

## Article

# Spatial-Temporal Analysis on Spring Festival Travel Rush in China Based on Multisource Big Data

Jiwei Li <sup>1,2</sup>, Qingqing Ye <sup>3</sup>, Xuankai Deng <sup>1,2</sup>, Yaolin Liu <sup>1,2,\*</sup> and Yanfang Liu <sup>1,2,\*</sup>

<sup>1</sup> School of Resource and Environmental Science, Wuhan University, Wuhan 430079, China; lijawei19850620@126.com (J.L.); xuankdeng@gmail.com (X.D.)

<sup>2</sup> Key Laboratory of Geographic Information Systems, Ministry Education, Wuhan University, Wuhan 430079, China

<sup>3</sup> School of Engineering Management and Real Estate, Henan University of Economics and Law, Zhengzhou 450002, China; yeqing0804@163.com

\* Correspondence: yaolin610@163.com (Y.L.L.); yfliu610@163.com (Y.f.L.); Tel.: +86-27-6877-8552 (Y.L.L.); +86-27-6877-8863 (Y.f.L.)

Academic Editor: Tan Yigitcanlar

Received: 21 September 2016; Accepted: 13 November 2016; Published: 17 November 2016

**Abstract:** Spring Festival travel rush is a phenomenon in China that population travel intensively surges in a short time around Chinese Spring Festival. This phenomenon, which is a special one in the urbanization process of China, brings a large traffic burden and various kinds of social problems, thereby causing widespread public concern. This study investigates the spatial-temporal characteristics of Spring Festival travel rush in 2015 through time series analysis and complex network analysis based on multisource big travel data derived from Baidu, Tencent, and Qihoo. The main results are as follows: First, big travel data of Baidu and Tencent obtained from location-based services might be more accurate and scientific than that of Qihoo. Second, two travel peaks appeared at five days before and six days after the Spring Festival, respectively, and the travel valley appeared on the Spring Festival. The Spring Festival travel network at the provincial scale did not have small-world and scale-free characteristics. Instead, the travel network showed a multicenter characteristic and a significant geographic clustering characteristic. Moreover, some travel path chains played a leading role in the network. Third, economic and social factors had more influence on the travel network than geographical location factors. The problem of Spring Festival travel rush will not be effectively improved in a short time because of the unbalanced urban-rural development and the unbalanced regional development. However, the development of the modern high-speed transport system and the modern information and communication technology can alleviate problems brought by Spring Festival travel rush. We suggest that a unified real-time traffic platform for Spring Festival travel rush should be established through the government's integration of mobile big data and the official authority data of the transportation department.

**Keywords:** Spring Festival travel rush; migration; spatial-temporal analysis; big data; complex network; urban-rural development; regional development

## 1. Introduction

Since 1989, hundreds of millions of passengers in China have been returning to their hometowns from their working cities before the Spring Festival to reunite with their families and traveling back after the holiday. This phenomenon generates the largest annual human migration on Earth, and the Chinese call this phenomenon as Chunyun, which is also called Spring Festival travel rush. Spring Festival travel rush is a special phenomenon in the urbanization process of China. Transport systems, particularly rail transport, experience great challenges during this period. Subsequently, various kinds

of social problems emerge, thereby causing widespread public concern and massive government investment. Therefore, research on the problem of Spring Festival travel rush has important practical significance, and many related works have been conducted [1,2]. However, before 2014, the majority of these works focused on qualitative theory study or large-scale time series analysis because of lack of data, thereby limiting further research. Small-scale quantitative studies can be conducted until Baidu published the Baidu Migration production based on its location-based service (LBS) in 2014. Since then, several small-scale quantitative studies on Spring Festival travel rush have begun to emerge.

Wang et al. used the Baidu Migration data to study Spring Festival travel rush for the first time in 2014 [3]. They explored the travel trend during the Spring Festival travel season, estimated and quantified the migration directions and routes, and visualized the travel flows. Li compared Spring Festival travel data of Baidu, Tencent, and Qihoo in 2015 from the aspects of data characteristics, overall travel trend, and spatial distribution [4]. Furthermore, he suggested the establishment of a set of open standards used to observe the population travel based on Internet user data. These two studies were not published in academic journals, and their analyses of Spring Festival travel data were relatively simple. However, our study uses methods such as complex network analysis to further analyze these data. In addition, some researchers used social media check-in data to study the inter-urban or inter-area population movements in China, and these were related significant studies [5,6]. The studies given above all benefited from big data technologies.

The topic of big data, which is characterized by volume, variety, velocity, and value, is currently one of the research focuses [7,8]. LBS data is one type of big data obtained from GPS and mobile phones. This type of data is now extensively used in spatial analyses, including spatial distribution, spatial trajectory, and spatial interaction [9]. These analyses based on LBS data can achieve real-time dynamic monitoring and prediction of population movements, which can help the formulation of relevant policies and countermeasures. This study uses the big travel data of Chinese Spring Festival derived from Baidu, Tencent, and Qihoo to investigate the spatial-temporal characteristics of Spring Festival travel rush.

Complex network analysis is a collection of quantitative methods for interaction data, and it provides an advanced analytical tool to study population travel and migration. The majority of complex network analysis methods are derived from graph theory, statistical physics, and social network analysis. The research contents of this analysis include centrality, assortativity, small-world and scale-free characteristics, community discovery, and network modeling [10–12]. The majority of human interactions occur in geographic space; thus, spatially embedded networks have become a hotspot of complex network research [13,14]. This type of complex network research mainly focuses on the characteristics of spatially embedded networks and on how spatial distance affects network formation. This study investigates the spatial characteristics of Spring Festival travel network.

Our research has three specific purposes. First, we compare three types of Spring Festival big travel data derived from Baidu, Tencent, and Qihoo, and their data characteristics are analyzed. This analysis is helpful to the future research. Second, we use these big travel data through time series analysis and complex network analysis to obtain the spatial-temporal characteristics of Spring Festival travel rush. Last, we investigate the influence factors of Spring Festival travel rush and propose some related policy suggestions on the problem.

## 2. Materials and Methods

### 2.1. Data Description

As mentioned earlier, data acquisition, particularly acquisition of small-scale data, is the most significant limitation to the research on Spring Festival travel rush. Big data technologies, particularly LBS data, provide a promising solution. However, big data are characterized by multiple sources and lack of structure. Moreover, each type of big data has its own characteristics and application scope. Therefore, the use of big data requires considerable caution compared with the use of traditional data,

and data validation and comparison are the premises. The Spring Festival travel data used in this study are obtained through the web crawler from special websites of Baidu, Tencent, and Qihoo. These three travel datasets have some differences in data sources, time range, time resolution, space resolution, and data volume (Table 1). Therefore, scientific data preprocessing is necessary for these three travel datasets before analysis.

These three travel datasets are representative in general. The data of Baidu and Tencent are typical LBS data. Baidu is the largest electronic map and LBS provider in China, and thousands of mobile applications use its services. Mobile applications, such as Baidu Maps and MoWeather, have millions of users. Tencent owns the two largest mobile social applications in China, namely, QQ and WeChat, and both of them have over 500 million users. The LBS built in the two applications handles millions of location requests. However, the data of Qihoo are obtained from its automatic train ticket booking platform built in a browser. This browser occupies a quarter of browser market share in China. During Spring Festival travel season of 2015, the booking platform provided booking services of more than 17 million times, accounting for approximately 10% of the total online bookings.

