IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received January 16, 2019, accepted February 18, 2019, date of publication February 22, 2019, date of current version March 12, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2900997*

# Innovative Spatial-Temporal Network Modeling and Analysis Method of Air Quality

**GUYU ZHAO, GUOYAN HUANG, HONGDOU HE, AND QIAN WANG**
College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
Computer Virtual Technology and System Integration Laboratory of Hebei Province, Qinhuangdao 066000, China

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

This work was supported by the National Natural Science Foundation of China under Grant 61772451.

**ABSTRACT** Air quality system is characterized by dynamism, dependency, and complexity. Scientifically representing the internal structure of air mass distribution and its relationship to reveal the dynamic evolution of air quality is the key to solve the air pollution problem. This paper abstracts the air quality system into the complex network innovatively by synthesizing spatial and temporal factors influencing air quality status. Based on quantifying the regional dynamic interconnection and interaction, our modeling approach is proposed to mine the relationship of different regions. First, the dynamic time-varying nature of air pollutant concentration is essential to get the interaction frequency of local air quality in the time dimension. The time correlation analysis of air quality nodes is conducted by calculating the time correlation matrix to construct the air quality network topology. Second, spatial distance and wind are the main factors influencing the diffusion of pollutants, which is used to characterize spatial homogeneity and heterogeneity. By computing the spatial correlation matrix, the spatial interaction intensity is quantified. Then, air quality spatiotemporal model is established by integrating the temporal and spatial correlation. Finally, based on the air quality spatiotemporal network model, community detecting algorithms are used to mine the local similarity and regional interaction. We evaluated our model with extensive experiments based on real data. The results show that our model is dynamic, reliable, and scalable. Utilizing the characteristics of the complex network community, our approach reflects the local and propagating characteristics of air quality and lays the foundation for air pollution prevention and further prediction.

**INDEX TERMS** Air quality system, the complex network, dynamic model, data mining.

## I. INTRODUCTION

The research on air quality has been widely watched in recent years, for air quality is critical to human health and urban governance [1], [2]. At present, various regions in many countries have established a wide range of air quality monitoring systems, resulting in a certain amount of monitoring data. However, there are many difficulties to be solved in order to indicate the air quality more scientifically. Firstly, air quality is affected by multiple factors [3]–[7], such as pressure, temperature, humidity, rainfall, illumination, etc. These factors will affect each other or produce physical and chemical reactions, making the analysis of air quality more dynamic, variable and complex. Considering the complicated data collection of these factors, studies need to be conducted to identify factors that play major roles in the evolution of air quality and should be further used rationally. Secondly, the amount of existing air quality monitoring stations is not substantial in a city due to the expensive cost of building and maintaining such a station [3], so that air quality monitoring data should be perceived properly before effective application. Finally, basic theoretical guidance and core technical support of dynamic analysis of air quality are still lacking, leading to the accuracy of air quality analysis difficult to guarantee. Due to the dynamism and complexity, a reliable model is needed to accurately characterize the distribution structure and evolutionary behavior, with the purpose of detecting the regional evolution trend of air quality and the interaction of pollutants between regions. According to data analysis, air quality evolution often follows some regularity with the development of time and space. Therefore, air quality analysis model should be established taking both spatially-related features and temporally-related features into account [3], [4], [8]. From the angle of time, air quality status in a

The associate editor coordinating the review of this manuscript and approving it for publication was Chun-Wei Tsai.

given area usually has no significant change over a consecutive period. Meanwhile, adjacent areas often appear similar variations in the spatial dimension. Accordingly, relying solely on temporal or spatial dimension cannot achieve the realistic requirement of air quality analysis. On the one hand, local air quality of geographically adjacent areas is possibly quite different, which means that small distance is not the sole criterion to estimate the interactive correlation among nearby areas. On the other hand, locations that are very far away from each other, with similar varying law of air quality in the same period, can hardly be related by cross-domain diffusion. Consequently, the key to realize accurate analysis of air quality by establishing an effective model is that scientifically representing the main influencing factors of spatial and temporal dimensions and characterizing complex spatial and temporal correlations.

The complex network model has led to a series of important discoveries in recent years [9], [10], [12], [13], which has played a significant role in the analysis of real world problems. Many complex systems in the real world either exist in the form of complex networks or can be transformed into complex networks [9], [11], such as interpersonal networks, scientists' cooperation networks and epidemiological networks in the social system, neural network, gene regulation network and protein interaction network in ecosystems. More complicated topology than the regular network, complex network in the real world possesses some basic statistical properties. The small world effect reflects the characteristic of complex networks with short path length and high clustering coefficients [12]. The characteristic of node degrees obeying power law distribution in the complex network is known as scale-free property [13]. The feature, that the nodes inside the community are closely connected and the nodes of different communities are less connected, always exists in complex network systems [9]. Complex network has become a powerful tool for systematic science, complexity science and statistical research. At present, researches of air quality system have not applied the complex network theory to establish model. Therefore, this paper uses the complex network to characterize the system in both spatial and temporal dimensions.

Note that economic factors, meteorological factors, geographical factors and their interactions form the complex relationships and constraints of air quality spatial-temporal surroundings. At the same time, regional air quality status shows the network topology dependency characteristic of dynamic interconnection and interaction. Therefore, this paper describes a novel complex network based air quality spatial-temporal dynamic model to dynamically measure spatial-temporal distribution and interaction.

The main contribution of this paper lies in the following three aspects:

(1) Spatiotemporal relations are illuminated by identifying temporal-related and spatial-related features to measure air pollution degradation and propagation. Temporal correlation is obtained by evaluating the pollutant evolutionary similarity of different locations. The spatial dependence is computed by assessing the propagation intensity of local sites using standard Gaussian diffusion model. Instead of treating the temporal correlation and spatial constraint separately, this paper combines them in a complex network for collaborative analysis.

(2) We propose a novel modeling method mainly based on the complex network for the mining of regional impacts of air quality. The distribution and regional diffusion of air quality are reasonably represented, and the subjectivity and deviation caused by grid-division or clustering-division methods are modified. By effectively detailing and formalizing the air quality system, regional monitor stations are mapped into nodes of the complex network, and the structure of complex network is constructed according to the temporal and spatial attributes.

(3) We evaluate our model dynamically by hour with real data using three different community detecting algorithms for the first time. According to the analysis of real data, the division results represented by communities well coincide with the criterions of community characteristics and requirements of real application. So, our model demonstrates that the complex network owns the reliability and scalability for better describing and understanding air quality system. By setting up the experiments based on different dates, the regional correlation of different cities and the interaction between them is clarified. Therefore, this model is favorable for accurate prediction of air quality relying on the entire area instead of just one place.

## II. RELATED WORK

Attributed to the large scale of air quality system, it is difficult to achieve an accurate description of the meteorological features, geographical features, economic features and the complex relationship among them during the dynamic evolution process. Accurate characterization and measurement of temporal and spatial distribution and interaction of air quality has become a key issue in the field of air quality research. At present, characterization methods of air quality model include grid-division based analysis method and cluster-division based analysis method.

Grid-division based analysis method which is considered as a traditional division way, partitions the entire area into adjacent grids to advance the spatial interpolation and air quality simulation. Shepard [14] used the grid-division method to achieve two-dimensional spatial interpolation of air quality, assuming that each grid is a unit with uniform air pollutant concentration. Bai *et al.* [15] defined uniform grids to accomplish near surface $PM_{2.5}$ concentration interpolation making use of the satellite observation data. In addition, Tang *et al.* [16] applied the divided grids to carry out the spatial interpolation for the $PM_{2.5}$ concentration, of which the data came from different sources. Zheng Y *et al.* divided the research area into disjoint grids (3km $*$ 3km), and then employed the temporal features and spatial features to conduct spatial interpolation of air quality taking the influence of adjacent grids. Moreover, Goodin *et al.* [17] constructed a three-dimension wind field of the urban area with the use of finely defined grids. Vardoulakis *et al.* [18] took advantage of grid-division method to complete air quality simulation of the street canyon. Besides, Syrakov *et al.* [19] simulated the European air quality through the pre-defined grids. Pisoni *et al.* [20] provided a hierarchical grid-division method, where the granularity changes with the distance, to increase the spatial flexibility of the "source-acceptor" relationship in the air quality model. In terms of air quality prediction, the grid-division based method is mainly used for the forecasting of air quality index called AQI or its main pollutant concentration. Ong *et al.* [7] simply selected *k* nearest neighbours to accomplish the prediction of $PM_{2.5}$, which used a dynamic pre-trained deep recurrent neural network. Eder *et al.* [21] and Byun and Schere [22] both employed the neighboring grids for predicting contaminants concentration. Yu *et al.* [23] constructed prediction model for the concentration of $PM_{2.5}$ based on random forest, utilizing the effect of air quality concentration of adjacent grids. Furthermore, Zheng *et al.* [4] defined a circular-division method, which is similar with the grid-division method, dividing the study region into three concentric circles to predict the AQI. The above analysis based on grid partition is largely influenced by subjective factors and limited by the critical conditions, which can be demonstrated by the empirical process. The grid partition procedure weakens the quality and efficiency of subsequent air quality analysis, because of lacking sufficient consideration of the constraint and correlation in air quality system.

