Text-based Analysis of the Geographic Spread of the COVID-19 Pandemic in China

Jiaobei Wang
Northwest University https://orcid.org/0000-0002-5316-694X
Gang Li (liglzu@gmail.com)
Northwest University
Annan Jin
Northwest University
Tingting Xu
Northwest University
Qifan Nie
The University of Alabama

Research Article

Keywords: COVID-19, Spatial diffusion, Text analysis, outside Hubei in China

Posted Date: March 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-258739/v1

Abstract

Background: The COVID-19 pandemic spread rapidly due to the mass migration during the Spring Festival in China, bringing a significant public health challenge. This study aims to explore the pattern of COVID-19 diffusion related to the Spring festival and provide some suggestions on controlling the epidemic.

Methods: We included 10316 cases from provincial or city Health Commission websites outside Hubei in China from December 30, 2019 to February 27, 2020. Data on the gender, age, occupation, inflow and outflow places were extracted from confirmed cases' detailed records.

Results: The number of confirmed cases who worked in commercial service was the largest, accounting for nearly 55%, followed by sales service (22%) and transport service (13.6%). "Wuhan", "close contact", "quarantine" (15.31) and "fever" (13.43) were the highest-linked centers within the text analysis. Most of the infected are relatives of confirmed cases and people mainly choose the self-driving transportation mode to go to hospitals. The clinical symptoms' expression varies among different age groups. Over time, infected areas and generations have experienced transitions from the nonlocal first to the local third group, and from more males infected to females. The source of the epidemic spreads to the south, then to the north, and finally evolves to the southeast, and the spread direction is from south to north afterward to west, over time. Wuhan and Shenzhen had the biggest influence on the flow as essential points of the flow network. The principal path often originated in Hubei or provincial capital cities and flowed to those cities which are densely populated and economically developed. “Returning home to visit relatives”, “migrant workers returning home” and “returning home after traveling” are the chief migration reasons. Compared with the long-term flows, short-term paths are of shorter distance. About half of confirmed cases migrated with their families. Finally, the flows are always opposite to the gradient mobility pattern of China's labor.

Conclusion: Based on the background of the reunion tradition in the Spring Festival, the diffusion of COVID-19 outside Hubei in China presented a pattern tied to person migration to return home. At the macroscopic scale, the pattern shows a linear relocation diffusion from vertical to horizontal. On the microscopic scale, radiation contact diffusion with transit from the nonlocal first to the local third group and male-dominated cases in the early stage to mostly female cases in the later stage were the apparent trend.

Introduction

The COVID-19 pandemic has brought great challenges to the world's public health and has spread rapidly worldwide. As of April 11, 2020, outbreaks have been spread in more than 210 countries and regions outside of China. The number of cumulative confirmed cases that have been reported has exceeded 10,000 in more than 18 countries and territories around the world, and it is known that the virus being transmitted from person to person and is extremely contagious. Large-scale human migration can

amplify localized disease into widespread epidemics to a certain extent[1]. In particular, there is a possibility of COVID-19 transmission between people during the asymptomatic incubation period, making it more difficult to control the spread of COVID-19[2]. Annual Chunyun mass migration in China can involve as many as three billion trips[3] and overlaps with the Chinese Lunar New Year holiday, the start of which coincided with the emergence of COVID-19 in 2020. Therefore, tracing the population moving on the eve of the Spring Festival was one of the most valuable ways to control the epidemic in China.

The earliest cases in China have led scholars to identify the Wuhan Huanan Seafood Market as the location from which the virus likely originated[4, 5]. Considering Wuhan's central position in China's railway and aviation networks, and the severity of COVID-19, the government decided to lock down Wuhan on Jan 23, 2020, to avoid the spread of COVID-19. Even so, about 5 million people emigrated from Wuhan before starting the travel ban[6]. The potential scale of the outbreak is very worrying. Many scholars use the basic regeneration number R_0, which is one of the classic indicators to measure the ability of pandemic transmission in infectious diseases, to construct SEIR[7, 8], SIR[9], EIR[10], and other models for anticipating the trend of the pandemic in the future. Nevertheless, the transmission of COVID-19 can be interrupted by timely detection, quarantine, treatment, and reliable traceability[11]. In order to control the spread of COVID-19, WHO and the US Centers for Disease Control and Prevention (CDC) proposed recommendations, including avoiding travel to high-risk areas and contact with symptomatic individuals [12, 13]. In China, celebrations related to China's annual Lunar New Year holiday were canceled and measures like Wuhan lockdown were implemented [14]. Gradually, China's R_0 has dropped to a low level, but the pandemic continues to spread globally. Scholars have turned their research to the most severely affected areas in the world, including Italy[15], Germany[16], and the United States[17]. Active public health interventions (early case detection, contact tracing, etc.) have played a massive role in Chinese's containment of this pandemic, an experience that will be closely focused on and analyzed by the world[18]. In contrast, countries such as the United States and others missed certain initial opportunities to implement measures to control the pandemic. The battle with COVID-19 may continue[19].

Analysis of different population and spatio-temporal spread models in China is a significant basis for the current prevention and control of the pandemic and response to similar future challenges. We conduct our research in three aspects by using the keywords extracted from detailed information to increase awareness and gradually restore normal life. The first aspect is demographic characteristics of confirmed cases, including cases' occupation, clinical symptoms and some composite characteristics. The second component is the temporal evolution period divided by the sex ratio and the infected areas and orders. Finally, the spatial pattern of diffusion in different scales, such as provincial and city-level, is observed.

Data And Method

Data description

This study's data are the details of confirmed cases collected by manual interpretation from provincial or city Health Commissions outside Hubei in China. 10,316 pieces of data involved in infection history and temporal trajectory of confirmed cases were obtained and analyzed. The average age of confirmed cases is 44.8 years old, and the median age is 45; the oldest confirmed case is 97 years old. In addition, the current youngest confirmed case in China was 5 days old, indicating there is a possibility of vertical maternal transmission in COVID-19. These results are consistent with the reports from the Health Commission, indicating typical and representative data. Among these, 9717 and 10165 pieces of data are obtained, respectively, containing gender and age records. In general, the male infected rate is slightly higher than that for females, and the average age and median age of females infected are 2 years older than that of males. The data are further divided by population into 5 age groups, which are 1–19 years old, 20–39 years old, 40–59 years old, 60–79 years old and ≥ 80 years old. The findings indicate that confirmed cases under 60 years old are mainly males, while people aged 60 and over are mainly females.

Table 1
The basic description of collected data

		N	Range	Average	Median	Sex ratio(%)
Total		10316	96.8	44.8	45	/
Gender	Male	5079	93	44.1	44	/
	Female	4638	97	45.7	46	/
Age	0–19	487	/	/	/	123.6
	20–39	3019	/	/	/	118.8
	40–59	4945	/	/	/	111
	60–79	1557	/	/	/	92.7
	≥ 80	157	/	/	/	85.7

Population definitions

Occupations: Combining the 10 vocational divisions of the China Association of Occupational Planners and the comprehensive information from the Health Commission website, we divided the 1,444 confirmed cases with full records of occupations in the information into 11 categories: waiter, staff, worker, retiree, farming, student, producer, politician, educator, homeless, salesman and freelancer. According to patients' workplaces, we also divide the records of confirmed cases into nonlocal and local. Among these, people who work nonlocally are outside the long-term residential city or domicile city, and the local's workplace is conversely. According to the International Service Trade Classification table, the service personnel is further divided into eight service types: transport service, commercial service, sales service, health and

social service, construction and related engineering service, tourism-related service, and telecommunication service.

Infected areas: Depending on whether the confirmed cases have a flow experience, they are divided into nonlocal infections and local infections. We found that the majority of people infected in nonlocal areas are primarily males and mostly aged between 20 to 59. Conversely, females account for a larger proportion of local infected people.

