IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 17, 2019, accepted August 18, 2019, date of publication August 21, 2019, date of current version September 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936613

Modeling Air Pollution Transmission Behavior as Complex Network and Mining Key Monitoring Station

CHEN SONG[1], GUOYAN HUANG[1], BING ZHANG[1], BO YIN[2], AND HUIFANG LU[1]
[1]School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
[2]32153 Troops of Chinese People's Liberation Army, Zhangjiakou 075100, China

Corresponding author: Guoyan Huang (hgy@ysu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61772451, Grant 61802332, Grant 61772449, Grant 61572420, and Grant 61807028.

ABSTRACT Air pollution is one of the most serious environmental problems in the world. At present, air quality models are focused on pollutant prediction and source discovery, lacking the analysis of transmission paths and independent monitoring station characteristics. In this paper, complex network based on statistics is applied in air quality field and a new model is proposed based on it. Path existential network and pollutant transmission path set are generated integrating the data of air quality and meteorological monitoring stations. Mapping relation is established among them to generate real-time air quality network, then a complex network in a whole cycle is obtained through statistics. Experiments based on PM2.5 pollution data in Jing-Jin-Ji region demonstrate the rationality of the proposed model. Three characteristics of complex network: scale-free, small-world and community aggregation are verified. Characteristic detection and key station mining provide guidance for air protection in reality. In the network, we can reduce pollution effect by blocking a few important transmission paths between communities. The results provide reference for site selection of new monitoring station, dynamic evolution and pollution degradation.

INDEX TERMS Air quality, complex network, transmission, key station.

I. INTRODUCTION

Clean air is the guarantee of human survival. Air pollution seems like an invisible killer which seriously affects human health and the sustainable development of economy and society. Every year, a lot of people die from air pollution related diseases. A large number of ecological and environmental problems such as acid rain, vegetation damage, climate warming are also closely related to air quality. ''Smoke killing incident'' of Britain resulted in at least 4,000 deaths. Acid rain in former West Germany leaded to the forest destruction of 800,000 hectares. In the winter of 2017, Beijing was covered with haze and tens of thousands people had respiratory tract infections. Fig.1 is the photo taken by NASA of fog and haze. For the serious harm of air pollution, more and more countries pay attention to air quality monitoring and control.

Air quality is complex, multidimensional and transferable, interacted by many factors, such as meteorology changes, characteristics of pollution, physical and chemical reaction processes in the atmosphere. Variations of meteorological factors, such as wind direction, wind speed, atmospheric turbulent motion and vertical distribution of air temperature, have wide impacts on the concentration and path selection in pollutants transmission [1] as well as the emission of pollutant sources. Complex physical and chemical reactions may happen within these factors, which bring great challenges to air quality analysis. Therefore, a reasonable way to integrate the affecting factors of air quality and establish an effective spatial-temporal characteristic model is of fundamental significance for the accurate analysis and improvement of air quality.

Complex network exists widely in biology, engineering, computer science and human society. It can describe a large number of systems in real world and find out common rules hidden in a large number of complex systems. Interpersonal networks [2], gene regulatory networks [3], meteorological interaction networks [4], and protein networks [5] are applications of complex networks. In the field of air quality research, many countries have formed a wide-spread, multi-level air quality system based on air quality monitoring

The associate editor coordinating the review of this article and approving it for publication was Dimitrios Katsaros.

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

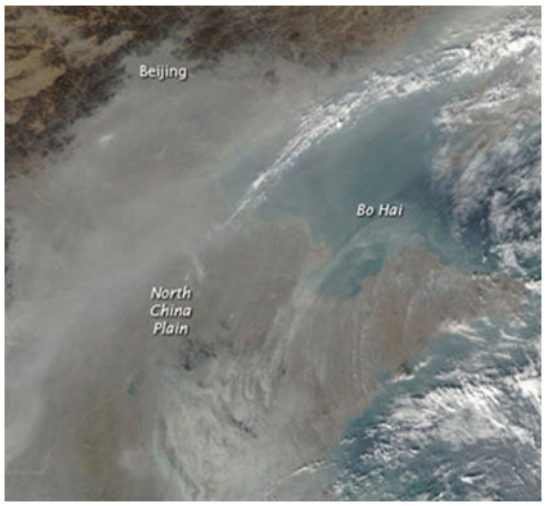


FIGURE 1. Photo of fog and haze in north China taken by NASA.

stations, which has obtained a certain scale of conventional monitoring data. Through the monitored data, we can find its evolution manners are quite similar as meteorological interaction which is complex and changeable, influenced by both meteorological and geographical factors. Monitoring stations on the map is similar as nodes on the graph while complex network is based on graph theory. If we can abstract the pollutant transmission relations into edges, air quality can be abstracted into a whole graph. At present, complex network has already been used in meteorology [6], [7], whereas few examples combine complex network and air quality together.

In this paper, a statistical model of air quality based on complex network is proposed, which integrates spatial and temporal factors together. Based on this model the law of pollutant transmission is explored, and periodic analysis is made by statistical and data mining methods. The main contributions of this paper are listed below: (1) A new air quality modeling method is proposed. The spatial and temporal factors affecting air pollution transmission are extracted and the correlation of local pollutants is analyzed. Mapping on graph is established to generate air quality complex network. (2) Model evaluation criteria system is established according to reality and characteristics of complex network. Network characteristics are analyzed from three aspects: scale-free, small-world and community aggregation. (3) Model characteristics and key node mining are used to guide dynamic analysis and pollution control of air quality in practice.

The main structure of this paper is as follows: Section 2 introduces the existing research methods and knowledge about complex network; Section 3 analyzes some key factors affecting pollutants transmission. Section 4 makes a detailed description about the method establishing complex network statistical model of air quality. Section 5 establishes a directed weighted meteorological network according to the data of Jing-Jin-Ji region. Then we analyze network characteristics and mine key stations. Section 6 discusses the significance and the limitations of the proposed method, pointing out future directions. Section 7 summarizes the full text.