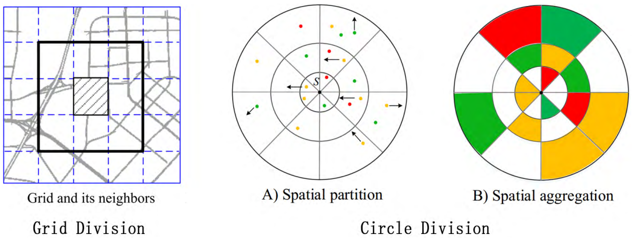
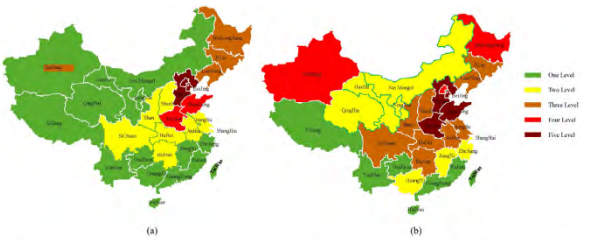


**FIGURE 1.** Grid division diagram.

Cluster-division based method is another pattern for basic air quality research, which exploits clustering algorithms to divide the research area into regions with internal similarity. Jiang *et al.* [24] used the BP neural network to predict the air quality based on the clustering result of available pollutants concentration. Reyes and Sánchez [25] enhanced the efficiency of clustering based on genetic algorithm for the following analysis of air quality. Sefidmazgi *et al.* [26] applied bounded variation clustering based on testing changes in air quality temporal series. Chen *et al.* [27] took k-means algorithm on the decomposed original temporal series to analyze the air quality evolution. Similarly, Austin *et al.* [28] employed k-means algorithm on the components of $PM_{2.5}$ to investigate the air quality situation in the US. Moreover, Liu *et al.* [8] did research on the prediction of $PM_{2.5}$ concentration using Geo-SOM cluster algorithm. Above methods can overcome partial influence of anthropic factor, but also ignore theoretical analysis of pollutant propagation between regions. Clustering algorithms often simply consider temporal features, which will result in inaccurate analysis and forecasting of air quality.



Clustering results based on air quality databases. (a) EPLS clustering result; (b) AQI clustering result.

**FIGURE 2.** Cluster analysis diagram.

In this paper, a model based on the complex network called air quality spatial and temporal network is proposed with a fresh perspective, abstracting regional air quality distribution into a complex network and scientifically representing the spatial and temporal correlation of air quality characteristics. Our work composites the temporal and spatial features, illuminating the relationship and effect intensity of different sites and providing a basis for further analysis and forecasting. This model will generate a reliable partition result called community set by dividing the network into communities to complete the analysis of air quality dynamic evolution according to the network connectivity.

## III. AIR QUALITY SPATIO AND TEMPORAL RELATIONSHIP ANALYSIS

In this section, reasonable analysis of air quality temporal and spatial features and their relations will be given, which is the basic theorization process of our work.

Inside the air quality system, geographical factors, meteorological factors and economic factors always have influence or interaction between each other, making the air quality analysis complicated. Among these factors, we consider air mass concentration as the main temporal factor that will reflect the correlations between sites. Geological features (such as site position and distance) and meteorological characteristics (such as wind direction) are regarded as spatial restraint to quantify the pollutant dispersion. The multidimensional characteristics of air quality system with complex temporal correlation and spatial restraint form the complexity and polymorphism of air quality evolution.

Obviously, air quality system covers a collection of monitoring sites, reporting the AQI or main pollutant concentration. Those sites combined with pollution transmission path can form an air quality network. Spatial interpolation algorithms based on detected data are usually applied to forecast air quality of areas possessing no monitoring stations in air quality network. This process often produces some errors and bias making the establishment of model imprecisely. Therefore, we construct an air quality network model, which utilizes the real data from monitoring sites without considering the spatial interpolation. Our model derives from the graph theory by mapping the monitoring sites into nodes of complex network and establishing the edges between nodes according to temporal and spatial correlations. The monitoring sites record the concentration by hour, resulting in a hierarchical model as shown in figure 3, of which each layer represents a circumstance of air quality at a given time.

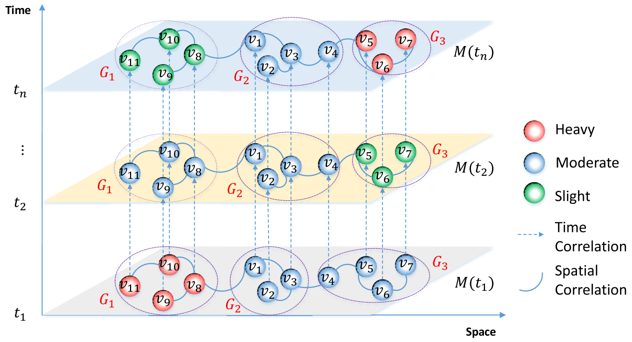


**FIGURE 3. Air quality spatial and temporal relation analysis diagram.**

The way that influences the air quality status of a place is usually categorized into two types, which can be termed as local pollution and propagation pollution. Local pollution means that the pollutants produce and disappear in a small range, which is mainly caused by the emission of local sources like automobile exhaust or industrial gas and can be dismissed by dilution settlement. Propagation pollution stands for the regional interaction of pollutants through the spreading process, which is influenced by geographical condition and meteorological situation. Besides, local pollution and propagation pollution are associated with each other. Propagation pollution is an external cause, which will bring contaminants from outside places influencing the regional air quality to different extent. Meanwhile, local pollution regarded as the internal factor of air quality can also be reduced by pollution diffusion, during which the pollutants are propagated to other places due to the effect of wind or other factors. Generally, downwind direction and lesser distance between two independent sites can make the propagation of pollutants easy.

Under the effect of local pollution and propagation pollution, air quality shows local similarity in a region of a certain size, which will form a distinct local structure called local air quality (LAQ). LAQ is composed of several regional sites, of which the air quality status and changing trends are very similar. The air quality network can be divided into several areas, according to tight interaction inside LAQ and limited reciprocal effect between LAQs. The division result can be detected by community mining algorithms in our work, which abstract LAQ into community structure in the complex network.

We construct a reasonable model for researching the correlation inside LAQ and distinction between LAQs, which involves the analysis of temporal features and spatial features to provide the basis of air quality study. The theoretical analysis process as shown in Figure 3 is detailed in the following description, which gives some concepts.

### A. REGIONAL AND LOCAL AIR QUALITY CHARACTERISTICS

During a certain period, the regional air quality (such as Beijing-Tianjin-Hebei region) appears internal polymorphism characteristic influenced by geographical factors, meteorological factors, economic factors and interactions between them. Thus, the regional area can be divided into several inner-similar local structures that are LAQs. As shown in Fig. 3, there are three LAQs represented as $\{G_1, G_2, G_3\}$ in the regional air quality $M(t_n)$ at time $t_n$:

$$M(t_n) = G_1\{v_8, v_9, v_{10}, v_{11}\} \cup G_2\{v_1, v_2, v_3, v_4\} \cup G_3\{v_5, v_6, v_7\} \quad (1)$$

There, $v_i$ stands for regional nodes (monitor stations), $G_i$ stands for the LAQ, and $M(t_n)$ respects the regional air quality at time $t_n$.

### B. TEMPORAL CORRELATION OF AIR QUALITY

Considering that air pollutant concentrations of regional nodes change with time, the temporal correlations between nodes can be expressed by the similarity of pollutants concentration and its variation trends over a certain period, which is caused by the interaction intensity between nodes. Thus, the basis for realizing the temporal correlation between regional nodes is computing correlations of pollutants concentration sequences. If concentration sequences of pollutants corresponding to different nodes have strong similarity, the temporal correlation between nodes is strong. Otherwise, the temporal correlation between nodes is weak. As shown in Fig. 3, it can be seen that the concentration of node $v_5$ and node $v_7$ at $t_n$ and their variation trends are very similar, which indicates that these two nodes have strong temporal correlation. Evidently, nodes with strong correlation are easy to form a LAQ.

We assume that $Q$ represents the collection of contaminant species in the model at time $t$.

$$Q(M(t_n)) = \{PM_{2.5}, PM_{10}, SO_2, NO_2, CO, O_3\} \quad (2)$$

### C. SPATIAL DEPENDENCE OF AIR QUALITY

Meteorological factors (such as wind direction) and geographical factors (such as distance) are important constraints for the spatial dependence analysis of regional nodes. The spatial dependency is caused by the interaction of pollutants between nodes, while the strength of air mass interaction is obtained by the cost of contaminant diffusion. According to the Gaussian propagation model, propagation cost of pollutants is computed mainly based on the distance of nodes pair and the wind direction. For example, when the distance between nodes $v_\mathrm{i}$ and $v_\mathrm{j}$ is large, or $v_\mathrm{j}$ is not within the range of wind direction of $v_\mathrm{i}$, the spatial interaction between them is weak. Under this circumstance, the propagation cost of these two nodes is expensive, which decreases the spreading effect. On the contrary, small distance and appropriate wind direction are attributable to strong interaction between nodes. Thus, spatial dependence of air quality is another important constraint on the formation of LAQs.

