of Top Ten Travel Websites in China: Ctrip Ranks First and Qunar Ranks Second*. [Online]. Available: http://www.bala.cc/shenghuo/pinpai/2020/7254.html
[47] M. Kaufmann, P. Siegfried, L. Huck, and J. Stettler, "Analysis of tourism hotspot behaviour based on geolocated travel blog data: The case of qyer," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 11, p. 493, Nov. 2019, doi: 10.3390/ijgi8110493.
[48] Y. Zheng, N. Mou, L. Zhang, T. Makkonen, and T. Yang, "Chinese tourists in Nordic countries: An analysis of spatio-temporal behavior using geo-located travel blog data," *Comput., Environ. Urban Syst.*, vol. 85, Jan. 2021, Art. no. 101561, doi: 10.1016/j.compenvurbsys.2020.101561.
[49] N. Mou, R. Yuan, T. Yang, H. Zhang, J. Tang, and T. Makkonen, "Exploring spatio-temporal changes of city inbound tourism flow: The case of Shanghai, China," *Tourism Manage.*, vol. 76, Feb. 2020, Art. no. 103955, doi: 10.1016/j.tourman.2019.103955.
[50] CPC Central Committee & the State Council of the PRC. (2019). *Outline Development Plan for the Guangdong-Hong Kong-Macao Greater Bay Area*. [Online]. Available: http://www.xinhuanet.com/politics/2019-02/18/c_1124131474.htm
[51] X. Y. Leung, F. Wang, B. Wu, B. Bai, K. A. Stahura, and Z. Xie, "A social network analysis of overseas tourist movement patterns in Beijing: The impact of the Olympic games," *Int. J. Tourism Res.*, vol. 14, no. 5, pp. 469–484, Sep. 2012, doi: 10.1002/jtr.876.
[52] IResearch. (2016). *Research Report on the Media Value of Users From Qunar.com*. [Online]. Available: https://www.iresearch.com.cn/Detail/report?id=2661&isfree=0
[53] Y. Yuan, Y. Lu, T. E. Chow, C. Ye, A. Alyaqout, and Y. Liu, "The missing parts from social media–enabled smart cities: Who, where, when, and what?" *Ann. Amer. Assoc. Geographers*, vol. 110, no. 2, pp. 462–475, Aug. 2019, doi: 10.1080/24694452.2019.1631144.

JIARUI MOU is currently pursuing the bachelor's degree with the College of Resources, Shandong University of Science and Technology, China. Her main research interests include spatial data mining and tourism geography.