II. RELEVANT RESEARCH

A. EXISTING AIR QUALITY MODELING METHODS

The existing modeling methods can mainly be divided into two types: numerical analysis and statistical analysis. The first kind is numerical analysis model. It relies on complex technical systems such as meteorological information, pollution source inventory and diffusion model. The representative achievement is model-3/CMAQ [8] which is developed by the US Environmental Protection Agency. This model takes all atmospheric factors into account and can comprehensively assess air quality. Denby *et al.* [9] used static interpolation method to model the observed data, in which the annual air forecast spatial map of ozone and sulfur dioxide is calculated by Kriging interpolation multiple linear regression method. Yu *et al.* [10] selected adjacent grids testing their influence on the concentration of air pollutants, and combine with stochastic forest algorithm to predict the concentration of PM2.5. Reference [11] configures model frameworks through one-way nesting, covering the whole Europe from small to large. Numerical analysis has high accuracy which can be adapted to various complex meteorological conditions. However, this method requires high scale and accuracy of input data. Huge amount calculation and high time complexity occur as consequent. The second type is statistical analysis model which relies on the deep understanding of air quality influence conditions, factor selection and model design, mainly including neural network and grey model. Neural network can be combined with genetic algorithm [12], nearest neighbor algorithm [13], FCM algorithm [14]. Grey theory model can be used when only a few samples can be obtained. Wang [15] predicted the concentration changing trend of local SO2 and PM10 by grey theory GM (1,1); Fang [16] established a model based on grey clustering and fuzzy evaluation method, and compared it with API measured data. The shortcoming of statistical model is lack of effective support from air quality mechanism theory, with inadequate constraints and insufficient correlation. Whether numerical or statistical analysis, neglect the correlation between monitoring stations and the independent characteristic of single monitoring station. They cannot give direct guidance to the governance of local pollutants in theory.

B. COMPLEX NETWORK

Complex network is not only a form of data expression, but also a research means. It has strong ability to integrate interdisciplinary fields. With the in-depth development and continuous expansion of research fields, the application of complex network is more and more extensive. A large number of complex systems in real world can be described by various networks. In real systems, different individuals can be abstracted into nodes, and the relationships between individuals can be abstracted into edges. Nodes and edges form the topological structure of complex network.

Complex network emerged in the 1990s. People found the vast majority of networks in real world not only conform to the partial characteristics of regular and random networks,

but also have scale-free [17] and small-world [18] characteristics. Scale-free is the description of the non-uniform interaction force in a network. Only a few nodes in the network have a large number of connections, while most nodes have a small number of connections. Small-world reflects the features of short-path length and high clustering coefficient. That is to say, complex network has much smaller average distance between nodes than regular network and much larger average clustering coefficient than random network. Community structure [19] is also an important characteristic of complex networks. Nodes in a same community connect closely while nodes between communities connect sparsely. Information spreads rapidly in complex networks.

A new air quality characterization method based on complex network is proposed in this paper which takes air quality monitoring stations as nodes and pollutant transmission paths as edges. The interaction relationship and intensity between nodes are characterized considering meteorological and geographical factors synthetically. Compared with the previous methods, less priori knowledge is needed in our model. Pollution source information is not required during modeling process. In the generated directed weighted network, representation of nodes and edges makes the pollutant transmission between regions clearer. The hierarchical structure of nodes, edges and networks enables the research to be carried out at different levels, not only for single monitoring station, but also for inter-regional transmission path and macro-laws.

III. RELEVANT INDICATORS OF AIR POLLUTION TRANSMISSION

Air pollution transmission is a complex process, which is mainly affected by geography, meteorology and pollution itself. Pollution transmission speed varies under different conditions. Chapter 3 analyzes the key factors affecting pollution transmission. After calculation, all data are standardized by min-max method.

A. DISTANCE

Distance is an important factor affecting pollutant spread. Without considering the influence of other factors, pollutant transmission is a radial diffusion from source to surroundings; pollution decreases with increasing distance. In fact, the closer the two places are, the higher pollutants correlation they will have. When a region is seriously polluted, the pollutants concentration of its vicinity tends to increase synchronously. In earth coordinate system, the distance between air quality monitoring stations can be calculated by longitude and latitude. Fig.2 is the diagram of the distance between two stations in the Earth's coordinate system. Assuming that the earth is a standard ellipsoid, and the average radius R is 6356.76km. The coordinates of station i are (lo_i, la_i) and station j are (lo_j, la_j). According to trigonometric theorem, the distance between two monitoring stations is as follows:

$$C_{ij} = \sin(la_i)\sin(la_j)\cos(lo_i - lo_j) + \cos(la_i)\cos(la_j) \quad (1)$$

$$Dis_{ij} = R \cdot \arccos C_{ij} \quad (2)$$

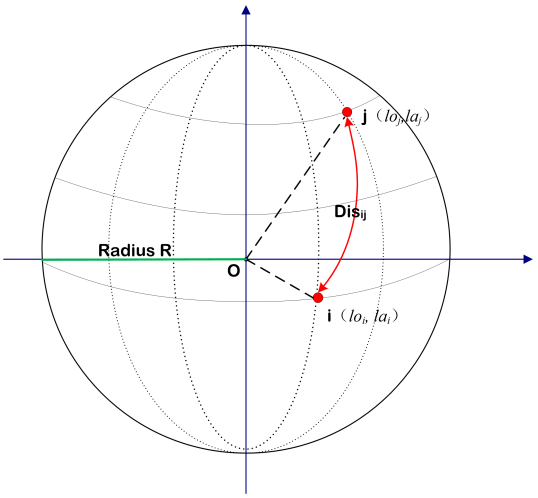


FIGURE 2. Distance between two stations.

In (1) and (2), Dis_{ij} is the distance between station i and j. C_{ij} is a intermediate value.

B. WIND FORCE DIFFERENCE

Wind has a significant impact on pollutant transmission, and pollutant spreads along wind direction. The greater the wind force difference, the faster the transmission speed. As shown in Fig.3, there are two stations i and j. At time t, the wind force of station i is $F_i(t)$ and site j is $F_j(t)$. The components of $F_i(t)$ and $F_j(t)$ in $\overrightarrow{D_{ij}}$ direction are $v_i(t)$ and $v_j(t)$ respectively. Wind force difference $\triangle F_{ij}(t)$ between station i and j at time t is:

$$\triangle F_{ij}(t) = |F_i(t)| \cdot \cos\theta_i + |F_j(t)| \cdot \cos\theta_j \quad (3)$$

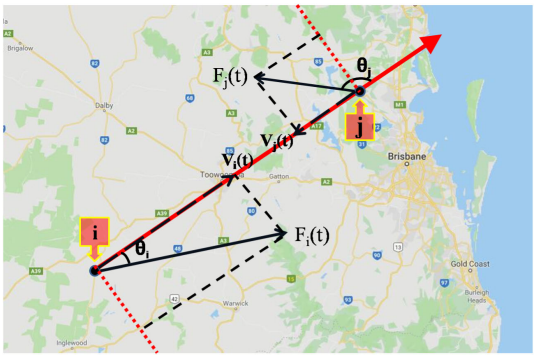


FIGURE 3. Wind force difference.