Table 2
Place attributes of confirmed COVID-19 cases outside Hubei in China

Age	Nonlocal		Local	
	Male(56%)	Female(44%)	Male (49%)	Female(51%)
0–19	109	92	112	87
20–39	822	591	507	498
40–59	807	628	737	759
60–79	256	238	311	378
≥ 80	10	11	46	56

Infected rank: As there is no detailed information about the infection course and activity tracking of the confirmed cases within Wuhan or Hubei at present, this paper regards the whole of Hubei as the "black box origin area" of the pandemic in China. The confirmed cases were roughly divided into three infected ranks by manually interpreting the infection history and activity trajectory (5554 cases) of the confirmed cases: confirmed cases living and working in Hubei for a long time are defined as the first group of infection, those who came into direct contact with the first group of infection or went to Hubei for a short period before or during the outbreak are defined as the second group of infection, and those who were not present in Hubei and could not be identified as the second group of infection are defined as the third group or above. In terms of age, cases are mainly aged 40–59 for each infection group, with the 20–39 age range being second for each infection group. In terms of gender, males in the first and second group are more prevalent while; conversely, the third group infections are majority females. There are apparent differences between infections in different stages according to the different people's abilities and mobility.

Table 3
Temporal series of confirmed COVID-19 cases outside Hubei in China

Age	1		2		3	
	Male (55%)	Female (45%)	Male (56%)	Female (44%)	Male (48%)	Female (52%)
0–19	60	51	64	46	89	74
20–39	400	295	434	339	401	411
40–59	387	320	468	368	554	607
60–79	148	152	137	123	252	286
≥ 80	4	3	11	12	38	45

Population with different flow purposes: The confirmed cases were divided into 3 major categories and 6 sub-categories according to their flow duration and purpose. Short-term flow includes business trips, gatherings, and tourism, while the long-term flow is divided into returning home to visit relatives, migrant workers return home, and students return home on holiday. Other flows and some paths have no reason records.

Analysis method

Social network analysis: It is assumed the node's importance is equivalent to its connection with other nodes. The province is regarded as a node in the network, using the migration path as the directed edge weighted by case quantity. These indicators were used to explore the properties of the flow network. In our study, we use degree centrality to evaluate the ability of a node's direct contact relationship[20]. Node betweenness is used to identify the control degree of the relationship[21, 22]. Edge betweenness is used to measure an edge's importance by counting the share of flows on the shortest paths that traverse that edge[23, 24].

Results

Composite characteristics

Inflow and outflow of population promoted the spread of the epidemic, and the type of occupation largely influences the population's mobility. In terms of vocation, workers, waiters, staff and students are most infected, accounting for nearly 80% of confirmed cases. Additionally, the highest proportion of nonlocal cases comes from the worker occupation, followed by waiters and students. For local occupations, the staff is the mainstay, followed by workers and waiters (Fig. 1A). The number of confirmed cases who worked in commercial service was the largest, accounting for nearly 55%, followed by sales service (22%)

and transport service (13.6%) (Fig. 1B). More people are generally infected with COVID-19 in nonlocal industries, which have high exposure and frequent contact with others, such as workers, staff, and servicers, especially business servicers. This servicer typically works in shopping malls or supermarkets, which are the main infection hubs of COVID-19. As a large group returning home during the winter vacation, students also demonstrate a very high risk of infection.

Furthermore, keywords are extracted from the manually collected text of 8,911 confirmed cases with detailed information, removing "patient", "diagnosis", gender, origin, and other irrelevant messages. Jieba is used to generate the top 100 keywords and word frequency after semantic combination screened relevant feature words, permitting the construction of the co-occurrence matrix of high-frequency words. Where the frequency is 0, there is no co-occurrence relation between such keywords. Gephi is used to describe the degree of dependence between these keywords, and the Force Atlas2 algorithm is utilized to construct the co-occurrence map of keywords. The node represents a high-frequency word. The larger the node, the greater the centrality of the keyword (Fig. 2). Among them, the average degree of each node is 28.222; the average path length is 1.194, and the average clustering coefficient is 0.868, indicating a high degree of connection among the keywords; that is, there is a susceptible population. Among them, "Wuhan", "close contact", "quarantine" (15.31) and "fever" (13.43) were the highest centers; "afternoon" (12.85), "consult" (10.95), "people's hospital" (10.95), "ante meridiem" (10.64) followed.

Thus, it can be observed that most of the confirmed cases are geographically related to "Wuhan", most of them are "returning home" people or a "permanent resident" of a nonlocal area, and most of the infected are "relatives" and "husbands" within the family, especially small core families, commonly composed of "couples". Most of the infected outside the family are acquaintances of the confirmed cases, such as "friends" and "colleagues".

The infection route is mostly "close contact", and the symptoms are mainly "fever", "positive", and "cough", which is consistent with the epidemiological characteristics of COVID-19. The disease symptoms are closely related to the patient's age, gender, and overall health status[25]. Thus, the age of each confirmed case and its clinical symptoms are combined (based on 4,859 records), irrelevant words are removed, and synonyms merged to calculate the frequency of each symptom keyword (the proportion of them accounting for all symptoms in a specific age group), and, finally, the keywords with a frequency of more than 0.01 in each age group are selected and united (Fig. 3). The results show that fever and cough were the core symptoms in all ages. In addition, cases in those aged from 0 to 19 years old also have multiple other symptoms, including diarrhea, rhinorrhea and feebleness. Feebleness, headache and ache are common in patients aged from 20 to 59 years old. Patients over 60 years of age have similar symptoms to those between 20 and 39 years of age, but "headache" is relatively rare.

Most confirmed cases are "centralized" in "community", "street" and "supermarket", and are associated with activities in the "morning" or "afternoon" like having "dinner" or going "shopping". Relatively independent transportation modes such as "self-driving" are used, which reduces the risk of re-transmission compared with "bus" and "train". People who have a distinct "onset" process, always "walk"

to "people's hospital", "designated medical institutions" or "outpatient" for "consult", and "nucleic acid detection" and "medical quarantine observation" are the main means of diagnosis and treatment. A few patients without obvious symptoms are taking "living at home" and other relevant measures; COVID-19 is highly infectious but has a low fatality rate, so patients' state is mostly "stable". Designated hospitals are usually arranged in the "people's hospital" and other types of medical institutions.

Temporal evolution properties

The number of cumulative confirmed cases showed a peaking trend over time. The number of cases is high from January 26 to February 6, 2020, with the number of cases on February 2 being the largest. The sex ratio is stable but exhibits significant fluctuation at early and late stages. The number of nonlocal cases is higher from January 26 to February 6, while the peak period of local infections is from February 1 to February 11, the lag between them, at 7 days, is equal to a half incubation. Overall, infected areas and orders have experienced a transition from the nonlocal first to the nonlocal second group, and then the local infection group.

Based on the above analysis, the temporal evolution is divided into the following four stages.

. Sporadic appearance period (December 30, 2019 to January 17, 2020). The COVID-19 pandemic had just spread from Wuhan to other areas before January 17, 2020. Most infected people were in the incubation stage, so the number of confirmed cases in this period is small, less than 20, and most are nonlocal infections with almost no local infection. There is no significant difference in sex ratio, so it is not representative.

. Fluctuating rise period (January 18 to January 25, 2020). The sex ratio of confirmed cases fluctuates in different degrees during the fluctuating rise period and the fluctuating decline period, but the fluctuations in the fluctuating rise period are relatively greater. More males than females were confirmed except for the days before January 13 and January 17; From January 18 to February 5, there were many people infected nonlocally, and the first infection group before January 25 had the highest proportion.

. Stable high-risk period (January 26 to February 6, 2020). The sex ratio in this period remained stable at 1, and the confirmed cases were gender-balanced. From January 26 to January 30, the nonlocal second group's proportion gradually increased, and the transition from the nonlocal first to the second group began; from January 31 to February 5, the trend began to change from nonlocal to local infections.

. Fluctuating decline period (February 7 to February 27, 2020). The sex ratio was relatively small during this period. Except for February 17, the confirmed cases were mostly females. After February 6th, the local-dominated, especially the third-group-dominated situation appeared.

Overall, the spread of the COVID-19 epidemic in China has experienced a transition from male-dominated cases in the early stage to mostly female cases in the later stage. At the same time, it has also transitioned from nonlocal to local infections. The specific manifestation is from the nonlocal first to the nonlocal second group, finally transitioning to the local third generation.