**Table 1.** Travel data characteristics of Baidu, Tencent, and Qihoo.

Data Types	Baidu	Tencent	Qihoo
Data sources	Baidu LBS	Tencent LBS	automatic train ticket booking platform of Qihoo
Time range	7 February 2015 to 31 March 2015	24 January 2015 to 12 March 2015	5 February 2015 to 5 March 2015
Time resolution	Per hour	Per day	Per day
Space resolution	Cities and provinces	Cities	Provinces
Data size per collection	Top 4000 intercity flows; 369 cities' top 10 inflows; 369 cities' top 10 outflows; 34 provinces' top 10 inflows; 34 provinces' top 10 outflows (Not provide intercity travel data within a province.)	365 cities' top 10 inflows; 365 cities' top 10 outflows	31 provinces' 31 inflows; 31 provinces' 31 outflows

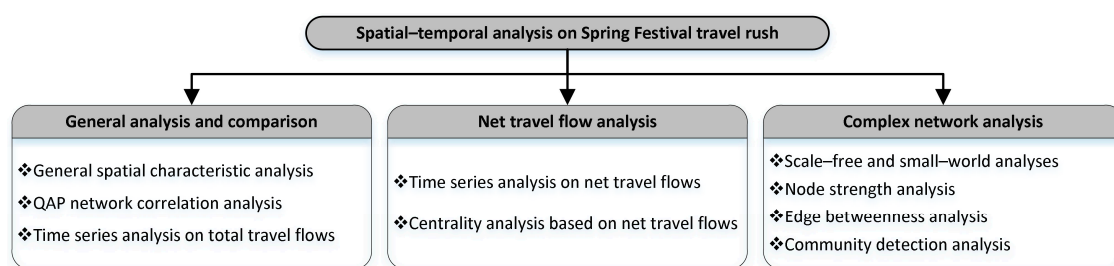
Validation and comparison of these three travel datasets need a uniform time range and a uniform spatial-temporal resolution under the premise of retaining data information as much as possible. Spring Festival travel season usually begins 15 days before the Spring Festival and lasts for approximately 40 days. Then, the period given by the government in 2015, that is, 40 days from 4 February 2015 to 15 March 2015, is adopted as the time range, and one day is adopted as the time resolution. Province, which is the study unit of this work, is adopted as the space resolution because of incompleteness or lack of intercity travel data. Baidu and Tencent provide 369 and 365 cities' data respectively, and their time resolution is hours for Baidu and days for Tencent. They all provide top 10 inflows and top 10 outflows per collection, and Baidu provides extra top 4000 flows per collection. Therefore, the raw data at city-level include  $40 \times 24 \times 369 \times 2 \times 10 + 4000$  flows for Baidu and  $40 \times 365 \times 2 \times 10$  flows for Tencent. The raw data are reorganized into network data at province-level, including  $31 \times 31$  flows for Baidu and Tencent, which is consistent with Qihoo (Supplementary Materials).

Moreover, migration data of 2010 national census are introduced as an authoritative reference to validate the three travel datasets. The census migration data is long-term macro migration data to some extent, whereas the three travel datasets can be considered short-term migration data. Furthermore, the acquisition times of the two kinds of data is different, that is, the census data is 2010 and the three travel datasets are 2015. This difference may bring some bias to the two kinds of data. However, studies on population migration or movements could only use the census data in China before 2014 when Baidu published the Baidu Migration production. Moreover, the high similarity between the two kinds of data, shown in the following QAP network correlation analysis, further proves the rationality of choosing the census data as a reference in our study. Thus, the census migration data are considered

a relative reference rather than an absolute standard. The census provides two kinds of migration data: one is based on the actual residence and registered permanent residence, and the other is based on the current residence and residence five years ago. The majority of migrants in the former migration data do not settle down in the cities where they work. This characteristic is an important one of the population migration in present China, which is more consistent with Spring Festival travel rush. Therefore, the migration data based on the actual residence and registered permanent residence is chosen as the reference.

## 2.2. Research Methods

Spring Festival travel rush is studied from three aspects in our research (Figure 1). First, spatial-temporal characteristics of the three travel datasets are generally analyzed and compared. Then, further in-depth analyses are conducted from aspects of net travel flows and complex network structures, respectively. Time series analysis and complex network analysis are the main methods used in the research. Furthermore, a centrality analysis method based on net travel flows is proposed to identify the function and population attractiveness of provinces in the travel network.



**Figure 1.** The analysis framework and methods for Spring Festival travel rush.

### 2.2.1. Time Series Analysis

Time series analysis comprises methods for analyzing time series data to extract meaningful statistics and other characteristics of data. The methods used in our research are simple but adequate. Time series line charts are used to compare the three travel datasets from the time dimension and to investigate their differences and the temporal characteristics of Spring Festival travel rush, such as peak and valley values of travel flows.

Time series analysis is mainly used in two places of our research: one is the overall trend analysis of the total flow, and the other is the trend analysis of the net travel flow of each province. The former compares the overall trend of the three travel datasets using original data and normalized data, respectively. Their peak and valley values of total travel flows are detected and corresponding similarities and differences are analyzed. Then, an overall trend of Spring Festival travel rush is obtained from these analyses. The latter puts the net travel flow lines of all provinces in three charts for the three travel datasets, respectively. The immigration and emigration clusters are identified and compared. Moreover, peak and valley values of the net travel flow for the three travel datasets are detected and compared with the overall travel trend.

### 2.2.2. Centrality Analysis Based on Net Travel Flows

Some issues of Spring Festival travel rush including identifying the immigration and emigration provinces and determining the strength of population attractiveness are public concerns. This study presents a method based on net travel flows to solve these issues. Net travel flows are outflows minus inflows. For immigration provinces, net travel flows are positive before the Spring Festival and negative after the Spring Festival. The opposite can be determined for emigration provinces. Therefore, we use net travel flows before the Spring Festival as the horizontal axis and net travel flows after the Spring Festival as the vertical axis to establish a four-quadrant diagram. The provinces of immigration

and emigration are located in the fourth and second quadrants, respectively, and the distance to the center of the four-quadrant diagram indicates the attractive strength.

In our research, this centrality analysis based on net travel flows is conducted after the time series analysis of net travel flows. Firstly, fitting line slopes and point distributions of scatter diagrams for the three travel datasets are compared to get an overview of their centralities. Then, the provinces of immigration and emigration are identified, and the reasons and subdivision types are further analyzed and presented. Next, centrality ranks of the three travel datasets, namely, the population attraction strength, are measured using the distances from provincial points to the center of the four-quadrant diagram. Finally, Spearman correlation coefficients among the travel and migration data, which is a nonparametric measure of rank correlation, are calculated to obtain a quantitative analysis result of different data's centrality similarity.

### 2.2.3. Complex Network Analysis

Systematic research on complex networks originated from the random graph proposed by Hungarian mathematicians Erdős and Rényi in 1960. At the end of the 1990s, Watts and Strogatz proposed the small-world network, whereas Albert and Barabasi proposed the scale-free network. Their works made a pioneering contribution to complex network research and inspired a wave of research fever. Since then, numerous research results have emerged. Typical methods of complex network analysis, including scale-free and small-world analyses, centrality analysis, and community detection, are used in our research to analyze the characteristics of Spring Festival travel network. Unlike the majority of previous research, our work uses weighted network methods rather than binary or topology ones because of three reasons. First, the binarization results in data loss. Second, determining the binarization threshold scientifically is difficult because of the scale-free property of edge weights, and the majority of binarizations are subjective. Third, research on weighted networks has become popular and has matured gradually [15–17].

Scale-free networks are defined as networks whose node degree distributions follow a power law. In weighted networks, node degree is replaced by node strength [17]. The strength of node  $i$  is defined as

$$s_i = \sum_{j \in V} a_{ij} w_{ij} \quad (1)$$

where  $V$  is the set of nodes in a network,  $a_{ij}$  takes the value 1 if an edge connects node  $i$  to node  $j$  and 0 otherwise, and  $w_{ij}$  is the weight of the edge connecting node  $i$  to node  $j$ .