### D. AIR QUALITY DISTRIBUTION AND DYNAMIC EVOLUTION

Temporal correlation and spatial constraint of regional air quality can reflect or establish the distribution of air quality and similarity of LAQs over different periods. As shown in Figure 3, under the influence of spatial-temporal relations, the air quality is developed as $M(t_1) \rightarrow M(t_2) \rightarrow \cdots \rightarrow M(t_n)$.

By learning spatial and temporal association of regional air quality, a complex network based model is established, whose main goal is rationally characterizing the distribution and interaction process of air quality and supporting further effective analysis.

## IV. AIR QUALITY SPATIAL AND TEMPORAL NETWORK MODEL

According to the above analysis of air quality temporal and spatial relations, establishment process and analysis procedure of air quality spatial and temporal network model will be detailed in this part as shown in Figure 4, which consist of three phases.

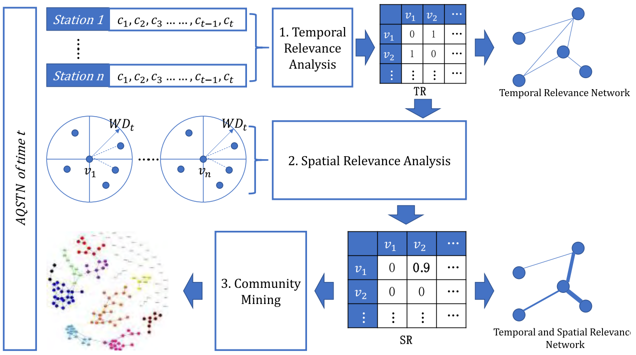

AQSTN of time t
Station 1 $c_1, c_2, c_3 \dots \dots, c_{t-1}, c_t$
Station n $c_1, c_2, c_3 \dots \dots, c_{t-1}, c_t$
1. Temporal Relevance Analysis
$v_1$ $v_2$ ⋯
$v_1$ 0 1 ⋯
$v_2$ 1 0 ⋯
TR
Temporal Relevance Network
$WD_t$
$WD_t$
$v_1$
$v_n$
2. Spatial Relevance Analysis
3. Community Mining
$v_1$ $v_2$ ⋯
$v_1$ 0 0.9 ⋯
$v_2$ 0 0 ⋯
SR
Temporal and Spatial Relevanc
Network

**FIGURE 4.** Air quality spatial and temporal network analysis framework.

The first phase is constructing the topology of air quality spatial and temporal network by mapping regional monitor stations in air quality system into nodes of a complex network with temporal and spatial attributes. Additionally, given the nodes distribution, edges in the network topology are created based on an adjacency matrix. The temporal correlation matrix is computed based on concentration sequences, which can be converted to the adjacency matrix, to store temporal relevance degrees of node pairs in nearest historical period. Secondly, considering the effect of spatial homogeneity and heterogeneity of air quality, spatial correlation intensity is calculated concerned with geography and meteorological factors, in hope of allocating weights of edges in the network topology. In this context, the spatial correlation matrix is constructed to evaluate the spatial constraint for each pair of nodes. Finally, the network is divided into clusters using corresponding community mining algorithms. Based on the mining results, the validation of the air quality spatial and temporal network model is undertaken with the characteristics of community structure, which proves the model credible and scalable.

### A. BASIC DIFINITIONS OF AIR QULITY NETWORK

In this section, basic definitions related to air quality spatial and temporal network are given. Node set and edge set present interconnection and interaction in air quality network. Besides, the air quality spatial and temporal network model is constructed dynamically, because the air pollutant concentration is recorded by hour.

*Definition 1 [Air Quality Spatial and Temporal Network (AQSTN)]:* AQSTN is an weighted undirected network, which represents the distribution and association of regional air quality. Depending on air quality temporal correlations and spatial constraints, we define air quality spatial and temporal network as follows.

$$AQSTN(t_n) = \{V, E, W, Q | TR \cap SR\} \tag{3}$$

In *AQSTN*, node set $V$ represents the regional nodes of different geographical areas. Edge set $E$ represents edges between different air quality regional nodes. The set $W$ denotes interaction intensity (edge weight) between nodes. Besides, set $Q$ denotes pollutant species to be analysed for node set. $TR \cap SR$ denotes the constraint condition under the impact of air quality temporal and spatial factors.

Node set $V$, the set of air quality nodes, can be defined as follows.

$$V = \{v_1, v_2, \dots, v_i, \dots, v_m\} \tag{4}$$

$$v_i = \langle t, Lng, Lat, wd_t, c_t \rangle \tag{5}$$

where $t$ is the current time, $m$ is the total number of nodes, $Lng$ is the longitude of node $i$, $Lat$ represents the latitude of node $i$, $wd_t$ is the wind direction of node $i$ at time $t$, $c_t$ is the contaminant concentration of node $i$ at time $t$.

Edge set $E$ describes the propagation and interaction relationship of air quality between nodes in a node set $V$.

$$E = \{e_{ij} | i, j = 1, 2, \dots, m\} \tag{6}$$

$$e_{ij} = \langle v_i, v_j \rangle \tag{7}$$

*Definition 2 [Air Quality Temporal Correlation (TR)]:* Air quality temporal correlation indicates temporal interaction intensity between air quality regional nodes for a certain period of time, which is defined as follows.

$$TR = \begin{pmatrix} r_{11}\langle C_1, C_1 \rangle & \cdots & r_{1m}\langle C_1, C_m \rangle \\ \vdots & \ddots & \vdots \\ r_{m1}\langle C_m, C_1 \rangle & \cdots & r_{mm}\langle C_m, C_m \rangle \end{pmatrix} \tag{8}$$

*TR* is the temporal correlation matrix, and $r_{ij}$ represents the partial correlation coefficient of air quality contaminant concentration temporal sequences corresponding to nodes $v_i$ and $v_j$, reflecting the degree of correlation between the two variables. Pollutant concentration temporal sequence characterizes the evaluation of pollutant concentration changing over time in a given period. The representation is as follows.

$$C_i = \{c_1, c_2, \ldots, c_{t-1}, c_t\} \tag{9}$$

Set $C_i$ donates the vector of contaminant concentration temporal sequence of node $i$. $c_t$ is the concentration of pollutant at time $t$ where the granularity is set to one hour.

*Definition 3 [Air Quality Spatial Correlation (SR)]: SR* stands for the spatial interaction intensity between nodes, which is measured by the standard model called Gauss diffusion model, under the effects of spatial distance and wind field. It is defined as follows.

$$SR = \begin{pmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mm} \end{pmatrix} \tag{10}$$

Here, $SR$ is the spatial correlation matrix. Element $w_{ij}$ in the matrix represents the degree of spatial correlation between nodes $v_i$ and $v_j$. Calculation of spatial correlation mainly considers spatial distance and wind direction angle.

The spatial distance ($Dist$) is presented as spatial distance matrix.

$$Dist = \begin{pmatrix} d_{11} & \cdots & d_{1m} \\ \vdots & \ddots & \vdots \\ d_{m1} & \cdots & d_{mm} \end{pmatrix} \tag{11}$$

Element $d_{ij}$ represents geographical distance between nodes $v_i$ and $v_j$.

The wind direction angle ($\Theta$) holds angles between wind direction $WD_t$ and the edge of node pair.

$$\Theta = \begin{pmatrix} \theta_{11} & \cdots & \theta_{1m} \\ \vdots & \ddots & \vdots \\ \theta_{m1} & \cdots & \theta_{mm} \end{pmatrix} \tag{12}$$

Notation $\theta_{ij}$ represents the magnitude of the wind direction angle corresponding to nodes $v_i$ and $v_j$.

### *B. AIR QUALITY AND TEMPROAL NETWORK ESTABLISHMENT*

At present, in China's Beijing-Tianjin-Hebei and Yangtze River Delta and other regions, wide range and multi-level air quality monitoring system has been constructed, which produces a large amount of data with various types and complex structures. By resolving regional air quality monitoring data and meteorological data, temporal evolution sequences and spatial characteristic matrices are acquired. After the foundation of air quality system dynamic topology based on temporal correlation analysis and weight allocation measured by spatial interaction, we establish the air quality spatial and temporal network, which is analyzed with community detecting algorithms. The specific process is shown in Figure 5.

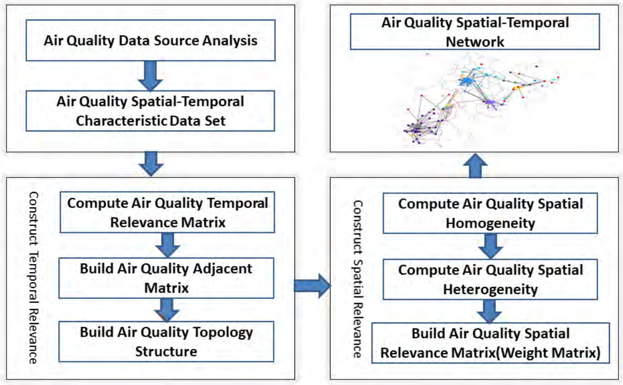

Air Quality Data Source Analysis
Air Quality Spatial-Temporal Characteristic Data Set
Construct Temporal Relevance
Compute Air Quality Temporal Relevance Matrix
Build Air Quality Adjacent Matrix
Build Air Quality Topology Structure
Air Quality Spatial-Temporal Network
Construct Spatial Relevance
Compute Air Quality Spatial Homogeneity
Compute Air Quality Spatial Heterogeneity
Build Air Quality Spatial Relevance Matrix(Weight Matrix)

**FIGURE 5. Air quality spatial and temporal network construction flow chart.**

#### 1) ANALYSIS OF REGIONAL AIR QUALITY

The data to be investigated consists of monitoring data, for example, concentration of $PM_{2.5}$, $PM_{10}$, $O_3$, $SO_2$, $NO_2$, CO, and so on, and geological data, such as longitude, latitude, altitude, topography, and so forth, and meteorological data, for example, wind direction, wind speed, temperature, humidity, pressure, rainfall, etc. These data can be classified as temporal and spatial characteristic data. For the temporal dimension, similarity of pollutant concentration sequences can be used to characterize the temporal correlation between two regional nodes. From the spatial perspective, in addition to the influence of geographical distance, the influence of meteorological factors should be taken into consideration. Through the analysis of associations between various types of meteorological factors and air quality evolution, the wind is proved to be the main factor that affects the propagation of pollutants.