Spatial flow network

Source and target provinces. The movement of people dramatically affects the spread of the epidemic, so the flow path information of confirmed cases is extracted from detailed data (3770 cases), using the degree centrality to identify the main inflow and outflow provinces in different periods and to explore the spread patterns of the epidemic in China (Fig. 5). In general, interprovincial flow accounts for many reported cases, as much as 86.7%. Within-province flows occur mostly in Hubei (17.9%), Guangdong (12.5%) and Anhui (11.8%). Standardized by the total confirmed cases, the provinces with more than 50% intra-provincial flows in descending order are Shanxi, Hainan, Heilongjiang, Gansu, Anhui, Jilin, Guangxi and Sichuan (there is no data for Qinghai). These provinces are low-risk in this pandemic and mostly located in central, western, and northeast China. Interprovincial flow is concentrated in the provinces with more confirmed cases, like Hubei, Jiangxi, Zhejiang, etc. As a whole, the outflow areas are distributed in a "one point and one area" pattern, concentrated in the southeast, where the "one point" is Beijing (1.4%) and the "one area" is a sector-shaped area centered on Hubei, which accounts for 70%, followed by Guangdong (3.6%), Henan (2.3%), Zhejiang (2.0%), Jiangsu (1.8%), Hunan (1.8%) and Jiangxi (1.4%). The case outflow from these provinces (except for Hubei) made up 14.2%. The main inflow areas are presented as "one point and one line". The “one point” is Guangdong (18.8%), and the "one line" is Anhui (10.8%), Henan (7.5%), and Zhejiang (6.5%) in descending order. 43.7% of the cases were imported into these four provinces in total, demonstrating that the spread of the COVID-19 has prominent regional agglomeration characteristics. Guangdong, Henan, and Zhejiang are provinces with both high out-degree and indegree, requiring special attention.

In order to identify the diffusion process of COVID-19, the specific inflow and outflow provinces are observed in the different periods. Because of the limited cases counts, the situation in the sporadic appearance period is not discussed. The diffusion path during the fluctuating rise period is represented by "one source and two sinks". Among this, "one source" is Hubei and its surrounding Anhui and Zhejiang provinces, with Hubei as the main outflow province. The "two sinks" have begun to take shape, mostly consisting of north and south. The northern sink is primarily composed of two provinces near Hubei, Anhui, and Zhejiang. The south sink is an area centered on Guangdong, including Guangdong, Guangxi and Hainan. During the stable high-risk period, the diffusion path changed to "two sources and two sinks". Compared with the previous period, the other source is Guangdong, with Guangdong only in the south sink and the north sink adding the province of Henan. Finally, in the fluctuating decline period, the source gradually spreads from the core provinces to southeast China, which becomes the high-frequency outflow area while the sinks add Sichuan located in the southwest.

Therefore, it is concluded that the overall diffusion path is represented by the distribution of “one area and two sinks”, the “one area” being a southeast sector-shaped area with Hubei as the center, and the “two sinks” are the provinces of Guangdong and provinces north of Hubei separately. Hubei and Guangdong are the chief outflow and inflow provinces.

City-level properties. Key intermediary cities are also identified by calculating the node betweenness (that is, the number of the shortest paths between any pairs of nodes that pass through an intermediary node.), as presented in Fig. 6 and Table 4. Affected by this pandemic's widespread, nearly all the cities in China have been involved, but the node betweenness is generally low. Wuhan and Shenzhen had the most substantial influence on the flow, with node betweennesses at 19.9% and 10.4%, respectively. Except for the southeastern coastal cities like Shenzhen, Wenzhou, and Suzhou, and Hubei neighboring cities, like Xinyang, the remainder of the top 20 cities are all provincial capitals or municipalities.

The edge betweennesses (that is, the number of shortest paths between any pairs of nodes that run along a path) of the key paths are listed in Table 4. The top 20 paths controlled only 21% of the total number of paths. In general, the out-flow cities were mainly distributed southeast of the Hu line, which is consistent with China's population distribution characteristics. The source is cities usually in Hubei or provincial capital cities, and the destination is mostly cities radiated by a sector-shaped formed with Wuhan as the center of the circle (Table 4). For example, the main path often originated in Wuhan, Beijing, Nanchang, Changsha and other cities. However, the destinations are most densely populated and economically developed cities such as Wuhan's neighboring cities, and the cities located along the southeast coast.

Table 4
Key confirmed cases' paths with the highest edge betweennesses in the city-level informal

Ranking	Source	Target	Edge betweenness(%)
1	**Wuhan**	Xinyang	2.41%
2	**Wuhan**	Wenzhou	1.92%
3	**Wuhan**	**Nanning**	1.52%
4	Xiangyang	Shenzhen	1.32%
5	**Shanghai**	Suzhou	1.31%
6	**Wuhan**	**Nanchang**	1.18%
7	**Taiyuan**	Jincheng	1.14%
8	Huanggang	Anqing	1.02%
9	**Wuhan**	**Taiyuan**	1.01%
10	**Beijing**	**Zhengzhou**	1.01%
11	Shenzhen	**Haikou**	1.00%
12	**Changsha**	Fuyang	0.93%
13	Shenzhen	Huizhou	0.89%
14	**Zhengzhou**	Shangqiu	0.89%
15	**Changsha**	Shenzhen	0.53%
16	**Nanchang**	Maanshan	0.53%
17	**Shanghai**	Xianyang	0.53%
18	**Nanchang**	Zhanjiang	0.51%
19	**Wuhan**	Chuzhou	0.51%
20	**Haikou**	Suzhou	0.51%
In this study, cities in bold are the provincial capital cities.			

The three major categories are short-term (11.3%), long-term (45.1%), and other flows (43.6%). Excluding other flows, returning home for visiting relatives ranked first (30.4%), followed by migrant workers returning home (13.2%), and returning home after traveling (8.3%). A visual analysis of short-term and long-term mobility was conducted for returning home to visit relatives and migrant workers returning home. No separate explanation is provided for returning home after travel, but only an overall analysis of short-term flows (Fig. 7). Inherently, compared with long-term flows, short-term flows are mostly over in

short distances, mainly to tourism cities like Sanya, Kunming, and Guilin, some developed cities and some cities near with Wuhan while long-term flows scattered outwards with Wuhan as the center, always over long-distance and involving most cities in China. At the same time, we also found that about half of confirmed cases migrated with their families while traveling, returning home to visit relatives, or attending gatherings. The ratio of the number of cases who migrated with their families to the total number of confirmed cases in each category is 43.3%, 25% and 17%, respectively. The average number of migrants per household is 3, 2.8, and 4.4, respectively.

There are also significant differences in the flows of visiting relatives and migrant workers returning home. For example, flows for visiting relatives are concentrated in the east with Wuhan as the boundary. The first four paths are Wuhan-Shenzhen, Wuhan-Xinyang, Wuhan-Haikou, and Wuhan-Nanyang, while the flows for returning home from work are concentrated in the north of Wuhan. Wuhan-Yichun, Wuhan-Xinyang, Wuhan-Anqing, and Wuhan-Chongqing, the top four flows are all cities near Hubei. Flows of visiting relatives are farther than returning home from work; 55% and 66% of the paths are from provincial capital cities respectively, which opposes the Chinese population's flow pattern.

Discussion

In order to get results of high credibility, the statistical results of the official case details are compared with the information reported by the Health Commission and found to be consistent, laying the foundation for follow-up research. This article primarily uses text analysis methods to extract relevant information from detailed case records, such as gender, age, occupation, symptoms, flow paths, etc. It discusses the basic pattern of the spread of the epidemic from comprehensive data characteristics and the spread of the epidemic in geographic terms.