Small-world networks are networks having a short average shortest path length and a high average clustering coefficient. Various kinds of average shortest path lengths and average clustering coefficients have been proposed for weighted networks [18,19]. The measures used in this research are as follows:

$$L_g = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)} \quad (2)$$

$$C_g = \frac{1}{\deg(u)(\deg(u)-1)} \sum_{uv} (\hat{w}_{uv} \hat{w}_{uw} \hat{w}_{vw})^{\frac{1}{3}} \quad (3)$$

where  $n$  is the number of network nodes,  $d(s,t)$  is the shortest path length from node  $s$  to node  $t$ ,  $\deg(u)$  is the degree of node  $u$ , and  $\hat{w}_{uv}$  is the edge weight normalized by the maximum weight in the network.

The criteria of the average shortest path length and average clustering coefficient are difficult to determine, and the corresponding random and regular networks are usually used as reference objects. However, the corresponding regular network for a weighted network is difficult to define. In 2008, Humphries and Gurney proposed the small-world-ness index to determine whether a network is a small-world one [20,21]. The index  $S$  is defined as



$$S = \frac{\gamma_g}{\lambda_g} = \frac{C_g}{C_{rand}} / \frac{L_g}{L_{rand}} \quad (4)$$

where  $C_{rand}$  and  $L_{rand}$  are the average shortest path length and average clustering coefficient, respectively, for the corresponding random network. A network is considered a small-world one if  $S$  is greater than 1.

Various measures of network centrality exist, such as degree (strength), betweenness, and closeness. Two types of centrality, namely, strength and betweenness, are used in this research [22]. The formula of edge betweenness is as follows:

$$c_B(e) = \sum_{s,t \in V} \frac{\sigma(s,t|e)}{\sigma(s,t)} \quad (5)$$

where  $\sigma(s,t)$  is the number of shortest  $(s,t)$ -paths, and  $\sigma(s,t|e)$  is the number of those paths passing through edge  $e$ .

Community detection is popular for current complex network research. This problem is an NP, and scholars have proposed various methods, particularly fast algorithms for large-scale networks, to solve it [23–26]. The multilevel algorithm, which is a bottom-up algorithm, proposed by Blondel et al. is adopted in this research to detect the community structure of Spring Festival travel networks by optimizing modularity [27].

### 3. Results

#### 3.1. General Analysis and Comparison

The general differences and characteristics of the three types of Spring Festival big travel data were investigated from three aspects: general spatial characteristics of travel networks, quadratic assignment procedure (QAP) correlation, and time series characteristics. The data used in the analysis were normalized through extremum method, and natural break method was used to classify data. Three travel networks and the census migration networks are shown in Figure 2, and the general spatial characteristics can be obtained. The networks of Baidu and Tencent had a similar characteristic, that is, strong travel flows mostly occurred between short-distance province pairs (Figure 2f,g). However, the Qihoo network had a different characteristic, that is, strong travel flows mostly occurred between relative long-distance province pairs, such as Guangdong-Sichuan, Guangdong-Hubei, Guangdong-Henan, Sichuan-Zhejiang, and Beijing-Heilongjiang (Figure 2h). The census network had a combination characteristic, that is, strong migration flows existed in both short- and long-distance province pairs (Figure 2e).

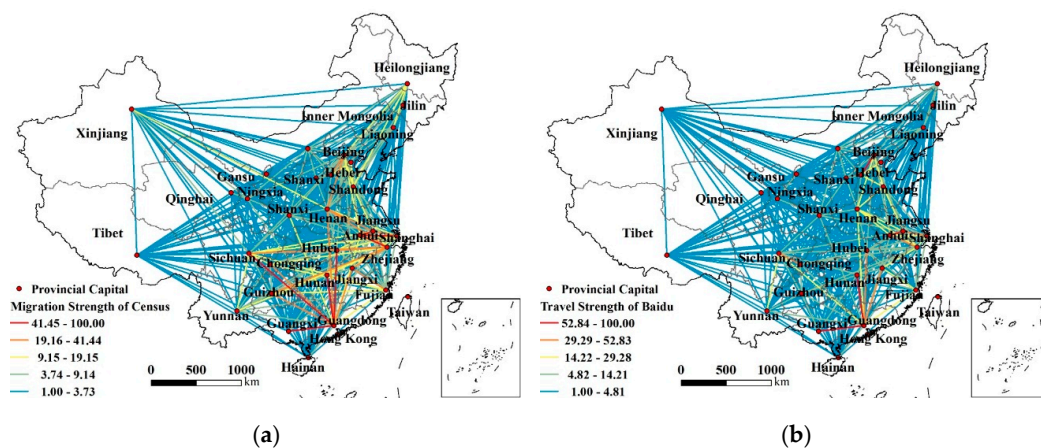
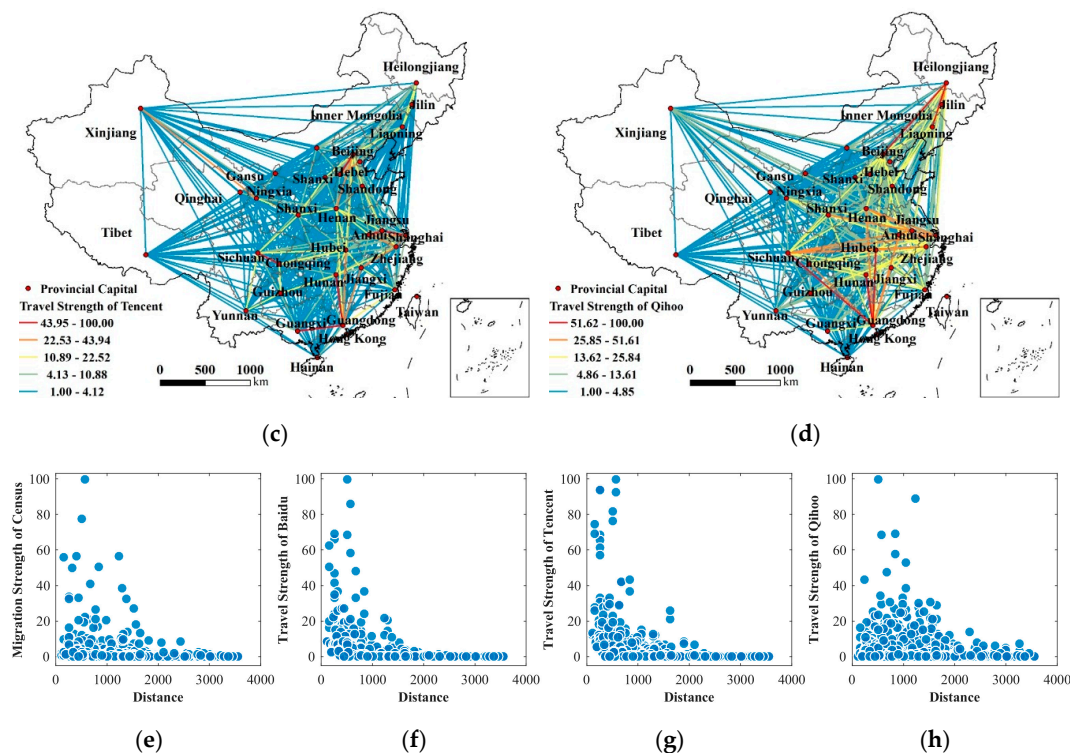


Figure 2. Cont.



**Figure 2.** Travel/migration networks of Spring Festival. (a–d) Travel/migration networks for the four data, namely, the census, Baidu, Tencent, and Qihoo, respectively; (e–h) the relation between distance and travel/migration strength for the four data. Each blue point represents a travel/migration edge between two provinces.