#### 2) ESTABLISHMENT OF AIR QUALITY TEMPORAL CORRELATION

By calculating the temporal correlation matrix and adjacency matrix, temporal correlation is considered to construct topological structure of air quality.

##### *a: CALCULATION OF THE TEMPORAL CORRELATION MATRIX*

To calculate $TR$, we first need to establish pollutant concentration temporal sequence $C_i$. Assuming that $t$ is the reference time to be analysed, the temporal sequence $C_i$ is generated by selecting continuous hourly concentration data and recorded as $C_i = \{c_1, c_2, \ldots, c_{t-1}, c_t\}$, where $t = T_0$ is the length of the time period. The pollutant concentration temporal sequences are generated for each nodes, for example, the temporal sequence of node $v_i$ at time $t$ is $C_i = \{c_1, c_2, \ldots, c_{t-1}, c_t\}$, where $c_1$ corresponds to the pollutant concentration at time $(t - T_0 + 1)$. The contaminant concentration temporal sequences of all sites constitute the set $C = \{C_1, C_2, \ldots, C_m\}$, where m is the total number of air quality monitor stations. To ensure data integrity and reliability, we set $T_0 = 72$.

The following step is calculating the temporal correlation matrix *TR* based on the established contaminant temporal sequence set *C*. This process obtains temporal correlation by calculating the partial correlation coefficient of temporal sequence vectors of node pairs, and records in the corresponding element of matrix *TR*.

For example, the temporal correlation $r_{ij}$ of $v_i$ and $v_j$ is acquired by calculating the partial correlation of $C_i$ and $C_j$. The partial correlation coefficient of pollutant concentration temporal sequence set ($C$) is calculated as follows.

$$r_{ij(k)} = \frac{r_{ij} - r_{ik} r_{jk}}{\sqrt{1 - r_{ik}^2}\sqrt{1 - r_{jk}^2}} \tag{13}$$

The matrix (*TR*), can indicate the temporal correlation intensity, which reflects the similarity and interaction frequency between each pair of nodes. Algorithm 1 gives the calculation process of *TR*.

**Algorithm 1** Air Quality Temporal Correlation Algorithm (ComputeTR)

```
Input: time t, node set V
Output: The time correlation matrix TR at time t
initialize $C_i$, $C_j$, TR
for each $v_i$ in $V$, do
        $C_i$ ←selectTimeSeries(t-T_0,t)
        for each $v_j$ in $V$, do
                $C_j$←selectTimeSeries($t - T_0$, t)
                $TR.r_{ij}$←
                computeTemporalRelevance($C_i$, $C_j$)
        end
end
    return TR
end
```

### b: CONSTRUCTION OF AIR QUALITY ADJACENCY MATRIX

Examining the temporal correlation matrix *TR*, the adjacency matrix is further constructed by 0-1 processing. In this process, the appropriate threshold $\rho$ is defined depending on the degree of temporal correlation. The 0-1 process deals with all elements $r_{ij}$ in the traversal matrix *TR*, assigning the element value greater than $\rho$ to 1, otherwise 0.

$$\begin{cases} s_{ij} = 1, & if\ r_{ij} \geq \rho \\ s_{ij} = 0, & if\ r_{ij} < \rho \end{cases} \tag{14}$$

After that, an adjacency matrix can be obtained.

### c: ESTABLISHMENT OF AIR QUALITY TOPOLOGY

In this part, we construct topological relations of the network based on the adjacency matrix. When the temporal correlation is greater than the threshold $\rho$, it indicates that the nodes pair has strong correlation, which can establish an edge between nodes. On the contrary, lower temporal correlation indicates that the association between the pair of nodes is weak, which is insufficient to create an edge. Therefore, if $s_{ij} = 1$, an edge $e_{ij} \in$ E is established between the nodes $v_i$ and $v_j$, otherwise $s_{ij} = 0$, no edge is founded between the corresponding nodes. Traversing the adjacency matrix to create all the edges can build the air quality spatial and temporal network topology, as shown in Figure 6.

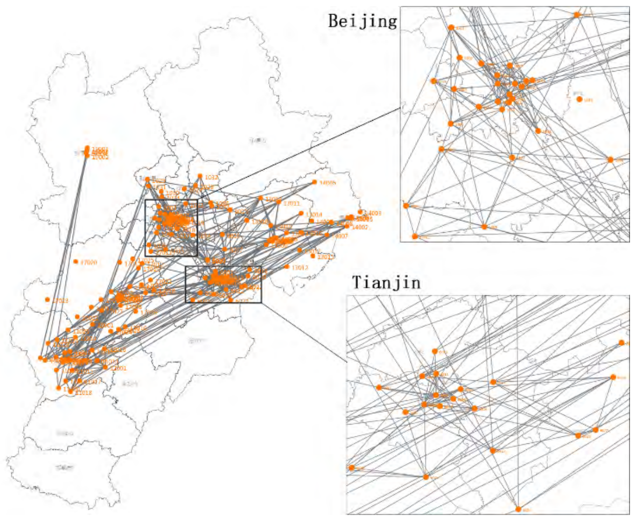


FIGURE 6. Air quality spatial and temporal network topology.

From the perspective of spatial dimension, the possibility of pollutants interacting between exceedingly distant nodes is very small. Thus, solely depending on the temporal correlation *TR* of air quality will result in cross-domain connections, which do not conform to the actual air quality environment. Consequently, the air quality system must be analyzed in conjunction with the spatial correlation *SR* of air quality to confine the occurrence of cross-domain phenomena.

## C. ESTABLISHMENT OF AIR QUALITY SPATIAL CORRELATION

This part discusses air quality spatial homogeneity and heterogeneity to compute the spatial correlation matrix *SR*. Based on *SR*, air quality spatial dependence is measured for the air quality network topology weight allocation.

### 1) CALCULATION OF AIR QUALITY SPATIAL HOMOGENEITY

Spatial homogeneity is presented through the geographic position relation between geographical nodes. Nodes of air quality spatial and temporal network have attributes *Lng* and *Lat*, which represent the latitude and longitude respectively. In our work, spatial homogeneity is given in distance, suggesting that larger distance between sites means larger propagation cost. *Haversine* spherical distance calculation method is utilized to form the distance matrix *Dist* in this paper, formulated as follows.

$$hav\left(\frac{d_{ij}}{r}\right) = hav\left(Lat_j - Lat_i\right) + cos\left(Lat_i\right) cos\left(Lat_j\right) hav\left(Lng_j - Lng_i\right) \tag{15}$$

Here, $hav\left(\beta\right) = sin^2\left(\frac{\beta}{2}\right) = \frac{1-cos(\beta)}{2}$, $d_{ij}$ is the spherical distance between two points, $r$ denotes the spherical radius, $Lat_i$ and $Lat_j$ are the latitudes of $v_i$ and $v_j$, $Lng_i$ and $Lng_j$ are the longitudes.

### 2) CALCULATION OF AIR QUALITY SPATIAL HETEROGENEITY

In general, simply relying on the distance to define the weight cannot effectively reflect properties of pollutant transmission. Since propagation of pollutants is mainly affected by the influence of wind, spatial heterogeneity of air quality is expressed by the wind angle.

Based on the above mentioned, we generate the wind direction angle $\theta_{ij}$, calculating the angle between wind direction $wd_t$, which is dynamically changing, and the edge direction $\overrightarrow{v_iv_j}$ between $v_i$ and $v_j$. On this basis, the wind angle matrix $\Theta$ is constructed. The wind direction angle is formulated as below.

$$\theta_{ij} = \left|wd_t - \overrightarrow{v_iv_j}\right| \tag{16}$$

where $\theta_{ij}$ denotes the wind direction angle between node $i$ and node $j$, and $wd_t$ is the wind direction of the node $i$ based on the north direction, and $\overrightarrow{v_iv_j}$ is the edge direction from node $v_i$ to node $v_j$ base on the north direction. As shown in Fig. 7, $N$ represents the northward direction. Suppose the wind direction $wd_t$ of $v_i$ is NE direction that is $45^o$, $\overrightarrow{v_iv_j}$ is $60^o$, then the wind direction angle $\theta_{ij} = 15^o$. Since the data of wind is collected by hour, the model is dynamically changing hourly.