Compared with other studies, this study provides a more in-depth analysis of the text by combining word frequency analysis with other commonly used analysis methods in the current sociology field to conduct word segmentation processing on the official text and uses the social network to visualize the word segmentation results. These results summarize the relevant characteristics of the confirmed cases. The centrality of "Wuhan", "close contact" and "quarantine" was found to have the highest frequency. This result reveals that "close contact" is the major driver of most infections; moreover, confirmed cases were geographically related to "Wuhan". Family gatherings exacerbated the spread of the epidemic before and after the Spring Festival in China, especially the family-style round table culture, which has promoted large-scale gatherings. Infections within small core families were especially pronounced. China has actively adopted "quarantine" measures for suspected cases to halt the infection chain [26]. This operation reduces the probability of potential case infection and thereby reduces the R_0 rate. Meanwhile, the information of confirmed case occupations is extracted; industries with high exposure risk and close contact with others are susceptible, such as workers, servicers and staff. The infection risk of migrant workers and students is also high during traveling. Workers are the majority population infected locally; the risk of resuming work and returning to school still requires careful evaluation. Older individuals have a high fatality rate once infected due to more basic diseases and symptoms[27]. According to our results,

the manifestation of clinical symptoms varies among different age groups, although fever and cough are the most frequent clinical symptoms mentioned in many recent studies. The symptoms of patients under 20 are easier for self-detection. In contrast, older people have more subtle symptoms such as "headache" and "fatigue", which are not easy for self-detection. As a result, it is more common for elderly patients to delay seeking medical treatment. As a result, they have become one of the most affected groups during the epidemic. If necessary, wearing a mask as a critical preventative measure for the elderly and other at-risk populations can minimize the risk. In summary, people related to "Wuhan" who have confirmed cases among relatives, and those who work as workers, staff and waiters are more susceptible. In the history of contact, persons who have returned home, travelled, gathered, and had close contacts with confirmed cases are also susceptible.

According to the above analysis, the susceptibility of various demographic groups has been identified, and the infection ranks of confirmed cases have been divided into first, second, and third infections. Simultaneously, the confirmed cases were divided into nonlocal and local infections based on whether they exhibited migration. By this, it has been found that, affected by the traditional family livelihood structure with "the male-dominant outside and the female host inside", more males are infected nonlocally, but most local infections are females. Infected areas and ranks have experienced a transition from the nonlocal first to the local third group over time. While Wuhan city has experienced a remarkable positive effect with its lockdown, the effort is still needed to prevent and control the virus at a micro-scale within the city. Although the current increase of infected cases in China has been controlled, the risk of imports from abroad rises, significantly as the overseas viruses have mutated at present, making it increasingly important to guard the country[28].

The pandemic is closely related to human activities and behaviors, especially the migration of people, promoting the spread of the epidemic to a certain extent[29]. Studies have shown that analyzing population outflow distribution based on mobile phone data in Wuhan can accurately identify and predict high-risk areas at an early stage[30]. The pandemic severity in other regions was primarily affected by emigration from Wuhan[31]. The analysis of the movement trajectory of confirmed cases presented in this study also demonstrated that the flow of confirmed cases originated from Wuhan, and the degree centrality showed the spread of the COVID-19 epidemic has evident spatial agglomeration. On a provincial scale, intra-provincial flows mainly occur in low- and medium-risk cities, while inter-provincial flows are concentrated in the high-risk provinces such as the neighboring provinces of Hubei and the provinces located along the southeast coast, indicating that provinces with more vital mobility traits have higher infection rates. Moreover, COVID-19 mainly spread from Hubei to the north and south and subsequently to the west. Generally speaking, it spreads vertically, then horizontally. This finding runs counter to the direction of China's migrant population, which runs from east to west and from south to north. Guangdong, Henan, and Zhejiang are provinces with high values of out-degree and in-degree; these provinces are in a tense atmosphere of internal and external trouble, and not only must nonlocal virus carriers be prevented from further aggravating the local situation, but the control of intra-provincial movement must be strengthened, including recognition of the health code in some cities, control of some local gathering places, etc. In terms of urban centers, the epidemic has spread to most cities across the

country. Compared with SARS in 2003, the outbreak of COVID-19 is prevalent. This is closely related to the large-scale population movement brought by the Spring Festival. Wuhan and Shenzhen, as the control centers of the flow network, require careful attention. Therefore, the purpose and duration of the flow of confirmed cases were distinguished, and it is evident that different purposes and different travel durations significantly affect the distance of migration and the spread of the epidemic. The Chinese Lunar New Year holiday is the most celebratory time of the year in China, and a long vacation often occurs during this period. Especially for those who live far away from their hometown, the Spring Festival provides a suitable opportunity to return home. Compared with short-term trips, long-term homecoming behavior is less restricted by distance, which is evident in the traditional Chinese New Year reunion, where many people chose to return to their hometown with their families. Many studies demonstrate that the characteristics of COVID-19 in China represented familial aggregation[32], which is likely related to the people returning home to visit their relatives and attend family gatherings. In this study, it is found that the average number of migrants per household is 2–4, and the number in gathering reaches 4.4, indicating that the family gathering is one of the main ways of spreading the pandemic. Among these, the shift from east to west, south to north, and from provincial capitals or municipalities to other cities in visiting relatives and migrant workers returning home is completely consistent with China's current status of labor mobility.

Therefore, it is imperative for China to strictly control the passenger flow as Spring Festival travel's return peak is approaching. Whether it is possible to avoid the cases' re-emergence due to multiple scattered in this special time node is particularly important and will become a significant challenge for epidemic prevention and control. China has implemented some related measures, including advocating for everyone to celebrate the Spring Festival locally to lessen travel flow, provide the nucleic acid test certificate before returning home, and so on. However, due to frequent nucleic acid testing errors, certain risks remain, so vigilance is still required. When facing similar major festivals to the Chinese Spring Festival, such as New Year and Easter, it is crucial to learn from China's measures to control macro-to-micro flows.

Although this study has revealed good knowledge of COVID-19 diffusion and provided insights for its prevention and control, the following limitations still exist: the data of this research might have inconsistencies in statistical caliber and completeness across different provinces. Our analysis is missing Hubei province since its information of some cases has not been made public. In the future, an in-depth analysis of the epidemic outbreak area in China should be conducted, and the spatial and temporal accuracy of this national study can be improved beyond the 10,316 cases presented in this research.

Conclusion

This study summarized the diffusion of COVID-19 in China. Based on the background of the reunion tradition in the Spring Festival, the diffusion of COVID-19 outside Hubei in China presented a pattern tied to migration to return home. At the macroscopic scale, there was a linear relocation diffusion from vertical to horizontal. At the microscopic scale, radiation contact diffusion with transit from the nonlocal

first to the local third group and male-dominated cases in the early stage to mostly female cases in the later stage were prevalent, demonstrating China's effectiveness macroscopical lock-down of Wuhan and microscopic "quarantine" actions. Finally, people related to "Wuhan" have confirmed cases among relatives, and those working as workers, staff, and waiters are more susceptible. In the history of contact, those who have returned home, traveled, gathered, and had close contacts with confirmed cases are also susceptible and are the key population to protected.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed during the current study are available in the website of National Health Commission of the People's Republic of China. [http://www.nhc.gov.cn/]

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by grants from Key Project of the Special Guidance Fund for Emergency Study on the Prevention and Treatment of COVID-19 by Northwest University (Grant No. 2020), Tang Scholar Program of Northwest University (Grant No. 2016) and Characteristic & Advantage Research Team Construction Project of Human-Environment Relations and Space Security of Northwest University (Grant No. 2019).

Authors' contributions

GL designed the study. JBW wrote and revised the paper. ANJ and TTX helped analyze the data. QFN helped revised the paper. All authors read and approved the final manuscript.

Acknowledgements

We thank all the medical staff who participated in treating patients. Additionally, we thank all the patients enrolled in this study.