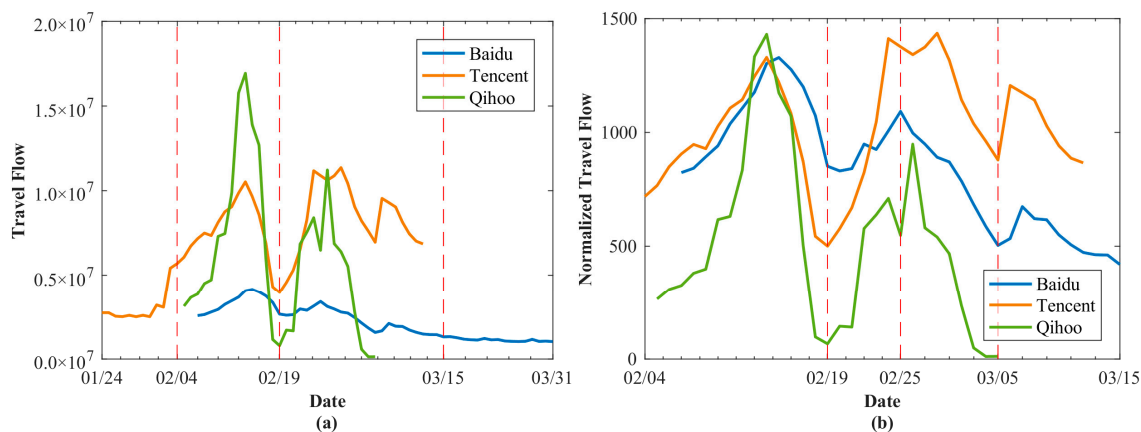
QAP is used to measure network correlation [28], and we use it to compare the linear and log-transformed correlations among the four networks (Table 2). The networks of the census, Baidu and Tencent, had high linear and log-transformed correlations with one another; in particular, the linear correlation between Baidu and Tencent reached 0.95. However, the network of Qihoo had a weak correlation with that of Baidu and Tencent and had a moderate correlation with that of the census. After logarithmic transformation, the correlations of Qihoo network with the others improved greatly, and the correlation with the census network reached 0.82.

**Table 2.** Correlation analysis of Spring Festival travel networks.

Linear/Logarithmic	Census	Baidu	Tencent	Qihoo
Census	1.00/1.00	0.88/0.81	0.79/0.81	0.67/0.82
Baidu		1.00/1.00	0.95/0.82	0.49/0.66
Tencent			1.00/1.00	0.46/0.70
Qihoo				1.00/1.00

As shown in the time series line charts of Spring Festival travel flows (Figure 3), the three original travel data had a great difference, which was possibly caused by platform user scale and data acquiring approach. However, these data had a similar overall travel trend, which was evident after data normalization. Furthermore, the travel strength of Tencent was generally stable, but the travel strength of Baidu and Qihoo showed a generally declining trend. The possible reason of this observation for Qihoo was that the train ticket supply after the Spring Festival was not as tight as it before the Spring Festival. Therefore, Qihoo could only provide data until the Lantern Festival (5 March 2015). The three travel datasets showed two travel flow peaks. One appeared approximately five days before the

Spring Festival, and the other appeared at approximately six days after the Spring Festival, which was the day the holiday finished. Two travel flow valleys appeared on two important festivals, namely, Spring Festival and Lantern Festival, which were the days for family reunion. As shown by Tencent and Qihoo data, two small peaks appeared, that is, one before and one after 25 February 2015, which was the day the holiday finished. However, Baidu data did not have this characteristic.



**Figure 3.** Time series of Spring Festival travel flows. (a) Original travel flows of Spring Festival; (b) Normalized travel flows of Spring Festival.

The following basic characteristics of the three travel datasets could be derived from the preceding analysis: First, the travel data of Baidu and Tencent had similar characteristics because both of them were obtained from LBS. Baidu and Tencent could acquire travel data including multiple modes of transportations, such as vehicles, trains, and planes, in real time. These modes of transportation covered short- and long-distance travel. Second, the Qihoo travel data, which were obtained from the third-party automatic train ticket-booking platform of Qihoo, might have two distinguishing characteristics. On the one hand, the platform could only acquire the railway passenger flow data characterized by long-distance travel, as mentioned earlier. On the other hand, the platform could not collect the whole train ticket booking data because this platform was a third party, thereby resulting in incompleteness of the data. Third, the census migration data used as a reference represented long-term macro migration. The characteristic of the census migration data—namely, strong migration flows—existed in both short- and long-distance province pairs, was reasonable. This finding was consistent with the Levy flight model.

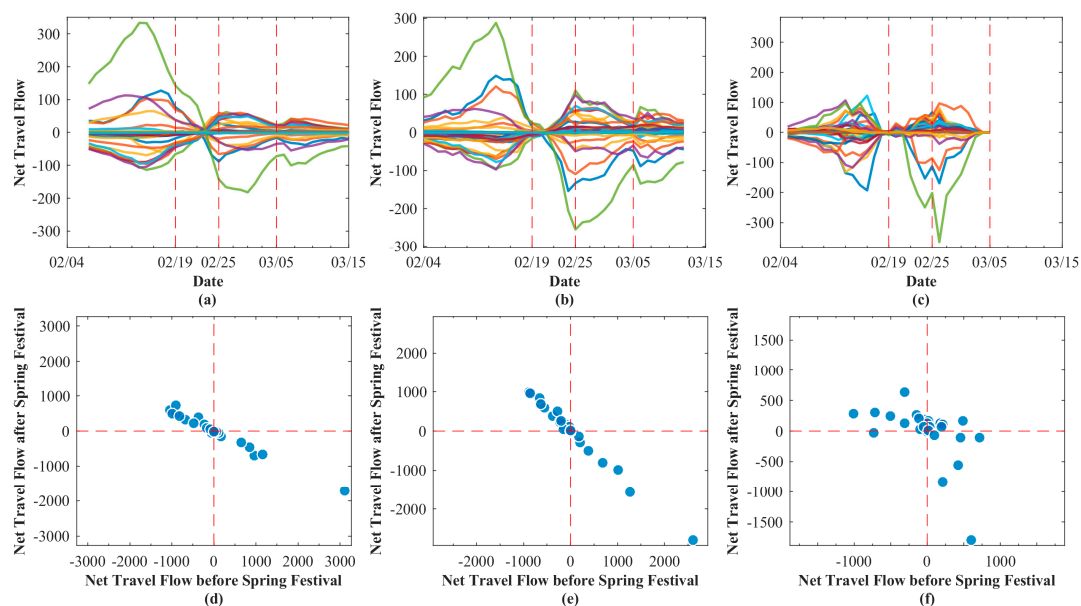
### 3.2. Net Travel Flow Analysis

Net travel flows could be used to determine that provinces are immigration or emigration ones in the whole travel network. The time series of net travel flows showed that the three travel datasets consistently formed two significant clusters (Figure 4a–c). One was immigration cluster whose net travel flows was positive before the Spring Festival and negative after the Spring Festival, whereas the other was emigration cluster, which behaved oppositely. The overall variation of immigration provinces was significantly larger than that of emigration provinces for Baidu and Tencent; however, this difference did not exist for Qihoo. This difference reflected the strong attractiveness of population for the few immigration provinces. Furthermore, the variation of Qihoo was more irregular than that of the others. Two variation peaks of net travel flows existed. The first variation peak appeared approximately five days before the Spring Festival, and the second variation peak appeared at approximately six days after the Spring Festival. This observation was consistent with the overall travel trend analyzed earlier. However, the variation valley of Baidu appeared in the middle of Spring Festival holiday, whereas the variation valleys of Tencent and Qihoo appeared on the Spring Festival.



For Baidu, the second variation peak of Guangdong Province, which was apparently delayed, appeared at approximately 12 days after the Spring Festival.