When the wind direction angle between $F_i(t)$ and $\overrightarrow{D_{ij}}$ exceeds $\frac{\pi}{2}$, the air pollutants at station i will not affect j. Or when $\triangle F_{ij}(t)$ is less than or equals with 0, it means the wind coefficient of i at $\overrightarrow{D_{ij}}$ direction is less than or equals to the wind coefficient at site j. In these two circumstances, we set the value of $\triangle F_{ij}(t)$ to 0.

C. HUMIDITY

For most gases, chemical reaction with water vapor will not happen at room temperature. In the range of daily air humidity, the diffusion coefficient increases with the increase

of air humidity. The greater the humidity is, the faster the pollutants diffuse.

D. OTHER FACTORS

Elevation difference between two places affects the spread of pollutants, for that low-lying area have little influence on the high-lying one. Continuous mountain blocks the propagation of pollutants. Air pressure also has impact on transmission. Pollutants tend to flow from high air pressure to low air pressure. The greater the air pressure difference is, the easier the pollutants spread. Wind is the direct manifestation of air pressure difference. In addition, temperature, precipitation, dust and so on also have impact on pollutant transmission.

IV. METHOD OF NETWORK ESTABLISHMENT AND STATION MINING

In air quality complex network, take monitoring stations as nodes and pollutant transmission paths as edges[20]. Edge direction is the direction of wind force difference, and edge weight is the occurrences of corresponding edge. Then a network can be abstracted into graph $G = (V, E)$ with n nodes and m edges. V represents the set of nodes (v_0, v_1. $v_2, \ldots v_n$,). HashMap E represents edges between nodes and the corresponding weight, e_n= (a_{xy}, w_{xy}). Model establishment is a statistical accumulation of PM2.5 transmission network at all times during a cycle. The overall framework is shown in Fig.4.

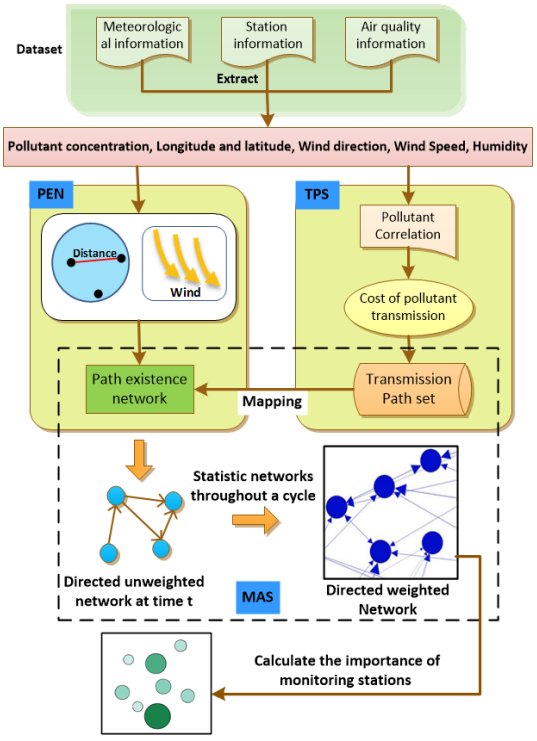


FIGURE 4. Overall framework of air quality network.

Firstly, networks are constructed in hours. Set the constraints that transmission path can exist between two nodes then construct directed weightless path existence network among nodes that conform to constraints. According to the spatial-temporal distribution of pollutants and correlation in practice, the pollutant propagation relationship is obtained, and the directed weightless complex network is generated by mapping to the existential network. Accumulate the occurrences of each edge as edge weight to generate transmission path matrix. Air quality statistical model of complex network during a cycle is constructed.

A. PATH EXISTENCE NETWORK (PEN)

Air quality is affected by many factors, as key factors play a decisive role in the transmission process. In practice, pollutants will not propagate against wind. When the wind force difference between two stations is less than 0 or in upwind direction, it is considered no transmission path will appear between the two stations. When the distance between two stations is too far, we assume the pollutants cannot be transmitted in real time. And if the distance is very close, pollutant information is basically synchronized then a large number of loops will form in the network. Therefore, too far or too close distances are both considered no transmission paths exist.

At time t, before modeling the transmission path between pollutants, relationship between stations should be summarized first. Considering both wind direction and actual distance, node pairs without transmission basis should be eliminated. Path existence network among all nodes is generated. The non-existent edges in the network indicate the transmission path cannot exist. It is necessary to control the reasonable distance according to region size and average distance between monitoring stations. Path existence network is established among nodes which conform to constraints. The direction of wind force difference between nodes is taken as edge direction.

B. TRANSMISSION PATH SET (TPS)

Path set at each time t is established in a cycle and corresponds to existence network one by one. The establishment of paths does not simply depend on meteorological factors or real-time performance of pollutants. Comprehend different situations and based on the dynamics of pollutant transmission, the paths at corresponding time is generated combining with actual distribution.

1) POLLUTANT CORRELATION

If the pollutant changing trend is the same between two stations, a transmission path is more likely to exist between them. According to the time series of air quality at each node, the correlation can be calculated. In this paper, Pearson coefficient is used to evaluate the linear correlation between pollutants. Assuming the pollutant concentration of monitoring station A is $X_i(t)$ at time t, and station B is $X_j(t)$. The average concentration of pollutants in monitoring station i is $\overline{X_i}$, while j is $\overline{X_j}$. The correlation of pollutants between

monitoring stations A and B can be obtained by (4).

$$\rho_{ij} = \frac{\sum(X_i(t) - \overline{X_i})(X_j(t) - \overline{X_j})}{(\sqrt{\sum_{i=1}^{n}(X_i(t) - \overline{X_i})^2})((\sqrt{\sum_{i=1}^{n}(X_j(t) - \overline{X_j})^2}))} \tag{4}$$

2) TRANSMISSION COST

Transmission cost is used to measure the difficulty about pollutant transmission between stations. The model established in this paper is a preliminary model based on statistics, which takes the correlation between real existing data as an important factor. Based on easily accessible and numerically represented information among key meteorological characteristics affecting transmission, a simple definition of transmission cost from station A to station B is defined combining with the Pearson correlation of pollutants.

$$cost_{ij} = \frac{Dis_{ij} * Hum_i(t)}{\rho_{ij} * \Delta F_{ij}(t)} \tag{5}$$

Before calculating, min-max method is used to standardize data. Dis_{ij} denotes the horizontal distance between two stations; $Hum_i(t)$ denotes the humidity in starting node i at time t; $\rho_{ij}(t)$ denotes the correlation between station A and B; $F_{ij}(t)$denotes wind force difference. In practice, more complex correlation may exist between these factors and transmission paths. Here we simply take them as proportionate.