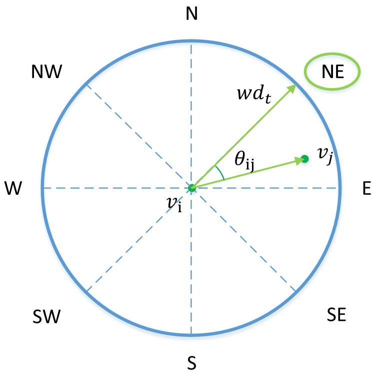


**FIGURE 7.** Calculation of wind direction angle.

### 3) CONSTRUCTION OF THE SPATIAL CORRELATION MATRIX (WEIGHT MATRIX)

Based on the analysis of spatial homogeneity and heterogeneity, the spatial correlation between air quality stations is negatively correlated with the distance $d_{ij}$ and the wind angle $\theta_{ij}$, which is denoted as the intensity $w_{ij}$. For instance, the impact of node $v_j$ on $v_i$ can be ignored, if $v_j$ is far away from the node $v_i$ or has a large wind direction angle between them. Consequently, it is necessary to set the distance threshold $c$ and the wind direction angle threshold $\varphi$ to cut edges with large distance or large wind direction angle between nodes. According to the influencing scope of wind, we set $\varphi = 90^o$ and give $c$ in section six.

After that, we establish the matrix $SR$ by analysing the spatial correlation for each nodes, which is executed according to the standard model - Gaussian diffusion model in the wind field to calculate the diffusion cost, called Gaussian vector weight (GVW). Considering the influence of distance and wind angle on the spatial propagation of pollutants, each element in SR matrix represents the degree of spatial correlation between a node pair. The basic formula of the Gaussian diffusion model [29] is as follows.

$$C_0\left(x, y, z, u\right) = \frac{Q}{\pi u\sigma_y\sigma_z}exp(-\frac{y^2}{2\sigma_y^2} - \frac{z^2}{2\sigma_z^2}) \tag{17}$$

where $x$ represents the downwind distance, and $y$ represents the horizontal distance between the point of departure and the centerline of wind direction. The element $z$ denotes the height of the pollutant release point. $C_0\left(x, y, z, u\right)$ represents air pollutant concentration. The symbol $u$ is the horizontal wind speed. Notations $\sigma_y$ and $\sigma_z$ represent the diffusion standard deviation in the horizontal and vertical directions, respectively.

In order to simplify the basic formula, Gaussian vector weight (GVW) is defined by combining the influence of wind direction and distance based on Gaussian kernel function, which is consistent with the propagation process of the air pollutant. GVW that is the weight of an edge is defined as follows.

$$\begin{aligned} &w_{ij}\left(d_{ij}, \theta_{ij}(t)|c, k\right) \\ &\quad= \begin{cases} e^{-\frac{d_{ij}^2*sin^k\theta_{ij}(t)}{2c^2}} & if\ 0^o \leq \theta_{ij}\left(t\right) \leq 90^o and\ if\ d_{ij} \leq c, \\ 0 & if\ 90^o < \theta_{ij}\left(t\right) \leq 180^o and\ if\ d_{ij} > c. \end{cases} \end{aligned} \tag{18}$$

where $d_{ij}$ and $\theta_{ij}$ represent the distance variable and the angle variable, respectively. Two conditional parameters $k$ and $c$ are used to control the influence degree of wind direction and distance. The function value ranges from 0 to 1.

The weight, which decreases with the increase of distance and reduces with the raise of wind direction angle, is obtained by spatial correlation analysis reflecting the interaction possibility and intensity between nodes. Algorithm 2 gives the air quality spatial correlation calculation process.

Subsequently, the spatial correlation matrix is utilized to assign weights of edges, accomplishing the construction of AQSTN. AQSTN integrates both temporal features and spatial features, of which the node pair frequently interacting during historical period establishes an edge. After that, appropriate weights are allocated for the edges according to pollutant propagation cost. Larger weight represents easier diffuse of pollutants, on the contrary, smaller weight means more difficult to spread contaminants.

## D. ESTABLISHMENT OF AIR QUALITY SPATIAL AND TEMPORAL NETWORK

Through the calculation and analysis of TR and SR, AQSTN is established. Accordingly, the community structure of it is mined using community mining algorithms.

**Algorithm 2** Air Quality Spatial Correlation Algorithm (ComputeSR)

```
Input: time $t$, node set $V$, time correlation matrix $TR$, time correlation critical value $\rho$
Output: The spatial correlation matrix $SR$ at time $t$
initialize $Dist$, $\Theta$, $k = 2$, $c$
$Dist \leftarrow computeDist(V, t)$
$\Theta \leftarrow computeAngle(V, t, TR)$
$c \leftarrow computeAvg(Dist)$
for each $v_i$ in $V$, do
        for each $v_j$ in $V$, do
                if $TR.r_{ij} > \rho$ & $0 <= \Theta.\theta_{ij} <= 90$
        $SR.w_{ij} \leftarrow computeSR(Dist.d\_ij, \Theta.\theta\_ij, k, c)$
                end if
        end
end
return $SR$
```

AQSTN can form a distinctive community structure, as shown in Figure 8. In this figure, we use different colors to represent different communities, and use straight lines to represent the interactions between nodes. We also detail the distribution of Beijing and Tianjin community structures. Indicated by the division result, the problem of cross-domain phenomenon has been addressed, which means nodes connected together are geographically close to each other.

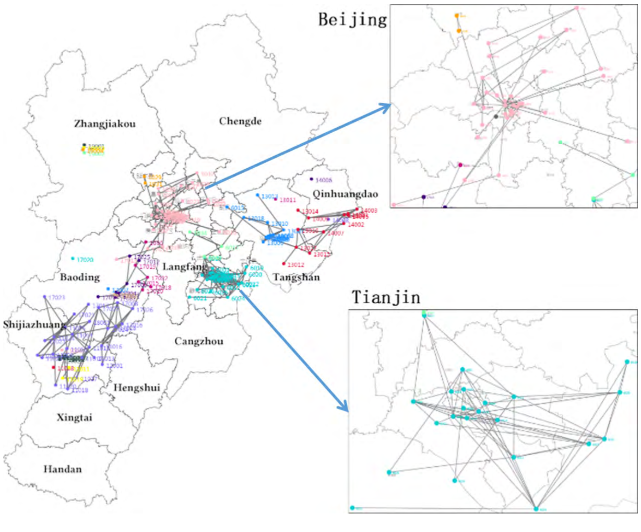


**FIGURE 8.** The model established by Walktrap algorithm for air quality on April 15, 2015.

## V. COMMUNITY EVALUATION CRITERIAS

Clustering coefficient is an important criterion for measuring the small world property of a topological graph. The clustering coefficient of the entire network is defined by Watts and Stogatz [12] as the mean of the local clustering coefficients of all nodes. If the average clustering coefficient of a graph is significantly higher than the random graph generated on the same set of nodes, and the average path length is close to the corresponding random graph, the graph can be considered to be of small world nature. Networks with higher average clustering coefficients are found to have significant module structures. Meanwhile, the average path length of them is usually smaller. The weighted local clustering coefficient of undirected graph is calculated as follows.

**Algorithm 3** Air Quality Modeling and Analysis Algorithm

```
Input: time $t$, node set $V$, time correlation critical value $\rho$
Output: The spatial and temporal network $AQSTN$ at time $t$, community division result $com$
initialize $TR$, $SR$, $AQTN$, $com$
$TR \leftarrow computeTR(V, t)$
$AQTN \leftarrow buildAQTN(TR, \rho)$
$SR \leftarrow computeSR$
$AQSTN \leftarrow buildAQSTN(SR, AQTN)$
$com \leftarrow communityMining(AQSTN)$
return $AQSTN$, $com$
```

$$C_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{w_{ij} + w_{ih}}{2} a_{ij} a_{ih} a_{jh} \tag{19}$$

Symbol $s_i$ is the intensity of node $v_i$, which is the sum of adjacent edge weights. Where $a_{\mathrm{ij}}$ is the corresponding element in the adjacency matrix, $k_i$ is the degree of node $v_i$, and $w_{ij}$ is the weight of edge $e_{ij}$. If the weight of all edges are the same, this formula represents the local transitivity of general unweighted graph.

The average path length is calculated as follows.

$$l = \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} d_{ij} \tag{20}$$

Here, $n$ is the number of nodes, $d_{ij}$ represents the shortest path length between two nodes. The shortest path length $d_{i \to j \to k} = w_{ij} + w_{jk}$, when it takes the weight into account.

A mass of studies focused on the complex network indicate that degree distribution has the characteristic of scale-free. The degree distribution of nodes describes the probability $P(s)$ of a node that possesses the degree of s in the network. Additionally, the probability distribution of node degree generally follows the power law distribution $P(s) \sim s^{-\alpha}$, of which the value of exponent $\alpha$ is usually around 2 [30]. To further study the ability of our approach, an analysis of scale-free feature for AQSTN is performed.