The four-quadrant diagram of net travel flows could easily identify the function and attractive strength of provinces. In general, the data points of Baidu and Tencent could significantly fit a straight line. The fitting line slope of Tencent was close to  $-45^\circ$ , whereas that of Baidu was obviously small (Figure 4d–f). This observation indicated that the travel flow strength of Tencent before and after the Spring Festival was nearly the same, whereas the travel flow strength of Baidu before the Spring Festival was stronger than those after the Spring Festival. The former was identical with theoretical expectation, but the latter was not. The points of Qihoo were discrete, and no obvious regularities existed. Furthermore, the immigration provinces could be divided into two categories. The first category was the eastern coastal developed provinces characterized by their manufacturing and service-oriented industries. These provinces included Guangdong, Beijing, Shanghai, Zhejiang, and Jiangsu. The second category was the central and western underdeveloped provinces characterized by their labor-intensive industries, such as agriculture and mining. These provinces included Xinjiang, Qinghai, and Inner Mongolia. The majority of the emigration provinces were populous and were developing provinces located in the central and western China, such as Anhui, Henan, Sichuan, Jiangxi, and Hubei. The three travel datasets had a high consistency with the census migration data; however, the travel data of Qihoo had a larger difference with the census data than the other two data. Spearman correlation coefficients among the travel and migration data of the census, Baidu, and Tencent were all above 0.95, but the correlation coefficients of Qihoo with the others were only approximately 0.85 (Table 3). This observation suggested that some drawbacks might exist for Qihoo travel data.



**Figure 4.** Net travel flows of Spring Festival. (a–c) Time series of net travel flows for Baidu, Tencent, and Qihoo, respectively. Different color lines represent net travel flows of 31 provinces, respectively; (d–f) Four-quadrant diagrams of net travel flows for Baidu, Tencent, and Qihoo, respectively. Each blue point represents a province.

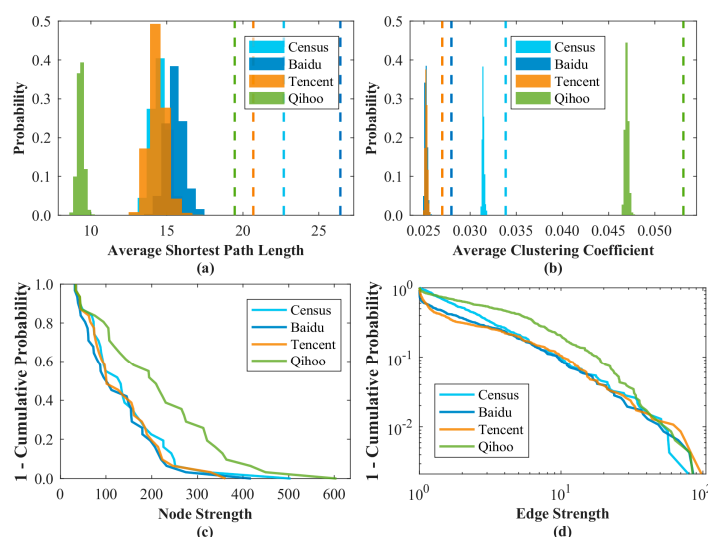
As analyzed earlier, Tencent travel data had a good performance in the overall trend, theoretical expectation, and the consistency with the census data. Baidu travel data had a deficiency, that is, the travel flows of Baidu before the Spring Festival were stronger than those after the Spring Festival. This observation did not meet the theoretical expectation. Qihoo travel data and the others had large differences which might be caused by acquiring railway passenger flow data only and by data incompleteness.

**Table 3.** Province rankings of Spring Festival immigration and emigration strength.

Ranking	Immigration				Emigration			
	Census	Baidu	Tencent	Qihoo	Census	Baidu	Tencent	Qihoo
1	Guangdong	Guangdong	Guangdong	Guangdong	Anhui	Anhui	Anhui	Heilongjiang
2	Zhejiang	Zhejiang	Beijing	Beijing	Henan	Henan	Hunan	Hubei
3	Shanghai	Beijing	Shanghai	Zhejiang	Sichuan	Hunan	Henan	Sichuan
4	Beijing	Shanghai	Zhejiang	Shanghai	Hunan	Jiangxi	Guangxi	Henan
5	Jiangsu	Jiangsu	Jiangsu	Jiangsu	Jiangxi	Guangxi	Jiangxi	Chongqing
6	Tianjin	Tianjin	Tianjin	Xinjiang	Hubei	Hubei	Hubei	Jiangxi
7	Fujian	Fujian	Xinjiang	Tibet	Guangxi	Sichuan	Sichuan	Shanxi
8	Xinjiang	Xinjiang	Liaoning	—	Guizhou	Hebei	Hebei	Anhui
9	Liaoning	Qinghai	Tibet	—	Chongqing	Shandong	Shandong	Guangxi
10	Inner Mongolia	Ningxia	Qinghai	—	Hebei	Guizhou	Guizhou	Inner Mongolia

### 3.3. Travel Network Analysis

Small-world and scale-free analyses are the first necessary programs of a complex network analysis. The average clustering coefficients of the four travel and migration networks were all larger than those in the random situation, and the average shortest paths of the four networks were all longer than those in the random situation (Figure 5a,b). Moreover, the small-world-ness indexes of the census, Baidu, Tencent, and Qihoo networks were 0.69, 0.65, 0.74, and 0.54, respectively, which were all smaller than 1. Therefore, the four networks all did not have the small-world characteristic. The node strength distributions of the four networks did not follow the power law. This observation proved that the four networks did not have scale-free characteristics. However, their edge strength distributions significantly followed the power law (Figure 5c,d). Furthermore, either node or edge strength distribution of Qihoo network showed a deviation from that of the others. These analyses showed that the travel network at the provincial scale had two characteristics. First, the large average clustering coefficient indicated that the network had a significant community characteristic, and the long average shortest path indicated that the travel was not nationwide. Second, the no scale-free characteristic indicated that the travel might have a multicenter trend, whereas the scale-free characteristic of the edge strength indicated that some strong travel flows existed among some provinces, which might be the key paths that should be paid more attention.



**Figure 5.** Small-world and scale-free analysis of Spring Festival travel networks. (a) Average shortest paths of the networks of the census, Baidu, Tencent, and Qihoo; (b) Average clustering coefficients of the networks; (c) Node strength distributions of the networks; (d) Edge strength distributions of the networks.

Some differences in node strength existed among the four networks, but an overall regional differentiation law could be found. This law indicated that high node strength appeared in the eastern and central China and low node strength appeared in the west China (Figure 6a,d,g,j). In particular, the four networks showed that Guangdong Province was the only one with ultrahigh node strength. Other provinces with high node strength mainly consisted of the eastern coastal developed provinces, such as Beijing, Shanghai, Zhejiang, and Jiangsu, and the central and western populous provinces, such as Henan, Sichuan, and Anhui. Therefore, the economic and social development level and population size might be important factors affecting travel strength.

The edge betweenness of the four networks had a similar spatial distribution with their edge strength. The spatial distribution was characterized by the short-distance travel trend of Baidu and Tencent, the long-distance travel trend of Qihoo, and a combination migration trend of the census (Figure 6b,e,h,k). Furthermore, the travel path chains of high betweenness, included Guangzhou-Hubei-Henan-Beijing, Guangdong-Henan, and Sichuan-Shaanxi-Gansu, geographically corresponded to important traffic backbones, such as Beijing-Guangzhou and Longhai railway lines. These travel path chains should be paid more attention and should be given a high priority.

The average clustering coefficient can be used to infer whether a network has a clustering characteristic, whereas the community identification can further identify the specific clusters. The community identification results of the four networks, whose identification process did not consider spatial factors, showed a high spatial contiguity trend (Figure 6c,f,i,l). This finding indicated that the travel or migration network might be affected by the geographical effect. Furthermore, the geographic boundaries of communities also showed a high similarity; in particular, the boundaries of Baidu and Tencent were nearly the same. In general, the identified communities could be divided into north and south. The north could be further divided into two subcommunities, namely, northeast and northwest, and the south could be further divided into three subcommunities, namely, Yangtze River Delta, south, and southwest. These results were highly consistent with the geographic zoning of China used in general, thereby further validating the availability of big travel data. However, the northwest and eastern communities were identified as one community in the identification result of Qihoo. This observation further indicated the possible deficiency of Qihoo data.