References

1. Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R: **Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges**. *International Journal of Antimicrobial Agents* 2020, **55**(3):105924.
2. Rothe C, Schunk M, Sothmann P, Bretzel G, Froeschl G, Wallrauch C, Zimmer T, Thiel V, Janke C, Guggemos W *et al*: **Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany**. *New England Journal of Medicine* 2020, **382**(10):970-971.
3. **Big data! The travel volume predictions during Lunar New Year holiday in 2020**. In. Edited by China MoTotPsRo. http://www.mot.gov.cn/fenxigongbao/yunlifenxi/202001/t20200109_3322161.html; 2020.
4. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y *et al*: **A new coronavirus associated with human respiratory disease in China**. *Nature* 2020, **580**(7803):E7-E7.
5. Wang C, Horby PW, Hayden FG, Gao GF: **A novel coronavirus outbreak of global health concern**. *Lancet* 2020, **395**(10223):470-473.
6. Chen Z-L, Zhang Q, Lu Y, Guo Z-M, Zhang X, Zhang W-J, Guo C, Liao C-H, Li Q-L, Han X-H *et al*: **Distribution of the COVID-19 epidemic and correlation with population emigration from Wuhan, China**. *Chinese Medical Journal* 2020, **133**(9):1044-1050.
7. Wan K, Chen J, Lu C, Dong L, Wu Z, Zhang L: **When will the battle against novel coronavirus end in Wuhan: A SEIR modeling analysis**. *Journal of Global Health* 2020, **10**(1).
8. Liu P-Y, He S, Rong L-B, Tang S-Y: **The effect of control measures on COVID-19 transmission in Italy: Comparison with Guangdong province in China**. *Infectious Diseases of Poverty* 2020, **9**(1).
9. Weissman GE, Crane-Droesch A, Chivers C, Mikkelsen ME, Halpern SD: **Locally Informed Simulation to Predict Hospital Capacity Needs During the COVID-19 Pandemic RESPONSE**. *Annals of Internal Medicine* 2020, **173**(8):680-681.
10. Xiong H, Yan H: **Simulating the Infected Population and Spread Trend of 2019-nCov Under Different Policy by EIR Model**. *Social Science Electronic Publishing* 2020.
11. **Novel Coronavirus(2019-nCoV): Situation Report - 10**. In. Edited by Organization WH. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200130-sitrep-10-ncov.pdf?sfvrsn=d0b2e480_2.; 2020.
12. **Novel Coronavirus (2019-nCoV) Advice for the Public**. In. Edited by Organization WH. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public; 2020.
13. **2019 Novel Coronavirus**. In. Edited by Prevention CfDCa. https://www.cdc.gov/coronavirus/2019-ncov/about/transmission.html.; 2020.
14. **Coronavirus in New York: Lunar New Year Events Canceled over Fears**. In: *The New York Times.* https://www.nytimes.com/2020/01/29/nyregion/coronavirusnyc.html; 2020.
15. Armocida B, Formenti B, Ussai S, Palestra F, Missoni E: **The Italian health system and the COVID-19 challenge**. *Lancet Public Health* 2020, **5**(5):E253-E253.

16. Wolfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Muller MA, Niemeyer D, Jones TC, Vollmar P, Rothe C *et al*: **Virological assessment of hospitalized patients with COVID-2019 (vol 581, pg 465, 2020)**. *Nature* 2020, **588**(7839):E35-E35.
17. Silverman JD, Hupert N, Washburne AD: **Using ILI surveillance to estimate state-specific case detection rates and forecast SARS-CoV-2 spread in the United States**. *medRxiv* 2020.
18. **Sustaining containment of COVID-19 in China**. *Lancet* 2020, **395**(10232):1230-1230.
19. **COVID-19 in the USA: a question of time**. *Lancet* 2020, **395**(10232):1229-1229.
20. Jun L: **Lectures on whole network approach: A practical guide to UCINET**. *Shanghai People's Publishing House* 2009.
21. Opsahl T, Agneessens F, Skvoretz J: **Node centrality in weighted networks: Generalizing degree and shortest paths**. *Social Networks* 2010, **32**(3):245-251.
22. Linton, C., Freeman: **Centrality in social networks conceptual clarification**. *Social Networks* 1978.
23. Girvan M, Newman MEJ: **Community structure in social and biological networks**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(12):7821-7826.
24. Bounova G, de Weck O: **Overview of metrics and their correlation patterns for multiple- metric topology analysis on heterogeneous graph ensembles**. *Physical Review E* 2012, **85**(1).
25. Gralinski LE, Menachery VD: **Return of the Coronavirus: 2019-nCoV**. *Viruses-Basel* 2020, **12**(2).
26. Bi Q, Wu Y, Mei S, Ye C, Zou X, Zhang Z, Liu X, Wei L, Truelove SA, Zhang T *et al*: **Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study**. *Lancet Infectious Diseases* 2020, **20**(8):911-919.
27. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y *et al*: **Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study**. *Lancet* 2020, **395**(10223):507-513.
28. Dawood AA: **Mutated COVID-19 may foretell a great risk for mankind in the future**. *New Microbes and New Infections* 2020, **35**.
29. Gang L, Jiaobei W, Tingting X, Xing G, Annan J, Yue Y: **Spatio-temporal evolution process and integrated measures for prevention and control of COVID-19 epidemic in China**. *Acta Geographica Sinica* 2020, **75**(11):2475-2489.
30. Jia JS, Lu X, Yuan Y, Xu G, Jia J, Christakis NA: **Population flow drives spatio-temporal distribution of COVID-19 in China**. *Nature* 2020.
31. Song W-Y, Zang P, Ding Z-X, Fang X-Y, Zhu L-G, Zhu Y, Bao C-J, Chen F, Wu M, Peng Z-H: **Massive migration promotes the early spread of COVID-19 in China: a study based on a scale-free network**. *Infectious Diseases of Poverty* 2020, **9**(1).
32. Jie L, Wanjun L, Zhihong D, Xiaojie W: **Clinical and Epidemiological Characteristics of 91Confirmed Children with COVID-19**. *Chinese Journal of Hospital Infectiology* 2020, **30**(11):1625-1629.

Figures

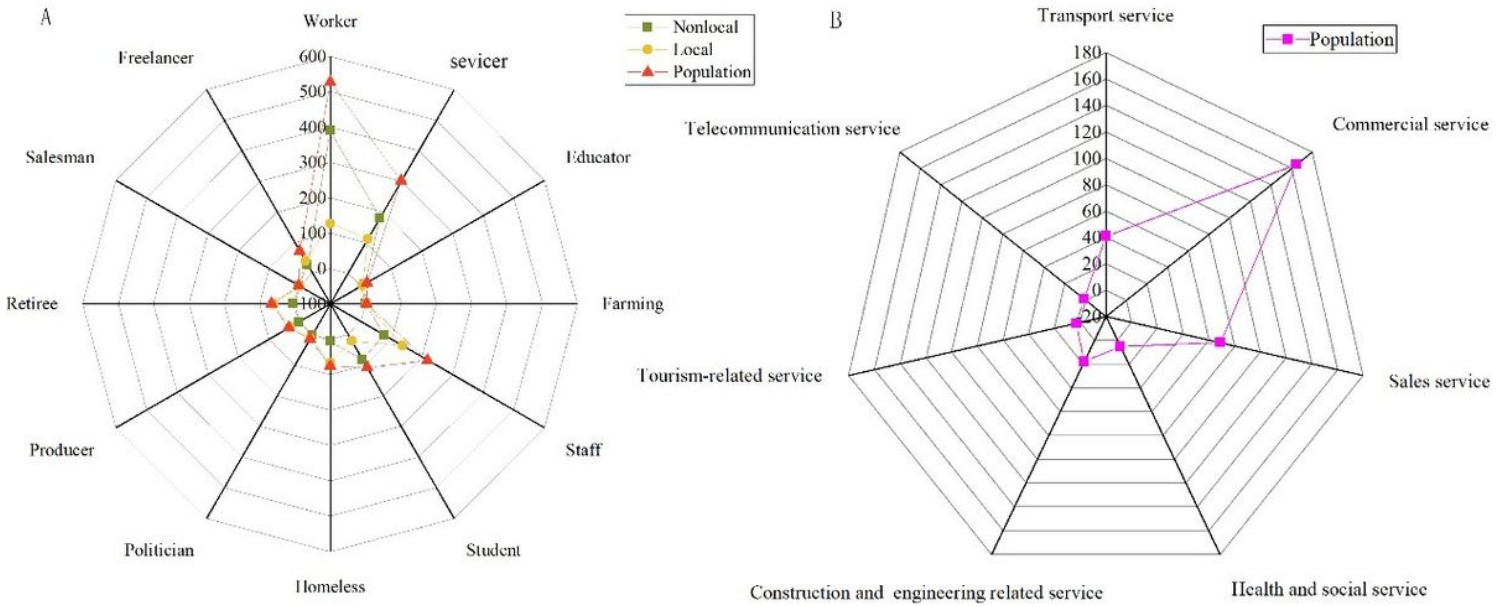

Figure 1

Personnel categories of confirmed COVID-19 cases outside Hubei in China (A: occupation categories; B: service types)