The community structure identified by the algorithms, is assessed using modularity $Q$ [32] that measures the significance of communities due to a selected null model. Let $l_i$ be the community (label) of node $v_i \in N$ and let $A_{ij}$ denote the number of edges incident to nodes $v_i$,$v_j \in$ N. Furthermore, let $P_{ij}$ be the expected number of incident edges for $v_{\mathrm{i}}$, $v_{\mathrm{j}}$ in the null model. The modularity then is defined as follows.

$$Q = \frac{1}{2m} \sum\nolimits_{n_i, n_j \in N} \left(A_{ij} - P_{ij}\right) \delta\left(v_i, v_j\right) \tag{21}$$

where $m$ is the number of edges, that is, $m = |\mathrm{E}|$. The symbol $\delta$ represents the Kronecker function, of which the independent variable (input value) is generally two integers, and if the pairwise numbers are equal, then the output value is 1, otherwise 0. $A_{ij}$ is the weight of edge between node $v_i$ and node $v_j$. $P_{ij} = \frac{k_i k_j}{2m}$, where $k_i = \sum_j A_{ij}$ represents the sum of weight of $v_i (v_i \in N)$, and $m = \frac{1}{2} \sum_{ij} A_{ij}$ represents the weight of all edges in the network.

## VI. EXPERIMENTS AND ANALYSIS

In this section, the dynamism, reliability, scalability and community characteristics of AQSTN model are verified by conducting various experiments on real data.

We evaluate our model with three community detecting algorithms to perform comparative analysis, which take advantage of data of air quality monitor stations in Beijing and Tianjin and the surrounding areas. Meanwhile, a large number of experiments, which consists of 12 months of targeted analysis, show that our experimental results are reducible.

The data set comes from the data in the [3] detailed in Table 1.

**TABLE 1.** Some details of datasets.

| Datasets | | Beijing | Tianjin |
|---|---|---|---|
| Time span | | 2014/5/1-<br>2015/4/30 | 2014/5/1-<br>2015/4/3 |
| Nearby cities | | 14 | 17 |
| AQI | In-city stations | 36 | 27 |
| | In-city instances | 278,023 | 189,604 |
| | Ave $PM_{2.5}$ | 106.4 | 104.3 |
| | Neighbor Sta. | 233 | 267 |
| | #. of instances | 1,272,979 | 1,436,051 |
| Meteorology | In-city sources | 17 | 20 |
| | In-city instances | 17 | 106,614 |
| | Nearby sources | 116,847 | 195 |
| | Nearby instances | 177 | 1,108,873 |
| Forecast | In-city sources | 1,006,814 | 20 |
| | In-city instances | 17 | 361,624 |

Time: 2014-05-01 ~ 2015-04-30;

Data Sheet: City, District, Air quality station, Air quality, Meteorological, Weather forecast;

Major cities: the 31 cities within Beijing, Tianjin and around 300 kilometers;

Air quality data: 220 sites per hour of data;

Real-time atmospheric data: zone level or city level;

Weather forecast data: district level or city level, the next two days, the time granularity of 3 hours, 6 hours, 12 hours.

### A. EXPERIMENTAL PARAMETERS

In the period of temporal correlation analysis, the threshold $\rho$ is to be set. The role of the threshold $\rho$ is to adjust total number of edges in the topology, as excessive edges lead to the existence of edges that reflect very weak correlations between nodes, and few edges remarkably reduce the connectivity of the entire topology. In that sense, we use 0.01 as the granularity to adjust $\rho$ for observing the change of the shortest path length of the topology. By contrast, with the enhancing of $\rho$, the shortest path length increases when the edge number is quite large, and the shortest path length decreases while the edge number reduces to a certain extent. That is to say, the shortest path length will peak with the change of $\rho$. Apparently, this phenomenon shows that the number reduction of edges of the topology at the beginning makes the partition appear gradually, consequently augmenting the length of the path between nodes pair. As the edge number is still decrease, the partition is more obvious, which leads to abnormal topology. This phenomenon weakens the interaction between the clusters greatly, which will affect the community mining results. Additionally, the topology of peak point has already shown abnormity, hence the best status appear before the peak point. By assessing the 12-month data, the peak appears when $\rho$ is between 0.06 and 0.09. Therefore, we choose a smaller $\rho$ value ($\rho = 0.05$). This threshold not only ensures the integrity of the topology, but also helps to complete the deletion of connections between nodes, of which the correlation is too low.

Furthermore, we consider the spatial homogeneity and heterogeneity as the main factors influencing regional diffusion in the spatial correlation analysis. In accordance with the definition of GVW, when the parameter $k$ regarded as a heterogeneous strength variable equals to 0, the GVW acts identically with traditional Gaussian kernel function. Another parameter $c$ is also known as bandwidth with the purpose of adjusting the smoothness of GVW. In our paper, according to the analysis of [8] and [29], the value of $k$ is set to 2, and the value of $c$ is the average of the distance of all edges.

### B. EXPERIMENTAL ANALYSIS

We firstly conduct experiments to investigate the community division results of AQSTN on April 15, 2015 using three community detecting algorithms (Walktrap algorithm [31], FastGreedy algorithm [32], GN algorithm [33]), as shown in Figure 8-10.

By contrast, division results of these three algorithms present generally stable and almost identical. With regard to most areas, such as Beijing, Tianjin, Tangshan and Qinhuangdao, community structures obtained by different algorithms are almost uniform. According to community results obtained by FastGreedy and GN, divisions are in most places the same except for several nodes in the Shijiazhuang area. Obviously, FastGreedy algorithm and GN algorithm divide Shijiazhuang and Baoding into two communities, but Walktrap algorithm generates a major community structure in these two cities.

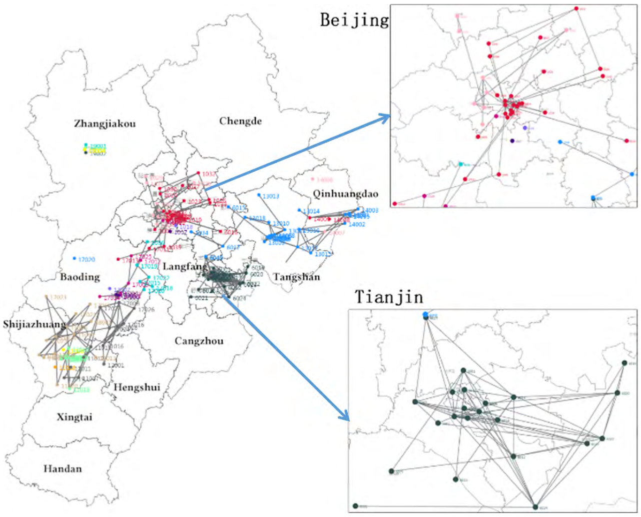


**FIGURE 9.** The model established by GN algorithm for air quality on April 15, 2015.

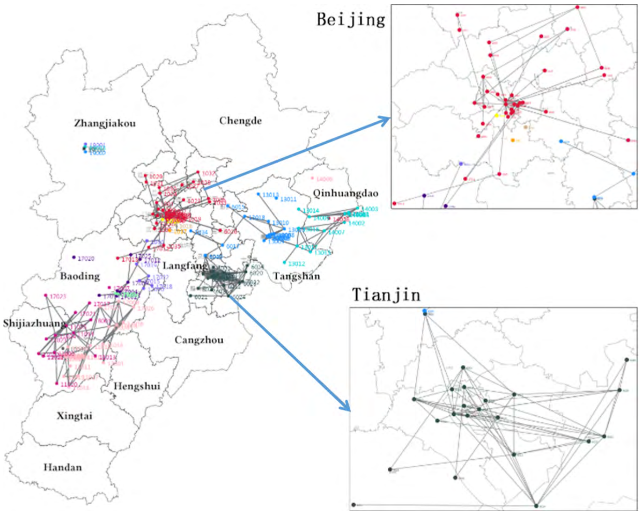


**FIGURE 10.** The model established by FastGreedy algorithm for air quality on April 15, 2015.

Furthermore, stations of Zhangjiakou city studied by three algorithms cannot form a community, which is to say the city shows a discrete state.

Through the analysis of the results, we can find that in April 15, 2015, the spread of air pollutants between Baoding and Shijiazhuang cities is continual. Although the impact between Qinhuangdao and Tangshan cities is also strong, the interaction between two cities comparing with inner propagation of their own, is not enough to form a union community. In addition, the interior correlation *SR* of Zhangjiakou city is stark, but there is no community structure, which indicates that the internal correlation *TR* between nodes is weak.

### C. DYNAMICS AND RELIABILITY ANALYSIS

Figure 11 shows community structure distribution of AQSTNs on November 14, 2014, December 10, 2014, and January 14, 2015, from which we can see that the air quality model that has been built possesses strong dynamics. The size and coverage area of communities, as well as the interaction within and between the clusters of AQSTN model are diverse in different dates. Take the case of Shijiazhuang and Baoding cities, there are four communities in this region on November 14, 2014, then this region is divided into six communities on December 10, 2014, however on January 14, 2015 the distribution contains five distinct communities and some discrete nodes.

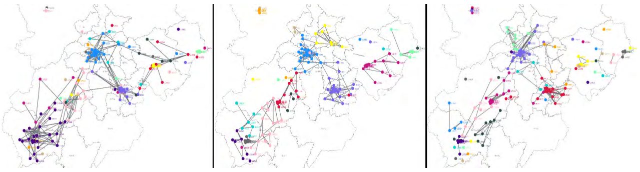


**FIGURE 11.** Community analysis of air quality spatial and temporal networks in November 2014, December 2014 and January 2015.