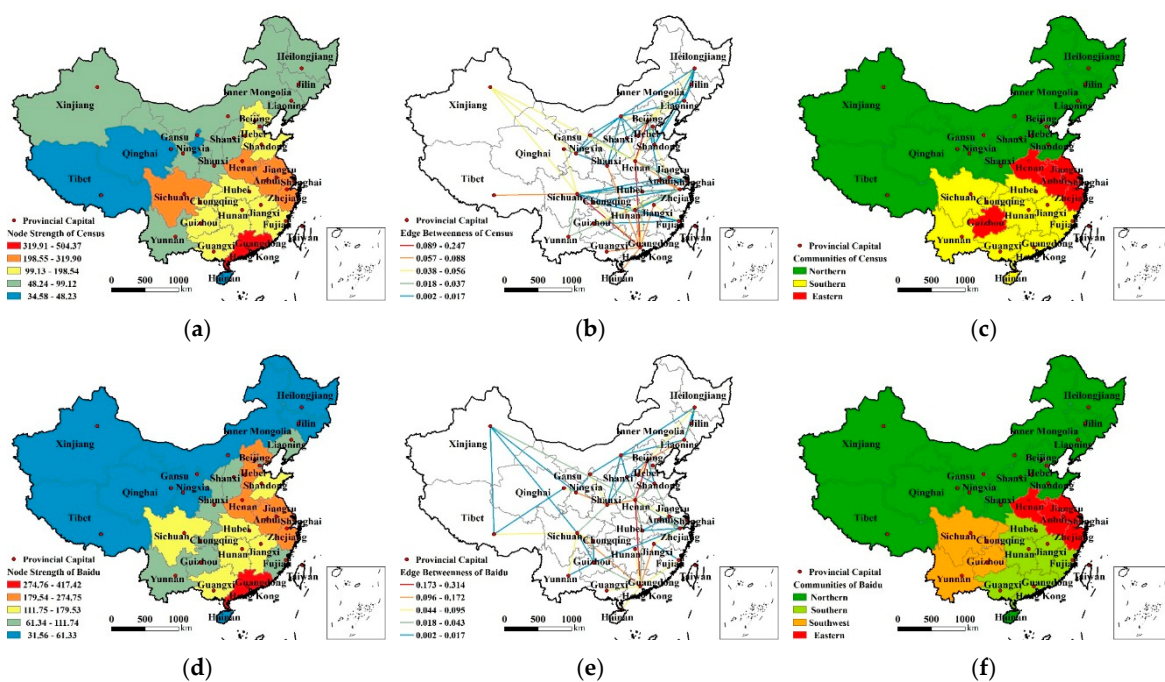
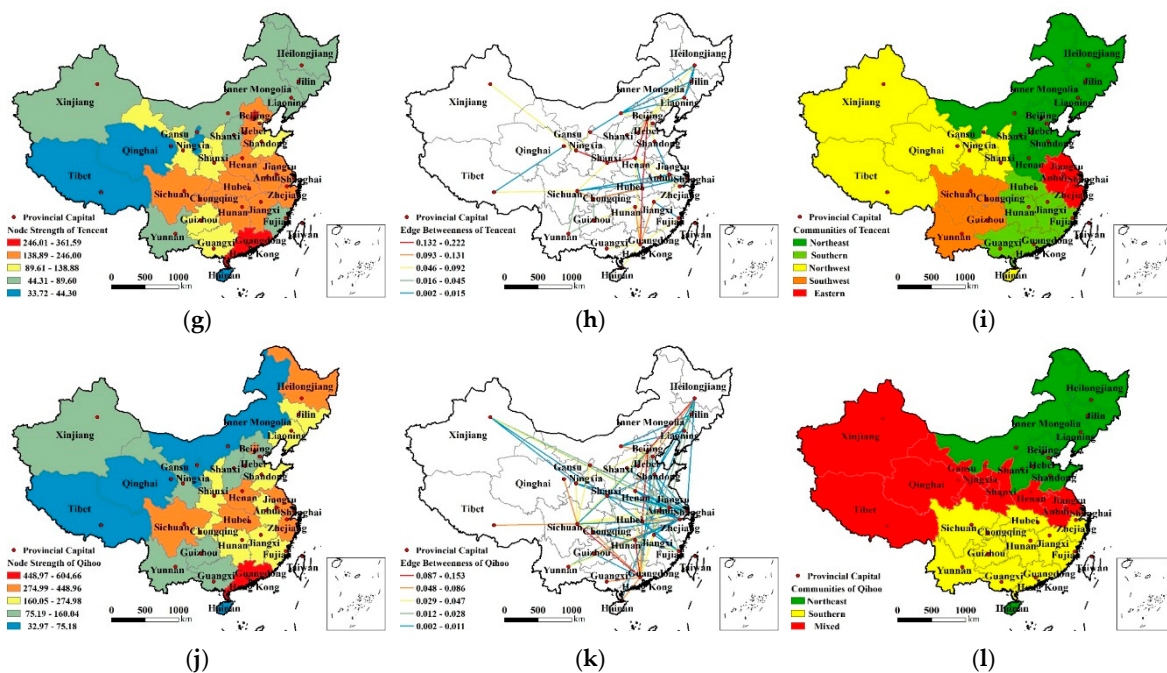


Figure 6. Cont.



**Figure 6.** Node strength, edge betweenness and community structures of Spring Festival travel networks. (a,d,g,j) Node strength of the networks for the census, Baidu, Tencent, and Qihoo, respectively; (b,e,h,k) Edge betweenness of the networks; (c,f,i,l) Community structures of the networks.

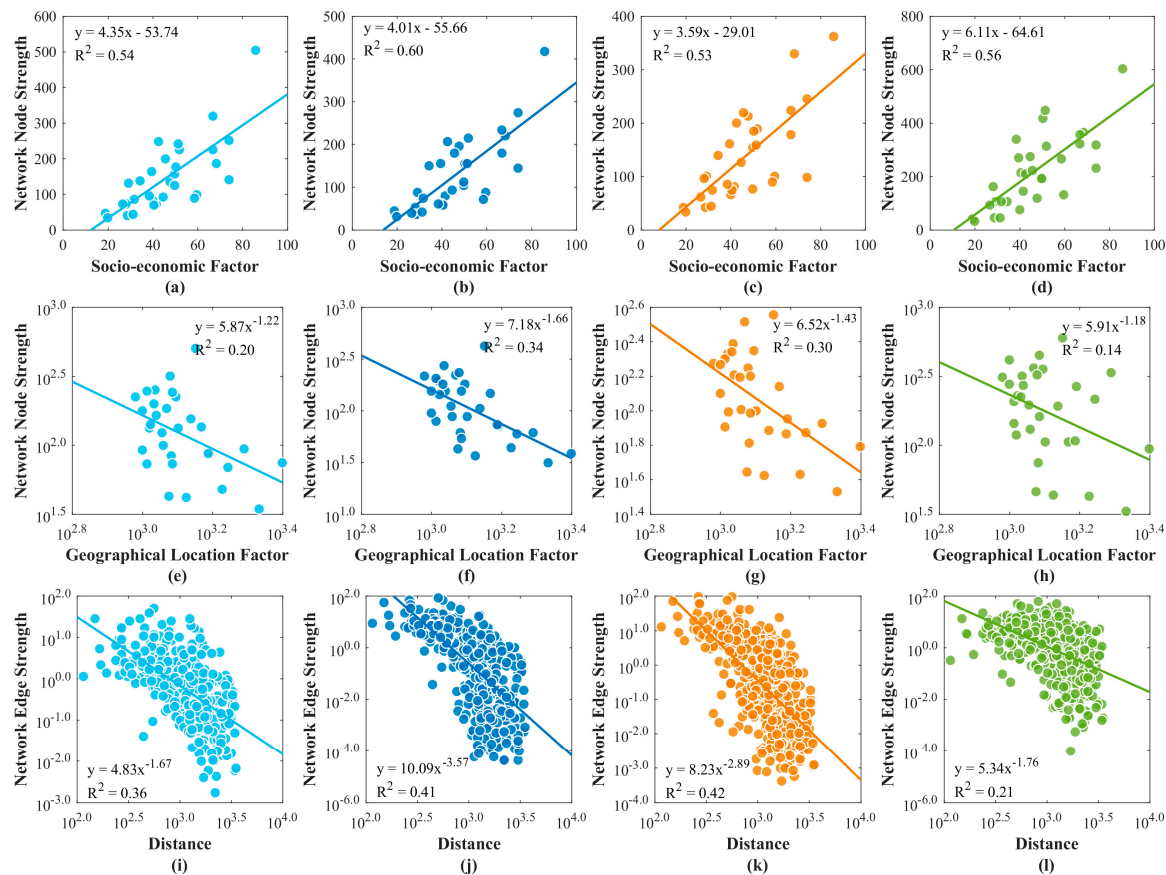
According to the preceding network analysis, the travel and migration networks of the census, Baidu, and Tencent have high consistencies with one another, whereas the travel network of Qihoo had a deviation with the others. The travel network showed a multicenter characteristic, that is, the eastern coastal developed provinces and the central and western populous provinces behaved as network centers, and a characteristic that some travel path chains played a leading role. Furthermore, the travel network presented a significant geographic clustering characteristic, which validated the geographic effect to the network.