For a fixed time t, the transmission cost between node A and other nodes in the network is calculated. When the cost is less than the threshold, there is a path between node i and the corresponding node. All nodes are traversed and the path set at time t is established.

C. MAPPING AND STATISTICS (MAS)

At time t, mapping path set to path existence network, retain the paths that both appear in existence network and path set, while the other paths are deleted. If a path only appears in the path set, it shows that the path does not have basic existent conditions. Another situation is when a path only exists in the existence network, it shows although the path has appearance basics, it does not happen for the affection of many actual factors in transmission process.

After the establishment of air quality network at each time, matrix conversion for storage should be done. Statistical analysis of the network matrix at each time in the cycle is executed. Nodes correspond one-to-one and directed paths are accumulated. Take statistical path occurrences as edge weight. We can get the preliminary generation of air quality complex network. Fig.5 shows an example of mapping and statistic.

In Fig.5, the obtained paths are stored in path set and map to path existence networks to generate complex network. Among the 5 paths in path set, e_{12} and e_{41} is included in path existence network. These two paths are retained and the other non-existent edges are deleted, generating a complex network at the corresponding time which is stored in matrix. The air quality complex network is obtained by accumulating the network matrix at each time in the cycle.

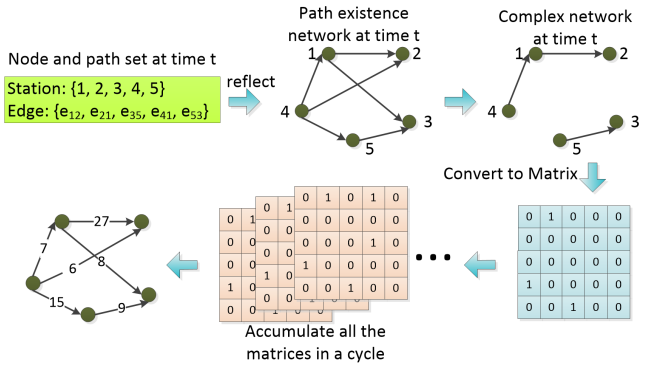


FIGURE 5. Process of mapping and statistics.

D. KEY STATION MINING

In complex network, node is a basic component. In this paper, we take air quality monitoring stations as nodes. Key nodes [21] in complex networks are some important nodes which affect network function and structure in a greater extent. Mining key nodes in air quality network can quickly find important areas in mass data which affect air quality greatly. These nodes often have important geographical location or serious pollution and make strong influence on other nodes. The mining method adopted in this paper is NPSI [22] which is based on similarity. Similarity is used to describe how similar two nodes are. When a node is similar to many nodes in the network, it is representative and can be treated as an influential node. Add the similarity between node x with every other node in the network and define the value as normalized probability similarity influence(NPSI). The greater of its value,the higher similarity node x has with more nodes in the network.

In order to get an accurate similarity result, probability walking model is used to simulate the active access process. Release walkers continuously with $\Delta t = 1$, until the step of the first walker is t. At the same time the number of walkers in the network is also t. Supposing the starting point is node x, walkers can only choose a path from the edges whose starting points are node x as the next step. After t steps, we can get the probability of walkers reaching y from x in $1 - t$ steps.Considering the influence on probability of both direction and weight, the one-step transition probability in directed-weighted network is defined as follows:

$$P_{xy} = \frac{a_{xy}}{k_{xout}} \cdot w_{pec} \tag{6}$$

Based on probability walking model, the formula of NPSI for evaluating node importance is defined as follows:

$$NPSI_x(t) = \sum_{i=1}^{n}\sum_{j=1}^{t} \frac{k_{xout}}{m} \cdot P_{xy}(j) + \frac{k_{yout}}{m} \cdot P_{yx}(j) \tag{7}$$

where n is the total number of nodes in the network and m is the total number of edges in the network. t is the number of walkers, as well as step number in probability walking model. k_{xout} and k_{yout} are the out-degree of nodes x and y.

$P_{xy}(t)$ is the corresponding value in probability transfer matrix which means the reachable probability from node x to y within j steps. NPSI importance is related to node location and adjacent nodes in the network. The adjacent nodes here include chain-in and chain-out nodes. The more adjacent nodes a node has, the more important the node is. The more important its adjacent nodes are, the more important the node is.

V. EXPERIMENT

In 2016, the World Health Organization announced 30 cities with the worst air quality in the world, among which 6 cities in Jing-Jin-Ji region were listed. The serious air pollution in this region is typical. The experiment researches air quality data of Jing-Jin-Ji, establishing air quality complex network model. Analyze its rationality and scalability and excavate the key nodes to reveal the evolution law of pollutants. The dataset includes hourly air quality and meteorological data from 220 air quality monitoring stations from May 1, 2014 to April 30, 2015. The East-West distance of the research region is about 500 km, and the North-South distance is about 950 km. When establish transmission paths, the selected longest distance [23], [24] is 200 km (the average distance between stations), and the shortest distance is 10 km. The number of monitoring stations in dataset is as Table.1.

A. GENERATING COMPLEX NETWORKS

The geographical locations of monitoring stations and the generated network in Jing-Jin-Ji region are shown in Fig.6.

TABLE 1. Monitoring station id in Jing-Jin-Ji cities.

City	Station Id	Total Number
Beijing	1001-1036	36
Tianjin	6001-6040	27
Shijiazhuang	11001-11025,12001	26
Tangshan	13001-13018	18
Qinhuangdao	14001-14009	9
Baoding	17001-17028,18001	29
Zhangjiakou	19001-19020	20
Chengde	20001-21014	14
Cangzhou	21001-21016	16
Langfang	22001-22012	12
Hengshui	23001-23013	13

After eliminating the default and redundant data, 691597 data at different times is obtained. There are 152 nodes and 925 edges in the generated complex network. The weight in the network is shown by edge color depth and corresponding arrow size. Darker color and bigger arrow represent bigger weight of the corresponding edge.

In order to intuitively analyze the pollutant transmission behavior of the abstracted network and verify the transmission paths between different regions and cities, the top 10 weighted edges between monitoring stations whose transmission distance is more than 30 km are extracted on the map, as shown in Fig.7.