The $PM_{2.5}$ pollution status on November 14, 2014 and April 15, 2015 is compared with the community distribution of AQSTN on the same date as shown in Figure 12 and Figure 13.

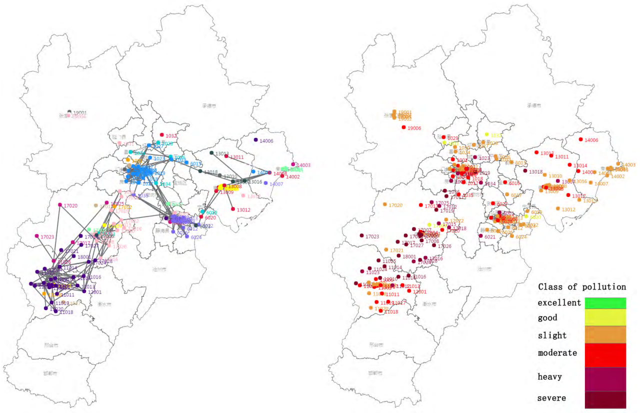


**FIGURE 12.** Comparison of air quality spatial and temporal network community structure vs. pollution status on November 14, 2014.

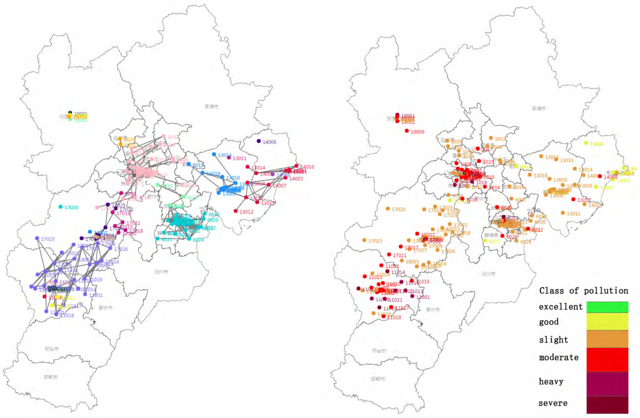


**FIGURE 13.** Comparison of air quality spatial and temporal network community structure vs. pollution status on April 15, 2015.

It is indicated that the community distribution is consistent with the distribution of $PM_{2.5}$ pollution level to a certain extent, which proves that AQSTN we have established is practical and reliable. Pollutant diffusion among adjacent areas on the impact of wind generates similar levels of pollution inside the community, which implies the internal interaction of nodes inside the same community is remarkable. Besides, there also exist edges between nodes belonging to different communities to represent the smaller interaction between the two communities. As a result, their pollution levels are infirmly comparable with certain diversity. Nevertheless, there is some discrimination between community structure and $PM_{2.5}$ distribution situation. As we can see, the pollutant concentrations in the same community of some areas appear different levels. The reason that nodes within the community hold a lower concentration is because pollutants have not reached yet or have been degraded. Moreover, the nodes possess higher concentration within the community describes the phenomenon that internal pollutants have not been degraded in time or proliferated to other regions.

### D. COMMUNITY CHARACTERISTICS ANALYSIS

Some basic statistic criterions of complex network are listed in Table2, where the first column means the dates of generated network model and others are the data obtained based on AQSTN. Remaining columns represent the number of nodes, the number of edges, degree distribution index, average aggregation coefficient, average path length, as well as the average aggregation coefficient and average path length of ER random graphs with the same number of nodes and edges, respectively. The ER random graph is the most used stochastic graph model, whose generation is random connecting a pair of nodes to satisfy the predetermined number of edges.

**TABLE 2. Basic statistics of the 12 months' AQSTN.**

| *date* | *n* | *m* | $\alpha$ | *l* | *C* | $C_{ER}$ | $l_{ER}$ |
|---|---|---|---|---|---|---|---|
| 20140514 | 125 | 238 | 1.493835 | 4.576536 | 0.351378 | 0.04026846 | 3.743569 |
| 20140614 | 127 | 223 | 1.44335 | 3.277295 | 0.496364 | 0.01811594 | 3.716391 |
| 20140720 | 121 | 216 | 1.428676 | 3.019749 | 0.573574 | 0.02685422 | 3.715196 |
| 20140814 | 124 | 162 | 1.46473 | 5.373576 | 0.411857 | 0.01442308 | 4.654873 |
| 20140914 | 91 | 189 | 1.505687 | 1.918327 | 0.588645 | 0.03875969 | 3.199438 |
| 20141031 | 145 | 109 | 1.462985 | 2.692607 | 0.274468 | 0 | 8.596795 |
| 20141114 | 143 | 246 | 1.474372 | 6.313118 | 0.378132 | 0.02054795 | 3.912636 |
| 20141210 | 146 | 251 | 1.409533 | 8.789646 | 0.422449 | 0.01716247 | 4.145376 |
| 20150114 | 147 | 185 | 1.547071 | 3.809253 | 0.496241 | 0.01750973 | 5.058073 |
| 20150214 | 147 | 222 | 1.828628 | 4.64437 | 0.538803 | 0.03065693 | 4.377031 |
| 20150314 | 147 | 194 | 1.43558 | 4.217323 | 0.408284 | 0.01821862 | 5.089072 |
| 20150415 | 149 | 249 | 1.401142 | 5.963319 | 0.321521 | 0.01694915 | 3.929047 |

According to Table 2, the average path length of AQSTNs is close to that of the random graph. Meanwhile, the average clustering coefficient is much larger, which indicates that AQSTN model possesses obvious ''small world'' property. Small-world property is related to the information propagation in the network, suggesting that in such a network that possesses small-world effect, information spreading speed is faster. Additionally, in such a network, small amount changes on the connectivity among nodes can dramatically alter the performance of the entire network. Therefore, it can be found that by adjusting the information propagation method existing in the air quality network, for example, cutting off several spreading path of air pollutants, can obtain noteworthy promotion of the air quality status of the whole network.

In Figure 14, the experimental results show that the probability distribution of node degree in AQSTN based on different dates obviously follows the power-law distribution. During the experiment process, we accomplish the data fitting and calculate the exponent $\alpha$ of the power law distribution, at different dates respectively. As shown in Table 2, the results present the value of $\alpha$, which is between 1 and 2. Since demonstrations of plenty of literatures, the complex network equips with scale-free property, which means the distribution of node degree obeys power-law distribution with its index $\alpha$ around 2. Accordingly, our model possesses distinct scale-free property, which means the connectivity among nodes is non-uniform, namely heterogeneous. Besides, it should be noted that scale-free property is closely relevant to the robustness of the whole network, contributing with the fault tolerance of the network. Nonetheless, for the selective attack based on the node degree value, its anti-attack capability is quite poor, and the existence of high degree nodes greatly weakened the network robustness. In practice, we can adopt degradation methods on the nodes with high degree and its adjacent areas based on the AQSTN, so that the overall pollution level can be significantly reduced.

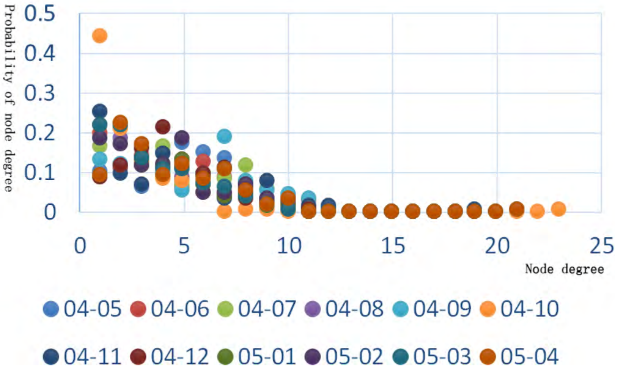


**FIGURE 14. The node degree probability distribution of air quality spatial and temporal network for each date.**

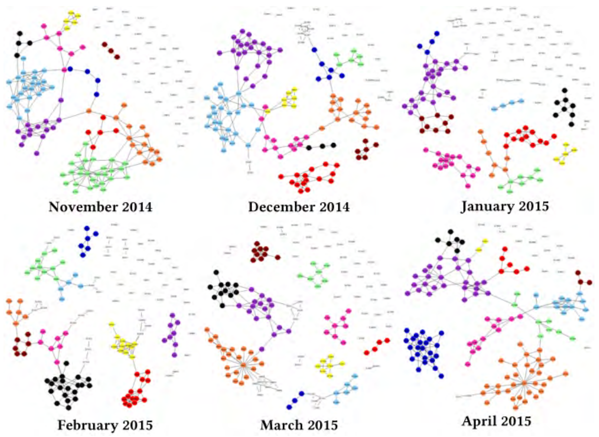
We take three community mining algorithms to compute the modularity of AQSTN for data of twelve months. Seen from the results in Table 3, the modularity of AQSTNs is almost more than 0.7, and the modularity greater than 0.3 can be regarded as a sign of obtaining significant community

**TABLE 3. Modularity Q obtained for AQSTN of different algorithms.**

| Date | FastGreedy | Walktrap | GN |
|---|---|---|---|
| 2014-05-14 | 0.766383 | 0.7520831 | 0.7684839 |
| 2014-06-14 | 0.7846528 | 0.7792938 | 0.7795854 |
| 2014-07-20 | 0.7731481 | 0.7846043 | 0.7930063 |
| 2014-08-14 | 0.7490665 | 0.7504191 | 0.7480948 |
| 2014-09-14 | 0.7357577 | 0.7357577 | 0.7357577 |
| 2014-10-31 | 0.7864658 | 0.786045 | 0.7914317 |
| 2014-11-14 | 0.7151001 | 0.7039709 | 0.7186529 |
| 2014-12-10 | 0.814995 | 0.8039555 | 0.8149474 |
| 2015-01-14 | 0.8061359 | 0.7829072 | 0.8061943 |
| 2015-02-14 | 0.8216561 | 0.8024817 | 0.8173241 |
| 2015-03-14 | 0.8107264 | 0.8114039 | 0.8152434 |
| 2015-04-15 | 0.7979065 | 0.7793341 | 0.7973742 |

structure. Particularly, some networks hold modularity above 0.8, providing an insight that the connectivity inside them is very tight. This phenomenon implies AQSTN has obvious local characteristics.