#### 4. Discussion

As mentioned earlier, the economic and social development level, population size, and geographical location factor of the provinces may be important factors of travel or migration network [29]. Therefore, we further discuss the influences of socio-economic and geographic location factors on the network node strength. The socio-economic factor is measured using a composite index consisting of provinces' GDP, proportion of nonagricultural industry, population size, urbanization rate, and administrative level. The geographical location factor is measured using the average distance from one province to the others. Then, it can be seen that the network node strength of the four travel and migration data has a strong correlation with the socio-economic factor, but it has a weak correlation with the geographical location factor (Figure 7a–h). This finding indicates that the socio-economic factor has a stronger effect than the geographical location factor. However, the significant correlation between the network edge strength and its geographical distance proves that the geographical location factor remains an important factor to the travel or migration network (Figure 7i–l). Furthermore, the fitting degrees of the regression curves of Baidu and Tencent were higher than those of the census and Qihoo. This scenario may result from the long-distance trends of the census and Qihoo, as mentioned in the preceding section. We use the reverse gravity model to estimate the geographical distance resistance coefficients of the four travel and migration networks. The coefficients are 1.32, 3.55, 3.01,



and 1.02 for the census, Baidu, Tencent, and Qihoo, respectively [30,31]. This observation further validates our inference.



**Figure 7.** The influences of socio-economic and geographic factors on Spring Festival travel networks. (a–d) The correlations between the socio-economic factor and the network node strength for the networks; (e–h) The correlations between the geographic location factor and the network node strength for the networks; (i–l) The correlations between the network edge strength and its geographical distance for the networks. Four different color points, namely, cyan, blue, orange and green, presents the census, Baidu, Tencent and Qihoo, respectively.

In general, Spring Festival travel rush has some differences with long-term migration, as shown by the census. Spring Festival travel rush is more affected by the geographical location factor and has a narrower scope of travel than long-term migration. The most significant characteristic of Spring Festival travel rush is that the travel intensively surges in a short time and brings considerable traffic burden and various kinds of social problems. However, Spring Festival travel rush has similar macro driving factors with long-term migration, that is, unbalanced urban-rural development and unbalanced regional development. These two types of unbalanced development result from dual development policies of urban and rural areas in the age of planned economy and from the preferential development of the eastern region after reform and opening-up policies [32]. Hence, large-scale rural surplus labor migrates to the eastern coastal developed provinces and urbanization area because of economic interests. However, the incompleteness of institutional, economic, and social factors, such as hukou, housing, and social insurance, and the constraint of population-carrying capacity of large cities impede the settlement of migrants [33–35]. Hence, an abnormal travel rush occurs around Spring Festival holiday and brings various kinds of social problems. In the future, with the urban-rural and regional balanced development, migrants, particularly rural migrant workers, will settle down in the cities where they work. Reasonable travel due to the tradition of family reunions on Spring Festival will



continue to exist, but abnormal travel caused by the unbalanced development will gradually decrease, and problems brought by Spring Festival travel rush will be effectively improved. Therefore, the government proposed some development strategies, including differentiated regional development, urban-rural integration development, and new-type urbanization development. However, these strategies cannot be achieved within a short time, and Spring Festival travel rush will still last for years.

Fortunately, the modern high-speed transport system characterized by expressways and high-speed rails and the modern information and communication technology characterized by the Internet and the mobile communication networks have achieved a great development and popularization. This development brings a favorable turn to the problem resolution of Spring Festival travel rush. The expressway network and the high-speed rail network of China have both become the largest ones in the world until 2014, and these networks can greatly enhance the passenger capacity during the Spring Festival travel season. The scales of the mobile communication network and mobile users in China are also the largest in the world. On the one hand, mobile big data technologies derived from the mobile communication network can obtain and analyze the Spring Festival travel status in real time. For example, the government can obtain and predict the travel rush trend from the time dimension, and identify hot routes and hot provinces and cities from a spatial dimension under the support of mobile big data at any time. Then, transport resources can be allocated and relocated by the transportation department reasonably. An early warning mechanism can be established to publish related information to the public through TV, mobile phones, and other ways in real time to guide the public expectations. On the other hand, the travelers can obtain real-time traffic early warning and traffic conditions through their mobile phones. Scientific and reasonable advice on travel schedules and routes can also be given to the travelers by mobile navigation applications based on mobile big data. All of these measures are helpful to alleviate problems brought by Spring Festival travel rush. However, as analyzed in this research, different big travel data have their own different characteristics and application scopes, and problems, such as different time and spatial resolution standards and inadequate details, exist. Therefore, we suggest that the government should integrate mobile big data and the official authority data of the transportation department, and a unified real-time traffic platform should be established for Spring Festival travel rush. This unified platform can assist the government to allocate passenger transport resources reasonably, and can also provide the public with a scientific and accurate travel guidance system to facilitate transport pressure relief.

## 5. Conclusions

This study investigated the spatial-temporal characteristics of Spring Festival travel rush in China through time series analysis and complex network analysis based on multisource big travel data derived from Baidu, Tencent, and Qihoo. Moreover, a method based on net travel flows was proposed to identify the function and population attractiveness of the provinces in the travel network. The main results are as follows: First, the travel data of Baidu and Tencent obtained from LBS had similar characteristics, and both of them were consistent with the census data used as a reference. This observation could validate the accuracy and scientificity of these two travel datasets. Qihoo travel data presented some deficiencies, such as a long-distance travel trend and data instability. These deficiencies might be caused by acquiring railway passenger flow data only and by data incompleteness. Second, two travel peaks appeared during Spring Festival travel season. One was approximately five days before the Spring Festival, and the other was approximately six days after the Spring Festival. The travel valley appeared on the Spring Festival. The complex network analysis indicated the Spring Festival travel network did not have small-world and scale-free characteristics at the provincial scale. The travel network showed a multicenter characteristic, that is, the eastern coastal developed provinces and the central and western populous provinces behaved as network centers, and a characteristic that some travel path chains played a leading role. Furthermore, the travel network showed a significant geographic clustering characteristic that was highly consistent with the geographic zoning of China used in general. Third, economic and social factors, such as economic

and social development level and population size, had more influence on the travel network than the geographical location factor. Furthermore, the problem of Spring Festival travel rush will not be effectively improved in a short time because of unbalanced urban-rural development and unbalanced regional development. However, the great development and popularization of the modern high-speed transport system and modern information and communication technology brings a favorable turn to the problem resolution of Spring Festival travel rush. Therefore, a unified real-time traffic platform for Spring Festival travel rush should be established through the government's integration of mobile big data and the official authority data of the transportation department. This platform can help the government and the public to alleviate the problems caused by Spring Festival travel rush.