Fig.7 shows the high-weighted transmission paths on Jing-Jin-Ji topographic map taken by Google Earth. Each path means the pollutants have been transmitted many times between the two stations in a year. From the topographic map, we can see that the western and northern parts of the region

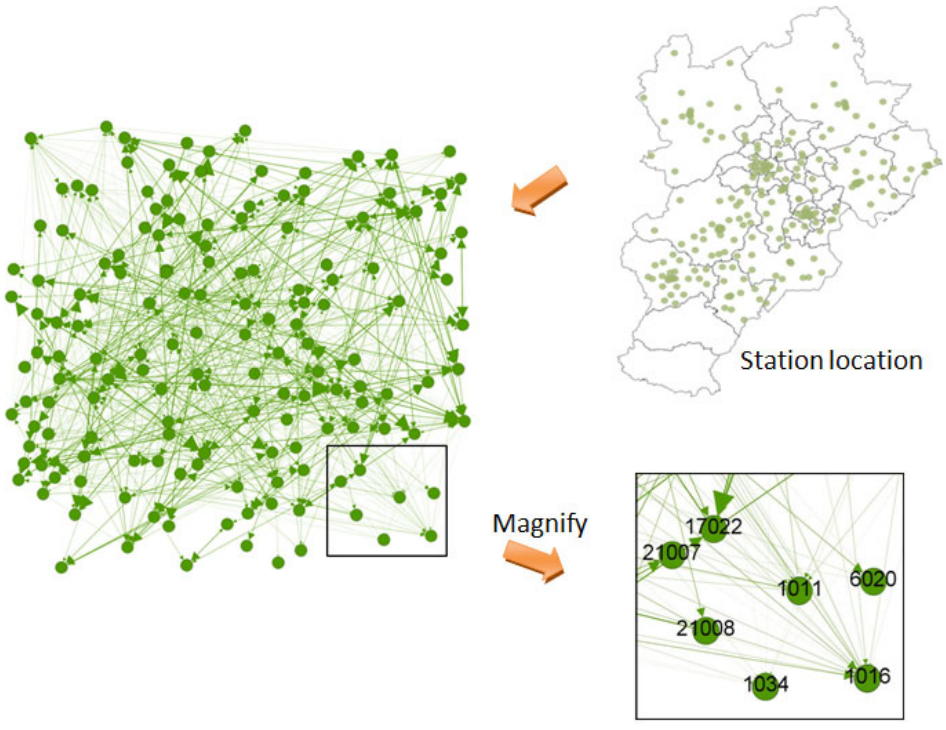


FIGURE 6. Complex network of Jing-Jin-Ji region.

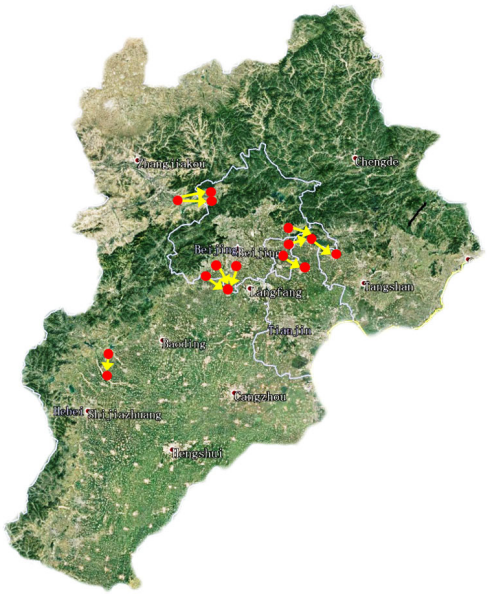


FIGURE 7. Top 10 weighted edges between stations.

are surrounded by Yanshan Mountain Range. To the east of the region is the Pacific Ocean, and its northwest is backed by the Siberian Plateau. Jing-Jin-Ji region belongs to the temperate monsoon climate in Northern Hemisphere. Winter is the most serious polluted season, in which the Siberian cold flows to the southeast and northwest wind prevails. In summer, the southeast monsoon from the ocean is relatively mild, so northwest wind is the dominant wind direction in a year. The top 10 weighted paths in Fig.7 generally transmit along the northwest to southeast, which is consistent with reality. Pollutants can be blocked by mountains and accumulate near them. All the paths on the topographic map do not spread across mountains, and almost all of them are at mountain foot, transmitting along mountain direction. This situation is also consistent with the reality.

B. CHARACTERISTIC ANALYSIS

The rationality of the air quality network generated by the experiment is analyzed from the general characteristics of complex network: scale-free, small-world and community structure. A network meets part or all of the three characteristics can be treated as a complex network. The characteristics are also used to explore the laws of air quality evolution.

1) SCALE-FREE AND SMALL-WORLD

Node degree in complex networks conforms to power law distribution. Fig.8 shows the occurrence frequency of node degree in complex networks. Node degree is discrete, and the occurrence probability of each degree is not uniform. Few nodes have high degree and many nodes have low degree. In general, node degree in air quality network conforms to power law distribution, which shows the network is obviously scale-free. In order to judge the scale-free characteristic of air quality network better, we compare its degree distribution

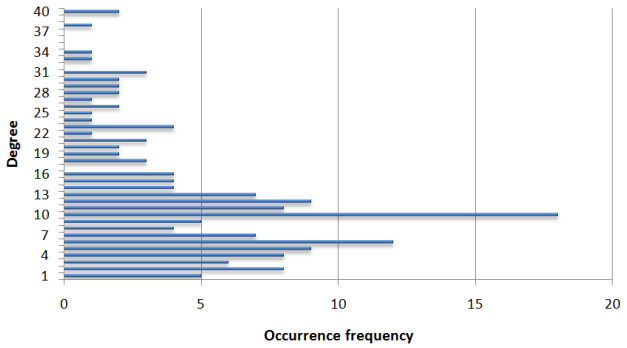


FIGURE 8. Occurrence frequency of node degree.

with some classical complex network datasets. Polbooks and American Football are currently widely used datasets in complex network field. As can be seen in Fig.9, air quality network has similar trend as the classical network and the same scale-free characteristics. The average path length is 7.766 and the average clustering coefficient is 0.507 which is obviously higher than random network whose average clustering coefficient is always lower than 0.1.

In air quality network, the average shortest path length is 6.80, and the average clustering coefficient of nodes is 0.40. In ER random networks, the average shortest path length is generally higher than 7, and the average clustering coefficient is lower than 0.2. Compared with random network, the network constructed in this paper has a high cohesion degree and obvious small-world characteristic.