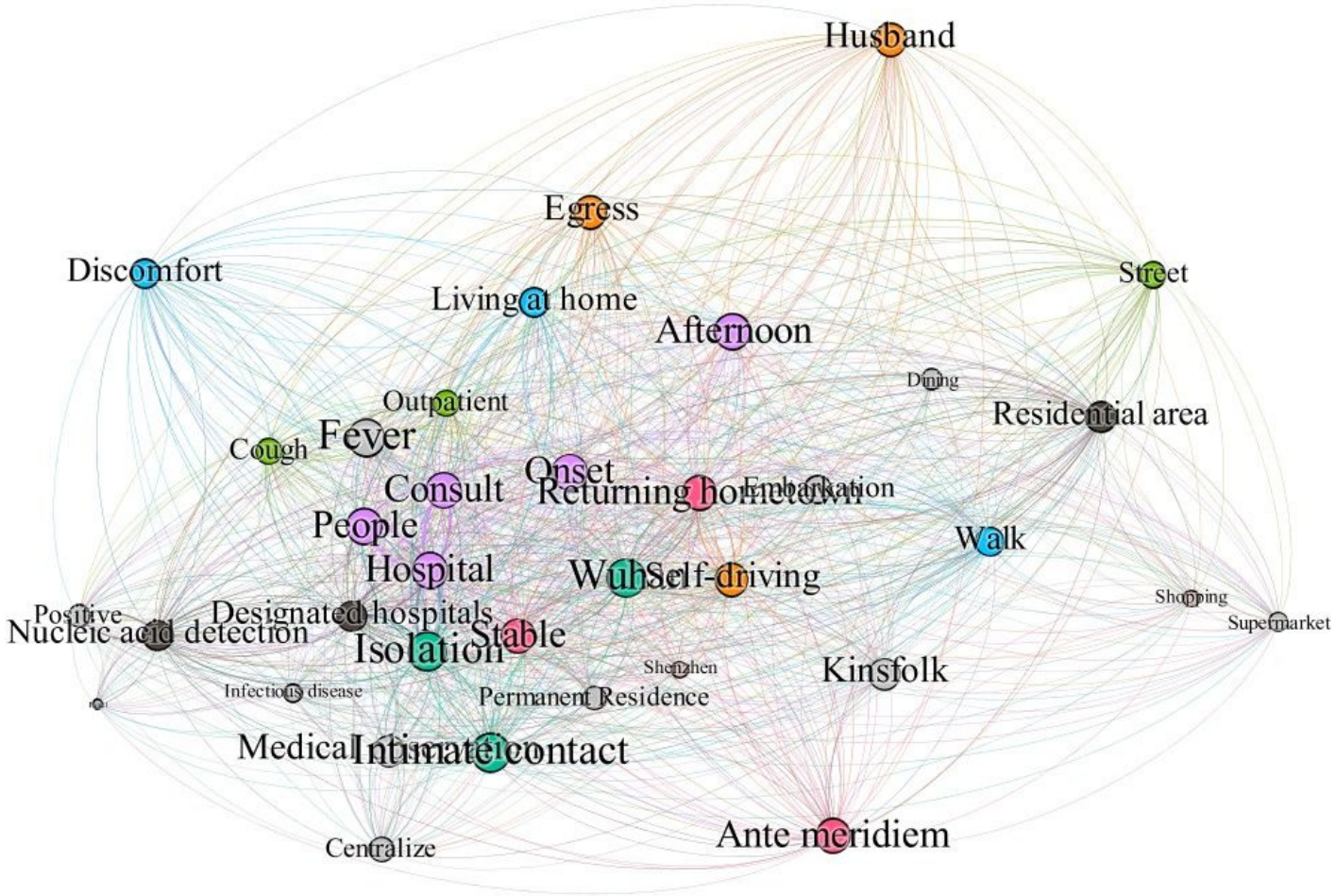

Figure 2

Co-occurrence keywords map of confirmed COVID-19 cases outside Hubei in China

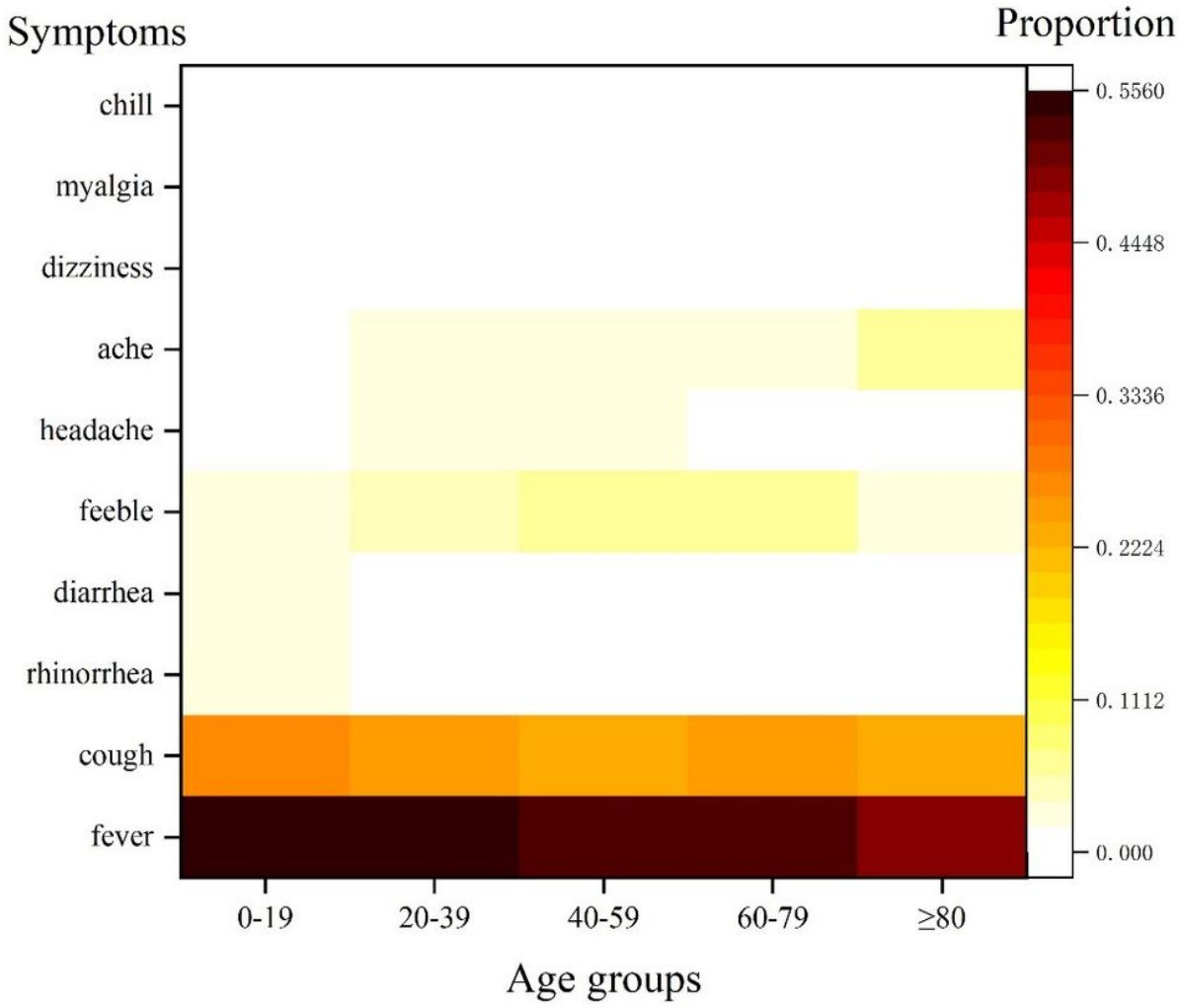

Figure 3

The proportion of symptoms in different age groups of confirmed COVID-19 cases outside Hubei in China