**Supplementary Materials:** The following are available online at [www.mdpi.com/2071-1050/8/11/1184/s1](http://www.mdpi.com/2071-1050/8/11/1184/s1).

**Acknowledgments:** This research was funded by the National Science & Technology Support Program of China (No. 2012BAJ22B02).

**Author Contributions:** Jiwei Li and Qingqing Ye conceived and designed the experiments; Xuankai Deng performed the experiments; Yaolin Liu and Yanfang Liu analyzed the data; Jiwei Li contributed analysis tools; Jiwei Li wrote the paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Duan, C.; Zhu, B.; Cui, C.; Chen, Y. Spring Festival Travel Rush and Floating Population. *Popul. Res.* **2009**, 30–45.
2. Dong, C.; Chen, C. A Sociological Interpretation of the “Spring Festival” Transportation Problem. *Popul. J.* **2008**, 30, 31–34.
3. Tracing the Largest Seasonal Migration on Earth. Available online: <http://arxiv.org/abs/1411.0983> (accessed on 13 November 2016).
4. Which Is the Better Migration Map of Chinese Spring Festival: Reviews Based on the Data Perspective. Available online: [http://www.thepaper.cn/newsDetail\\_forward\\_1321590](http://www.thepaper.cn/newsDetail_forward_1321590) (accessed on 1 May 2016).
5. Zhengwei, S.; Lun, W.; Yu, L. Study on Interactive Network among Chinese Cities Based on the Check-in Dataset. *Geogr. Geo-Inf. Sci.* **2013**, 29, 1–5. [[CrossRef](#)]
6. Liu, Y.; Sui, Z.; Kang, C.; Gao, Y. Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data. *PLoS ONE* **2014**, 9, e86026. [[CrossRef](#)] [[PubMed](#)]
7. Hilbert, M. Big Data for Development: A Review of Promises and Challenges. *Dev. Policy Rev.* **2016**, 34, 135–174. [[CrossRef](#)]
8. Mayer-Schönberger, V.; Cukier, K. *Big Data: A Revolution that Will Transform How We Live, Work, and Think*; Houghton Mifflin Harcourt: Boston, MA, USA, 2013.
9. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Ann. Assoc. Am. Geogr.* **2015**, 105, 512–530. [[CrossRef](#)]
10. Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D. Complex networks: Structure and dynamics. *Phys. Rep.* **2006**, 424, 175–308. [[CrossRef](#)]
11. Newman, M.E.J. The structure and function of complex networks. *Siam. Rev.* **2003**, 167–256. [[CrossRef](#)]
12. Albert, R.; Barabási, A.L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, 74, 47. [[CrossRef](#)]
13. Barthélemy, M. Spatial networks. *Phys. Rep.* **2011**, 499, 1–101. [[CrossRef](#)]
14. Bullock, S.; Barnett, L.; Di Paolo, E.A. Spatial Embedding and the Structure of Complex Networks. *Complexity* **2010**, 16, 20–28. [[CrossRef](#)]
15. Antoniou, I.E.; Tsompa, E.T. Statistical Analysis of Weighted Networks. *Discrete Dyn. Nat. Soc.* **2008**, 2008, 375452. [[CrossRef](#)]
16. Li, W.; Lin, Y.; Liu, Y. The structure of weighted small-world networks. *Phys. Stat. Mech. Appl.* **2007**, 376, 708–718. [[CrossRef](#)]
17. Barrat, A.; Barthélemy, M.; Pastor-Satorras, R.; Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **2004**, 101, 3747–3752. [[CrossRef](#)] [[PubMed](#)]
18. Saramäki, J.; Kivelä, M.; Onnela, J.; Kaski, K.; Kertész, J. Generalizations of the clustering coefficient to weighted complex networks. *Phys. Rev.* **2007**, 75, 27105. [[CrossRef](#)] [[PubMed](#)]

19. Onnela, J.; Saramäki, J.; Kertész, J.; Kaski, K. Intensity and coherence of motifs in weighted complex networks. *Phys. Rev.* **2005**, *71*, 65103. [[CrossRef](#)] [[PubMed](#)]
20. Humphries, M.D.; Gurney, K. Network ‘Small-World-Ness’: A Quantitative Method for Determining Canonical Network Equivalence. *PLoS ONE* **2008**, *3*, e2051. [[CrossRef](#)] [[PubMed](#)]
21. Humphries, M.D.; Gurney, K.; Prescott, T.J. The brainstem reticular formation is a small-world, not scale-free, network. *Proc. R. Soc. B* **2006**, *273*, 503–511. [[CrossRef](#)] [[PubMed](#)]
22. Freeman, L.C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1979**, *1*, 215–239. [[CrossRef](#)]
23. Newman, M.E.J. Communities, modules and large-scale structure in networks. *Nat. Phys.* **2011**, *8*, 25–31. [[CrossRef](#)]
24. Fortunato, S. Community detection in graphs. *Phys. Rep.* **2010**, *486*, 75–174. [[CrossRef](#)]
25. Lancichinetti, A.; Fortunato, S. Community detection algorithms: a comparative analysis. *Phys. Rev. Stat. Nonlinear Soft Matter Phys.* **2009**, *80*, 56117. [[CrossRef](#)] [[PubMed](#)]
26. Danon, L.; Diaz-Guilera, A.; Duch, J.; Arenas, A. Comparing community structure identification. *J. Stat. Mech.-Theory E* **2005**. [[CrossRef](#)]
27. Blondel, V.D.; Guillaume, J.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [[CrossRef](#)]
28. Krackardt, D. QAP partialling as a test of spuriousness. *Soc. Netw.* **1987**, *9*, 171–186. [[CrossRef](#)]
29. Zhang, K.H.; Song, S. Rural-urban migration and urbanization in China: Evidence from time-series and cross-section analyses. *Chin. Econ. Rev.* **2003**, *14*, 386–400. [[CrossRef](#)]
30. Song, W. Nodal attractions in China’s intercity air passenger transportation. In *Papers and Proceedings of the Applied Geography Conferences*; Harrington, L.M.B., Harrington, J.A., Eds.; Applied Geography Conferences Inc.: Denton, TX, USA, 2006; Volume 29, pp. 443–452.
31. Shen, G. Reverse-fitting the gravity model to inter-city airline passenger flows by an algebraic simplification. *J. Trans. Geogr.* **2004**, *12*, 219–234. [[CrossRef](#)]
32. Shen, J. Increasing internal migration in China from 1985 to 2005: Institutional versus economic drivers. *Habitat Int.* **2013**, *39*, 1–7. [[CrossRef](#)]
33. Long, H.; Zou, J.; Pykett, J.; Li, Y. Analysis of rural transformation development in China since the turn of the new millennium. *Appl. Geogr.* **2011**, *31*, 1094–1105. [[CrossRef](#)]
34. He, C.; Chen, T.; Mao, X.; Zhou, Y. Economic transition, urbanization and population redistribution in China. *Habitat Int.* **2016**, *51*, 39–47. [[CrossRef](#)]
35. Li, Y.; Liu, H.; Tang, Q.; Lu, D.; Xiao, N. Spatial-temporal patterns of China’s interprovincial migration, 1985–2010. *J. Geogr. Sci.* **2014**, *24*, 907–923. [[CrossRef](#)]



© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).