2) COMMUNITY STRUCTURE

Community structure [25], [26] is a subset of node group that in which nodes closely related to each other and between which the relationships are relatively sparse. Nodes in the same community often share some common characteristics. Community structure is usually evaluated by modularity Q [26] which ranges in [0,1]. The closer to 1 Q is, the stronger community structure is. Specific formula is as follows:

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2|E|})\delta(c_i, c_j) \tag{8}$$

In (8), m is the total number of edges. A_{ij} is the adjacency matrix. k_i is the degree of node i. For node v_i, it has a community label ci. δ is Kronecker function and the independent variables are usually two integers. If the two variables are equal, the function value is 1, otherwise it is 0. Significant community structure is considered when Q is greater than 0.3. Figure 10 is the community detection result of air quality network [27]. Its Q is 0.766. Therefore, air quality network has community structure characteristic, and it is significant.

As can be seen from Fig.10, the whole network is divided into 11 communities. For community, internal relations are dense and external relations are sparse. In air quality network, the organization of communities has obvious regional characteristics, but not limited to region at the same time.

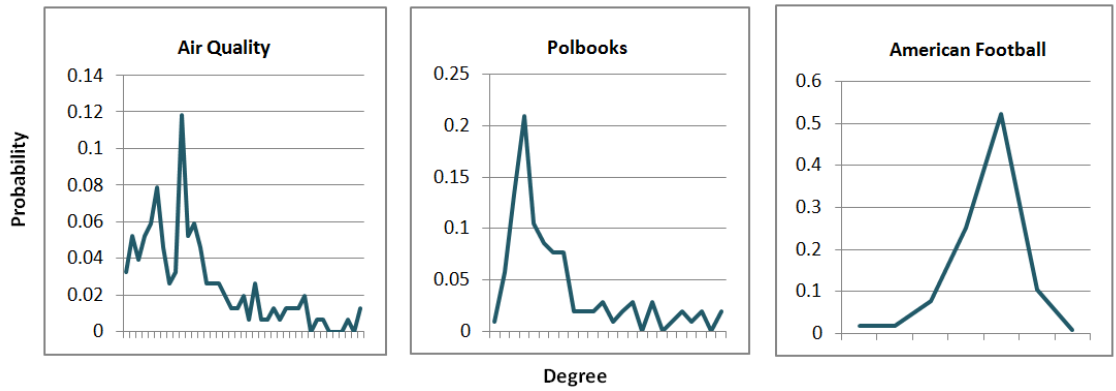


FIGURE 9. Comparison of network degree distribution.

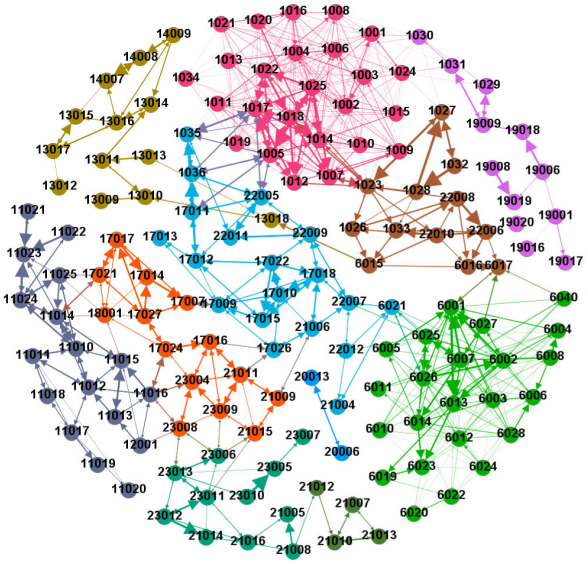


FIGURE 10. Community structure after detection.

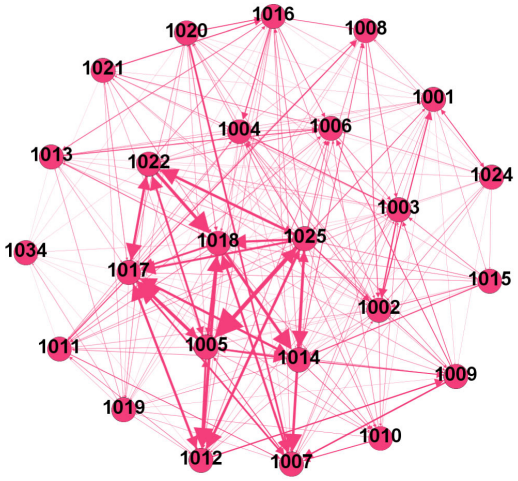


FIGURE 11. A community in Beijing.

For the sparse connection among communities, these inter-community connections are very important for information exchange. Some bridging nodes located between communities play a significant role in information spreading between different communities, such as nodes 1017, 17009, 6001.

In the community structure graph of Fig.11, color depths and arrow size indicate edge weight. It can be seen that the communities in Beijing is very closely linked. At the same time, there are several high-weight edges. This shows that pollutants cannot be effectively degraded after aggregation in Beijing, resulting in repeated transmission. Analyze with the geographical conditions in Beijing area. The city is high in the northwest and low in the southeast and surrounded by mountains in west, north and northeast. After arriving in Beijing, pollutants are blocked and retained by mountains, accumulating continuously. They can only spread in a small area, causing serious pollution. This phenomenon also can be treated as a validation of model rationality.

C. KEY STATION MINING

According to NPSI algorithm, rank node importance of the obtained air quality network and the top 10 key nodes are shown in Table.2.

TABLE 2. Top 10 key stations in the whole network.

Rank	Station Id	Station Name	NPSI	Degree	Over standard rate of PM2.5
1	1005	Liangxiang, Fangshan	10.00	40	30.34
2	1017	Huangcun Town, Daxing	9.93	40	31.35
3	1023	Shunyi New Town	8.47	29	24.74
4	1025	Longquan Town, Mentougou	8.27	38	20.43
5	1018	Yizhuang Development Zone	7.36	33	31.76
6	6001	Huaihe street, Tianjin	6.04	23	26.68
7	22005	Gu'an Party School	5.96	14	34.10
8	6007	Xinhua Road ,Tianjin	5.76	30	22.04
9	11010	Zhengding Unicom	5.68	13	22.05
10	1007	Chaoyang Olympic Center	5.64	25	25.22

The degree of nodes in Table.2 is the total degree which sums in-degree and out-degree up. The mined key nodes often have high degree or serious pollution rate, which can affect

the overall structure of the network to a greater extent. The number of key nodes in Beijing is the most, which indicates Beijing is a key area.

The most important node in each city is listed Table.3. The key nodes in Beijing, Tianjin and Langfang rank front. These cities also are the most polluted ones in Jing-Jin-Ji region. Zhangjiakou and Chengde rank behind in the key nodes, indicating little contribution to pollution the two cities made in the whole air quality network. In fact, these two cities have large area of forest and grassland so their environment is of good quality. The discovery of key stations in each city has valuable reference for the observation and protection inside a city as well as the site selection of new monitoring stations. Emphasis should be laid on key nodes both in the control of pollution sources and air purification.