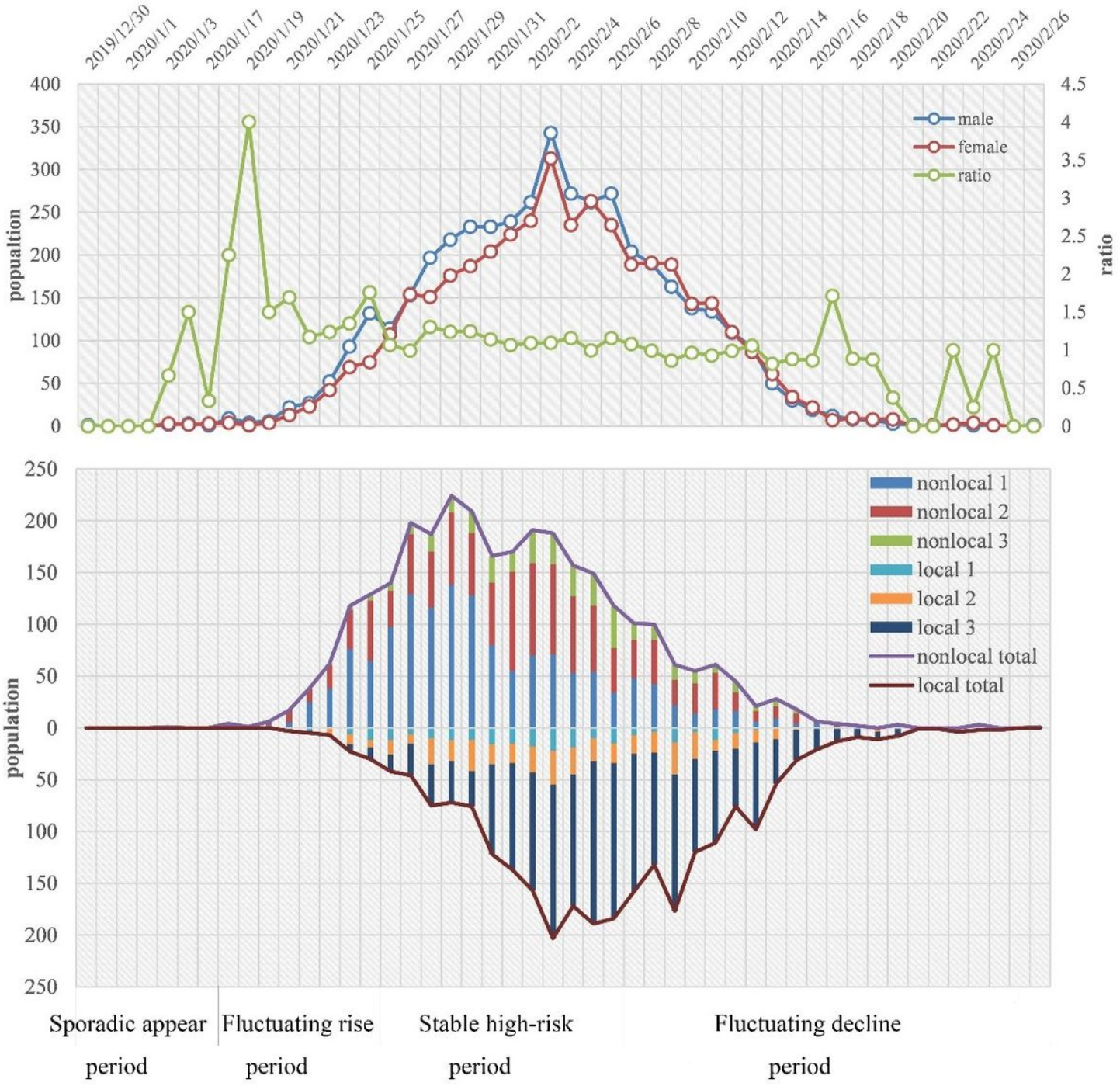

Figure 4

Temporal evolution patterns of the infectious place property and generation

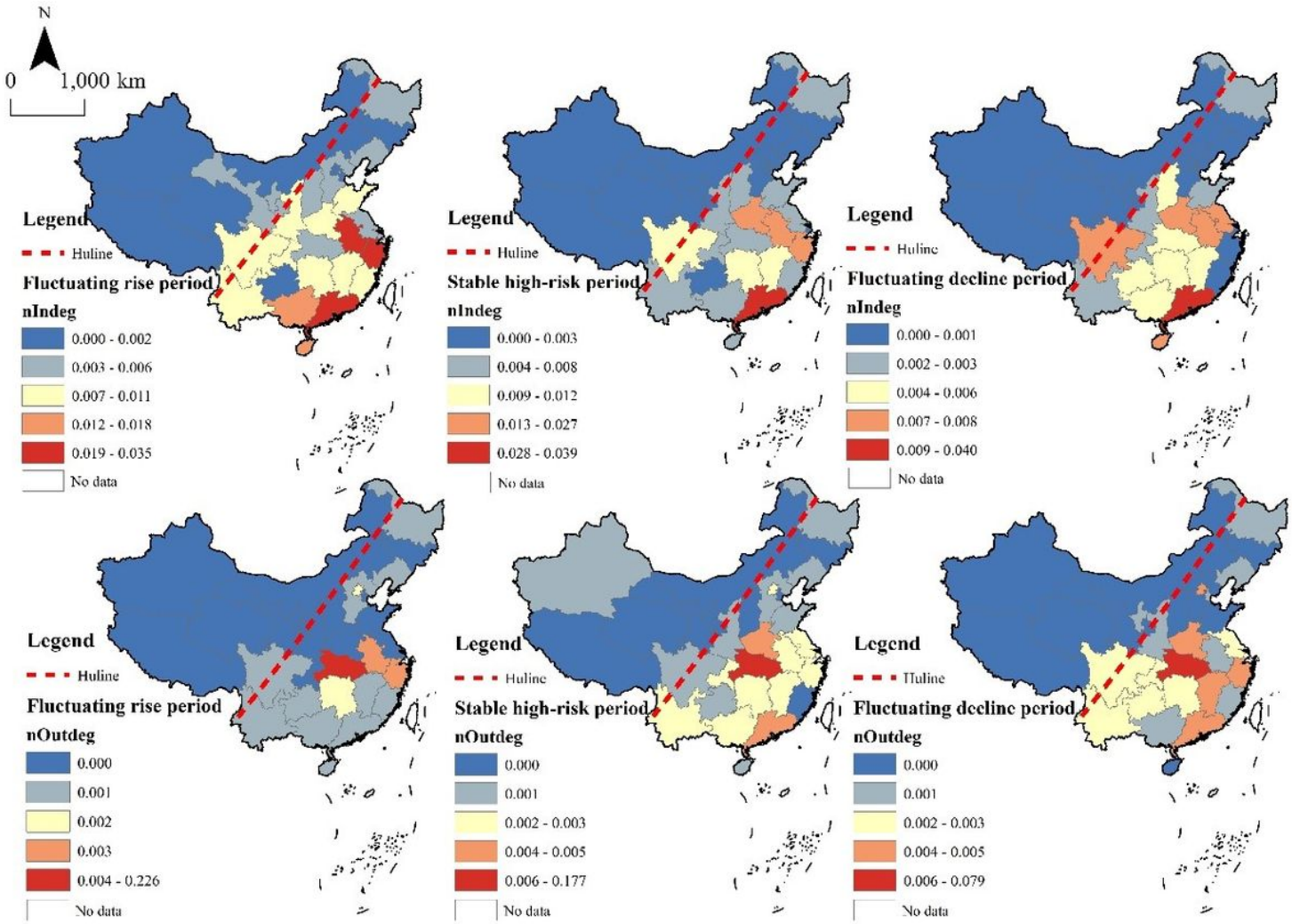

Figure 5

Geographical distribution between normalize in- and normalize out-degree for each province in different period. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