TABLE 3. Top 1 key station in each city.

City	Id	Network Rank	NPSI	Degree	Over standard rate of PM2.5
Beijing	1005	1	10.00	40	30.34
Tianjin	6001	6	6.04	23	26.68
Shijiazhuang	11010	9	5.68	13	22.05
Tangshan	13016	59	2.75	10	25.53
Qinhuangdao	14007	60	2.74	15	26.22
Baoding	17018	14	5.13	12	46.86
Zhangjiakou	19009	95	1.69	5	11.27
Chengde	20006	139	0.59	2	14.32
Cangzhou	21011	32	3.97	12	46.05
Langfang	22005	7	5.96	14	34.10
Hengshui	23004	45	3.42	11	42.66

VI. DISCUSSION

This paper abstracts transmission path based on influencing factors and actual pollutants distribution. Complex network is used to describe the evolution of air quality. Air quality network is proved conform to the general characteristics of complex network with good reliability and expansibility. The key edges and nodes are consistent with reality. Previous studies on air quality are mostly based on machine learning and statistical analysis whose purposes are focused on air quality prediction and source mechanism. These methods lack analysis about interregional interaction of pollutants and characteristics of independent monitoring stations.

We combine air quality evolution with statistical complex network. Nodes and edges clearly display the evolution process of air quality. The model in this paper focuses more on governance than prediction. Its scale-free and small-world characteristics indicate that air quality network pollutants are relatively concentrated while easy to spread. In such a network, its anti-destructive ability is weak since a few changes in connection can dramatically change the performance of the network. This feature can be fully utilized in air control. A small number of nodes and paths can be controlled to make maximum effect on air quality control in the region. In the process of community detection, pollutants correlation is high between monitoring stations within the same community. When the cohesion degree of a community is much higher than that of other areas, it indicates that pollutants accumulate at local and spread repeatedly. Pollutants in this area need urgent degradation. Connections between communities and bridging nodes act as liaison officers, which are the links of pollutants between regions. Cutting off the path between communities can effectively prevent the transmission of pollutants in different regions and control pollution in a certain range. Key nodes mining in the network can help us quickly find stations with serious pollution and strong dissemination ability. Key nodes should be selected out for key protection and governance. Controlling the emission of pollution sources, planting green plants and actively degrading harmful gases in these areas will significantly improve the air quality of the whole network. In the site selection of new monitoring station, density is the most important factor. Then priority should be given to the junction of communities and the perimeter of key nodes.

Complex network is a means of scientific research, its combination with air quality can help us reveal more laws. The model established in this paper is only a preliminary exploration; many limitations remain to be solved. More data in different cycles is needed to get accurate results. A better model should be applied to simulate pollutant transmission cost. In future, we will try to use link prediction in complex network in air quality prediction.

VII. CONCLUSION

Spatial-temporal variations of air quality can be modeled in complex networks. The model has reasonable structure and reliable results. The scale-free, small-world and community aggregation properties of complex networks are well reflected in the network. Based on these basic characteristics, we can better analyze the pollutant transmission mechanism and air quality evolution law, simulating the similarity and difference between stations accurately. The analysis of network characteristics and mining of key stations enable us to have a better understanding about the transmission law of air quality and pollutants treatment. This model has an objective application prospect in city planning, monitoring station positioning and pollution control.

REFERENCES

[1] M. Tang, X. Wu, P. Agrawal, S. Pongpaichet, and R. Jain, "Integration of diverse data sources for spatial PM2.5 data interpolation," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 408–417, Feb. 2017. doi: 10.1109/TMM.2016.2613639.

[2] M. A. Al-Garadi, K. D. Varathan, S. D. Ravana, E. Ahmed, G. Mujtaba, M. U. S. Khan, and S. U. Khan, "Analysis of online social network connections for identification of influential users: Survey and open research issues," *ACM Comput. Surv.*, vol. 51, no. 1, p. 16, Jan. 2018. doi: 10.1145/3155897.

[3] F. M. Lopes, R. M. Cesar, Jr., and L. D. F. Costa, "Gene expression complex networks: Synthesis, identification, and analysis," *J. Comput. Biol.*, vol. 18, no. 10, pp. 1353–1367, May 2011. doi: 10.1089/cmb.2010.0118.

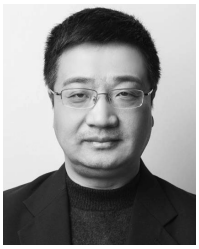
[4] C. J. Fang, F. J. Shao, Y. Sui and R. C. Sun, "Analysis of meteorological network based on complex network," *J. Qingdao Univ.*, vol. 30, no. 4, pp. 50–57, Nov. 2017.

[5] U. K. von Schwedler, M. Stuchell, B. Müller, D. M. Ward, H.-Y. Chung, E. Morita, H. E. Wang, T. Davis, G.-P. He, D. M. Cimbora, A. Scott, H.-G. Kräusslich, J. Kaplan, S. G. Morham, and W. I. Sundquist, "The protein network of HIV budding," *Cell*, vol. 114, no. 6, pp. 701–713, Oct. 2003. doi: 10.1016/s0092-8674(03)00714-1.

[6] C.-J. Fang, F.-J. Shao, W.-P. Zhou, C.-X. Xing, and Y. Sui, "Construction and analysis of meteorological elements correlation network," presented at the Int. Symp. Neural Netw., Muroran, Japan, Jun. 2017.
[7] L. Zhou, R. Zhi, A. X. Feng, and Z. Q. Gong, "Topological analysis of temperature networks using bipartite graph model," *Acta Phys. Sin-Ch Ed.*, vol. 59, no. 9, pp. 6689–6696, Sep. 2010. doi: 10.3724/SP.J.1077.2010.01263.
[8] M. C. Bove, P. Brotto, F. Cassola, E. Cuccia, D. Massabò, A. Mazzino, A. Piazzalunga, and P. Prati, "An integrated $PM_{2.5}$ source apportionment study: Positive matrix factorisation vs. the chemical transport model CAMx," *Atmos. Environ.*, vol. 94, pp. 274–286, Sep. 2014. doi: 10.1016/j.atmosenv.2014.05.039.
[9] B. Denby, I. Sundvor, M. Cassiani, P. de Smet, F. de Leeuw, and J. Horálek, "Spatial mapping of ozone and SO_2 trends in Europe," *Sci. Total Environ.*, vol. 408, no. 20, pp. 4795–4806, Sep. 2010. doi: 10.1016/j.scitotenv.2010.06.021.
[10] R. Yu, Y. Yang, L. Yang, G. Han, and O. A. Move, "RAQ—A random forest approach for predicting air quality in urban sensing systems," *Sensors*, vol. 16, no. 1, pp. 86–104, Feb. 2017. doi: 10.3390/s16010086.
[11] L. S. Neal, M. Dalvi, G. Folberth, R. N. McInnes, P. Agnew, F. M. O'Connor, N. H. Savage, and M. Tilbee, "A description and evaluation of an air quality model nested within global and regional composition-climate models using MetUM," *Geosci. Model. Develop.*, vol. 10, no. 11, pp. 3941–3962, Nov. 2017. doi: 10.5194/gmd-10-3941-2017.
[12] H. Zhao, A. Liu, K. Wang, and Z. Bai, "An improved air quality prediction model based on GA-ANN," *Environ. Sci. R*, vol. 22, no. 11, pp. 1276–1281, Nov. 2009. doi: 10.13198/j.res.2009.11.42.zhaoh.008.
[13] P. Perez, "Combined model for PM10 forecasting in a large city," *Atmos. Environ.*, vol. 60, pp. 271–276, Dec. 2012. doi: 10.1016/j.atmosenv.2012.06.024.
[14] M. Wang, Z. Yuan, X. Zhang, D. Zheng, and D. Ji, "Construction of air quality evaluation system based on FCM algorithm and BP neural network," *Agricult. Biotechnol.*, vol. 7, no. 5, pp. 274–276, Jul. 2018. doi: 10.19759/j.cnki.2164-4993.2018.05.072.
[15] Y. F. Wang, Z. F. Liu and J. Xu, "Evaluation and prediction of atmospheric environmental quality by grey system theory," *Environ. Sci. Manage.*, vol. 34, no. 5, pp. 162–166, May 2009. doi: 10.3969/j.issn.1673-1212.2009.05.044.
[16] H. Ding, Y. H. Liu and S. X. Cao, "Study on urban air quality assessment based on fuzzy-grey clustering method," *Environ. Sci. Tech.*, vol. 36, no. 12m, pp. 374–379, Dec. 2013.
[17] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999. doi: 10.1126/science.286.5439.509.
[18] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1999. doi: 10.1038/30918.
[19] X.-K. Zhang, C. Song, J. Jia, Z.-L. Lu, and Q. Zhang, "An improved label propagation algorithm based on the similarity matrix using random walk," *Int. J. Mod. Phys. B*, vol. 30, no. 16, May 2016, Art. no. 1650093. doi: 10.1142/S0217979216500934.
[20] Y. Wu, "Research on key propagation paths and important node mining methods of regional air pollutants," M.S. thesis, Yanshan Univ., Qinhuangdao, China, 2018.
[21] D. J. Robinaugh, A. J. Millner, and R. J. McNally, "Identifying highly influential nodes in the complicated grief network," *J. Abnormal Psychol.*, vol. 125, no. 6, pp. 747–757, Jun. 2016. doi: 10.1037/abn0000181.
[22] C. Song, G. Huang, B. Zhang, J. Ren, and X. Zhang, "A node influence ranking algorithm based on probability walking model," *Int. J. Mod. Phys. B*, vol. 33, no. 13, May 2019, Art. no. 1950132. doi: 10.1142/S0217979219501327.
[23] B. Liu, S. Yan, J. Li, and Y. Li, "Forecasting PM2.5 concentration using spatio-temporal extreme learning machine," presented at the 15th IEEE Int. Conf. Mach. Learn. Appl., Anaheim, CA, USA, Dec. 2016.
[24] L. Li, J. Gong, and J. Zhou, "Spatial interpolation of fine particulate matter concentrations using the shortest wind-field path distance," *PLoS ONE*, vol. 9, no. 5, May 2014, Art. no. e96111. doi: 10.1371/journal.pone.0096111.
[25] J. Ma, J. Fan, F. Liu, and H. Li, "A community detection algorithm based on Markov random walks ants in complex network," *J. Shanghai Jiaotong Univ. (Sci.)*, vol. 24, no. 1, pp. 71–77, Feb. 2019. doi: 10.1007/s12204-019-2041-2.
[26] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 2, Mar. 2004, Art. no. 026113. doi: 10.1103/PhysRevE.69.026113.
[27] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 10, Apr. 2008, Art. no. P10008. doi: 10.1088/1742-5468/2008/10/p10008.

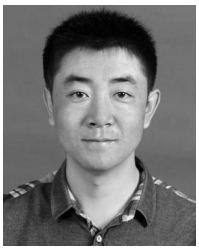
CHEN SONG was born in Shandong, China, in 1988. She received the B.S. and M.S. degrees in computer science and engineering from the Tianjin University of Science and Technology, in 2016. She is currently pursuing the Ph.D. degree in computer science and technology with the School of Information Science and Engineering, Yanshan University, China.

From 2016 to 2018, she was an Assistant Lecturer with the 32153 Troops of the Chinese People's Liberation Army, Hebei, China. Since 2018, she has been a Laboratory Technician with the School of Information Science and Engineering, Yanshan University. Her current research interests include evolution of complex networks, random walk algorithm, and air quality.

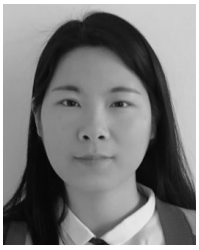


GUOYAN HUANG was born in Heilongjiang, China, in 1969. He received the Ph.D. degree from Yanshan University, China, in 2006, where he is currently a Professor with the School of Information Science and Engineering. His research interests include data mining, temporal data modeling, and software security. His research was supported by the National Natural Science Foundation of China.

BING ZHANG received the bachelor's degree from the College of Computer and Information Technology, Three Gorges University, China, in 2012, and the Ph.D. degree from the School of Information Science and Engineering, Yanshan University, China, in 2018, where he is currently a Lecturer and holds a postdoctoral position. He has ever been with the Norwegian University of Science and Technology as a Visiting Scholar. His research interests include data mining, machine learning, and software security.



BO YIN was born in Shanxi, China, in 1986. He received the B.S. degree in electrical engineering and automation form the Armored Forces Engineering College, in 2013. He is currently an Assistant Lecturer with the 32153 Troops of Chinese People's Liberation Army, Hebei, China. His research interests include data mining and software security.



HUIFANG LU is currently pursuing the M.S. degree in computer science and technology with the School of Information Science and Engineering, Yanshan University, China. She is currently involved in an air quality project. Her research interests include machine learning and data mining.