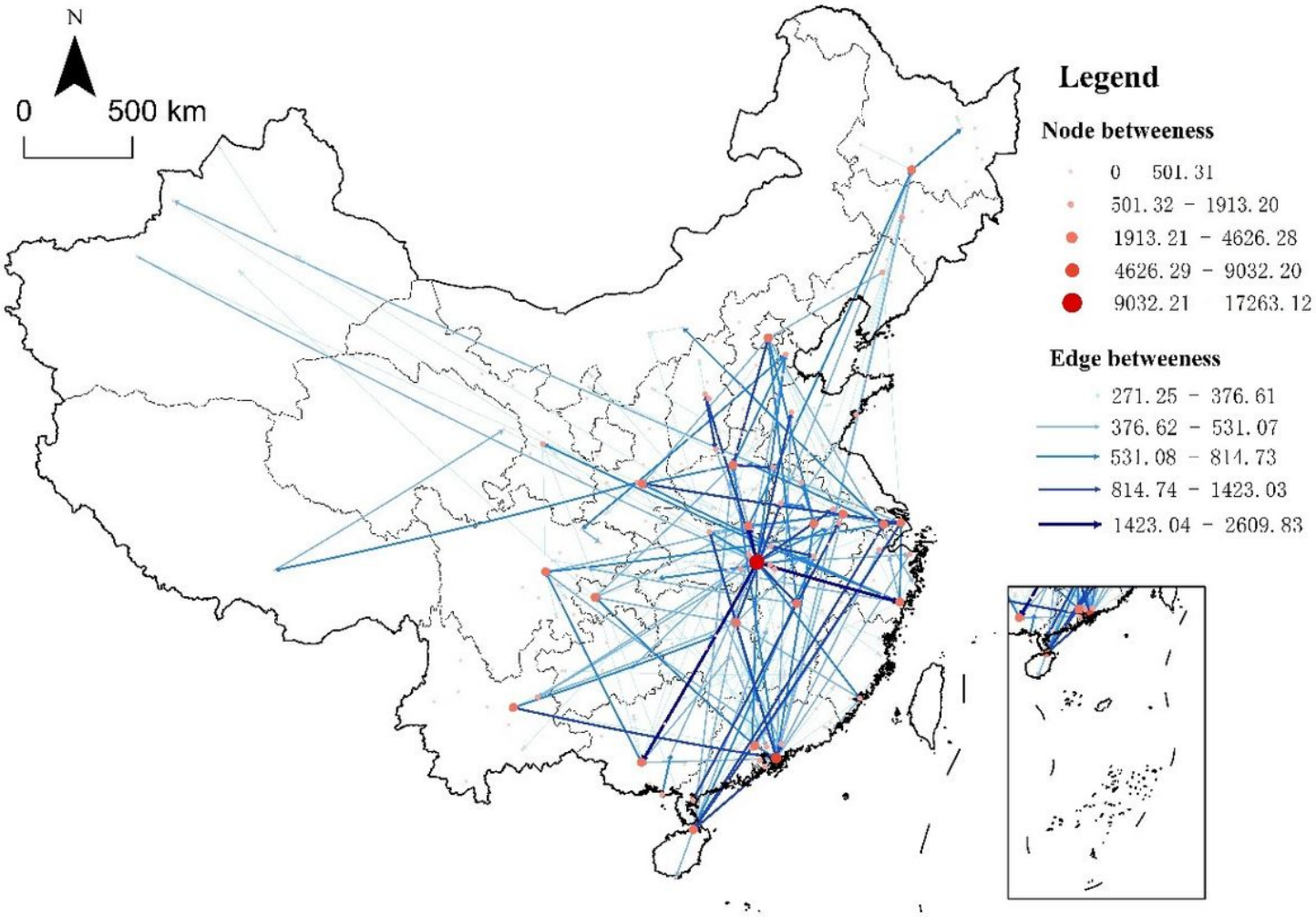

Figure 6

Information flow network between cities. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

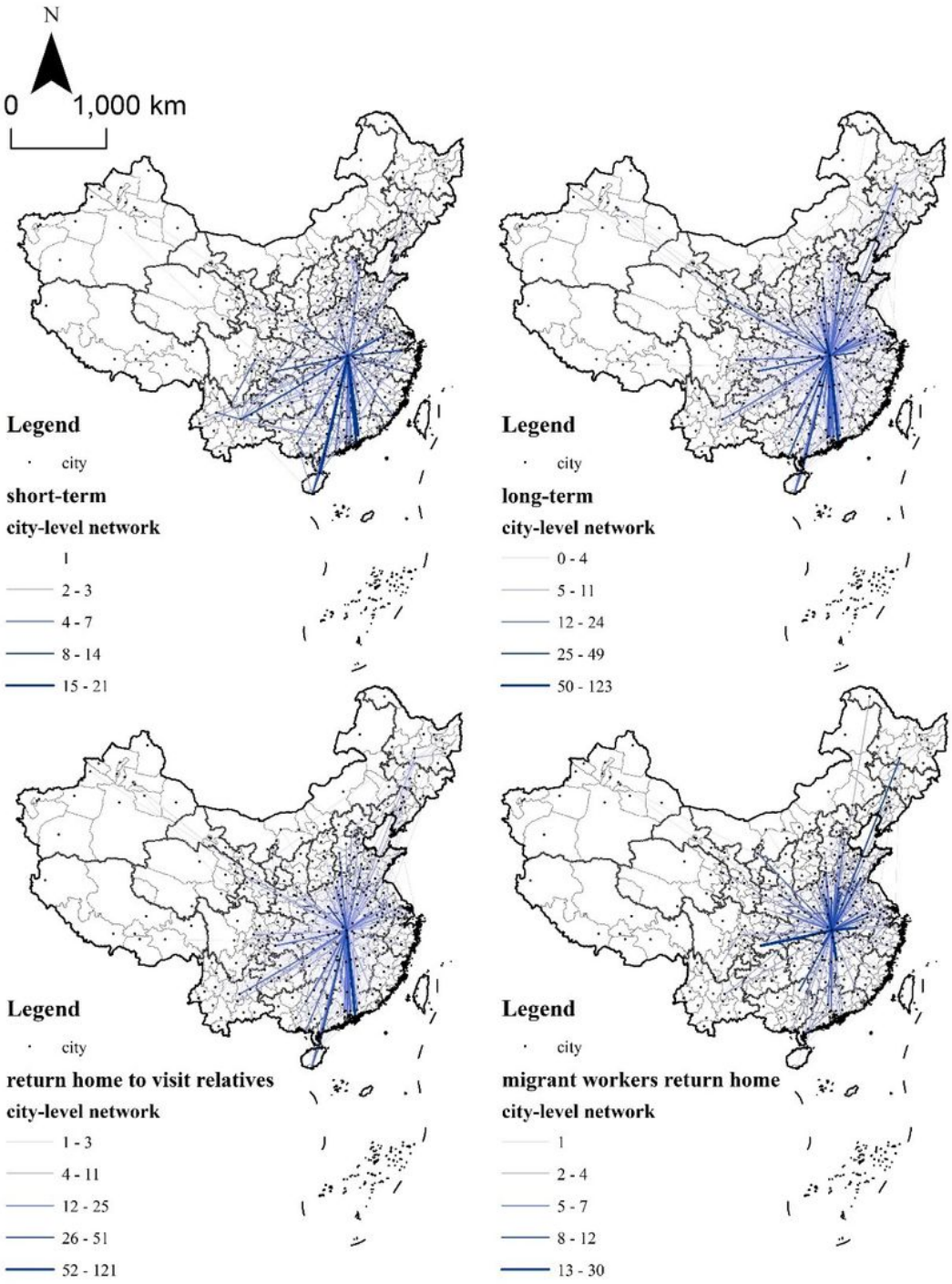

Figure 7

City level network in different purpose. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.