With the aim of performing the comparison and contrast of modularity, we construct the community structure diagram, which is created by Walktrap algorithm based on dates from November 2014 to April 2015. As the experiment results of date November 2014 are shown in Figure 15, when the modularity Q is smaller, there exist more scattered nodes and communities with more connections between them. In April 2015, despite less scattered nodes, interactions between communities are still relatively strong, corresponding to a smaller Q. Subsequently, we see the internal structure within community of the network is tightly connected and the relationship between communities is weak in February and March 2015, though many scattered nodes exist at the position of upper left, of which the modularity is higher. Hence, in the actual application scenario of AQSTN, when Q is high, air pollutants are more likely to spread in a small region, where local degradation measures should be taken.



November 2014 December 2014 January 2015

February 2015 March 2015 April 2015

**FIGURE 15. Community structures created by Walktrap algorithm from November 2014 to April 2015.**

## VII. CONCLUSION

This paper studies the construction process of air quality spatial and temporal network model based on the complex network, and demonstrates the dynamics and reliability of AQSTN. Besides, we analyze the community structure of our model, revealing the potential relationship between air quality regions. Through analysis, we find AQSTN has small-world effect, scale-free property and community structure characteristic. These characteristics are consistent with the observed properties of complex networks in other fields, which provide a theoretical basis for the study of air quality.

According to the investigation of this model, the application of AQSTN is proved to be highly promising. The key contributions of our work for practical application lies in three aspects. Firstly, mining algorithms can be employed on AQSTN to find core nodes, which are influential in seeking pollution source, namely the places that seriously polluted. Secondly, the edges of communities provide a basis for finding the propagation path, which can be cut off to prevent the spreading of pollutants. Finally, it is more conducive to the monitoring of air quality that establishing monitor stations in areas where pollution is more serious or interaction is more obvious.

Future work will mainly focus on the following two aspects. First, we will further research the air quality spatial and temporal network model by better combining temporal and spatial features with the aim of enhancing the adaption of real applications. For example, distinguishing the positive or negative effect of interaction to construct directed edges among nodes can contribute to accurately simulating the similarities and differences between sites. We will also concern with the latest community detecting algorithms, which can be applied to community mining in order to facilitate discussion of feasibility of AQSTN. Second, we will employ reasonable approaches to perform air quality analysis and prediction based on the division results of AQSTN. Thus, the key path of controlling air quality pollution will be explored from a macroscopic perspective, and reasonable suggestions for setting locations of monitoring stations will be given.

## ACKNOWLEDGMENT

The authors are grateful to valuable comments and suggestions of the reviewers.

## REFERENCES

[1] M. Á. Olvera-García, J. J. Carbajal-Hernández, L. P. Sánchez-Fernández, and I. Hernández-Bautista, "Air quality assessment using a weighted fuzzy inference system," *Ecol. Inf.*, vol. 33, pp. 57–74, May. 2016.

[2] L. pan, B. Sun, and W. Wang, "City air quality forecasting and impact factors analysis based on grey model," in *Proc. Procedia Eng.*, 2011, pp. 74–79.

[3] Y. Zheng, F. Liu, and H. P. Hsieh, "U-Air: When urban air quality inference meets big data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1436–1444.

[4] Y. Zheng *et al.*, "Forecasting fine-grained air quality based on big data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 2267–2276.

[5] J. J. Carbajal-Hernández, L. P. Sánchez-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "Assessment and prediction of air quality using fuzzy logic and autoregressive models," *Atmos. Environ.*, vol. 60, pp. 37–50, Dec. 2012.

[6] M. Oprea, S. F. Mihalache, and M. Popescu, "Applying artificial neural networks to short-term PM 2.5 forecasting modeling," in *Artificial Intelligence Applications and Innovations*. Cham, Switzerland: Springer, 2016.

[7] B. T. Ong, K. Sugiura, and K. Zettsu, "Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM*2.5*," Springer-Verlag, New York, NY, USA, Tech. Rep. 6, 2016.

[8] B. Liu, S. Yan, J. Li, and Y. Li, "Forecasting PM2.5 concentration using spatio-temporal extreme learning machine," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 950–953.

[9] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, 2002.

[10] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, pp. 268–276, Mar. 2001.

[11] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75–174, 2010.

[12] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[13] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[14] D. Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *Proc. 23rd ACM Nat. Conf.*, 1968, pp. 517–524.

[15] Y. Bai, L. Wu, K. Qin, Y. Zhang, Y. Shen, and Y. Zhou, "A geographically and temporally weighted regression model for ground-level $PM_{2.5}$ estimation from satellite-derived 500 m resolution AOD," *Remote Sens.*, vol. 8, no. 3, p. 262, 2016.

[16] 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.

[17] W. R. Goodin, G. J. Mcrae, and J. H. Seinfeld, "An objective analysis technique for constructing three-dimensional urban-scale wind fields," in *Journal of Applied Meteorology*, vol. 19, no. 1, pp. 98–108, 1980.

[18] S. Vardoulakis, B. E. A. Fisher, K. Pericleous, and N. Gonzalez-Flesca, "Modelling air quality in street canyons: A review," *Atmos. Environ.*, vol. 37, no. 2, pp. 155–182, 2003.

[19] D. Syrakov, M. Prodanova, E. Georgieva, I. Etropolska, and K. Slavov, "Simulation of European air quality by WRF–CMAQ models using AQMEII-2 infrastructure," *J. Comput. Appl. Math.*, vol. 293, pp. 232–245, Feb. 2016.

[20] E. Pisoni, A. Clappier, B. Degraeuwe, and P. Thunis, "Adding spatial flexibility to source-receptor relationships for air quality modeling," *Environ. Model. Softw.*, vol. 90, pp. 68–77, Apr. 2017.

[21] B. Eder, D. Kang, R. Mathur, S. Yu, and K. Schere, "An operational evaluation of the Eta–CMAQ air quality forecast model," *Atmos. Environ.*, vol. 40, no. 26, pp. 4894–4905, 2006.

[22] D. Byun and K. L. Schere, "Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (CMAQ) modeling system," *Appl. Mech. Rev.*, vol. 59, no. 2, pp. 51–77, 2006.

[23] 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, p. 86, 2016.

[24] Z. Jiang, B. Mao, X. Meng, X. Du, S. Liu, and S. Li, "An air quality forecast model based on the BP neural network of the samples self-organization clustering," in *Proc. 6th Int. Conf. Natural Comput.*, Aug. 2010, pp. 1523–1527.

[25] J. Reyes and A. Sánchez, "Analysis of air quality data in Mexico city with clustering techniques based on genetic algorithms," in *Proc. 23rd Int. Conf. Electron., Commun. Comput.*, Mar. 2013, pp. 27–31.

[26] M. G. Sefidmazgi, M. M. Kordmahalleh, A. Homaifar, and S. Liess, "Change detection in climate time series based on bounded-variation clustering," in *Machine Learning and Data Mining Approaches to Climate Science*. Cham, Switzerland: Springer, 2015, pp. 185–194.

[27] Y. Chen *et al.*, "Air quality data clustering using EPLS method," *Inf. Fusion*, vol. 36, pp. 225–232, Jul. 2017.

[28] E. Austin, B. A. Coull, A. Zanobetti, and P. Koutrakis, "A framework to spatially cluster air pollution monitoring sites in US based on the $PM_{2.5}$ composition," *Environ. Int.*, vol. 59, pp. 244–254, Sep. 2013.

[29] 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, 2014, Art. no. e96111.

[30] M. E. J. Newman, "Detecting community structure in networks," *Eur. Phys. J. B*, vol. 38, no. 2, pp. 321–330, 2014.

[31] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Computer and Information Sciences*. Berlin, Germany: Springer, 2004, pp. 284–293.

[32] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, 2001.

[33] 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, Feb. 2004, Art. no. 026113.

**GUYU ZHAO** was born in 1993. She received the B.S. degree from Hebei Normal University, China, in 2015. She is currently pursuing the Ph.D. degree with the College of Information Science and Engineering, Yanshan University, China. She is currently focusing on the project on data mining with air quality, which has been supported by the National Natural Science Foundation of China under Grant 61772451. Her research interests include data mining and machine learning. She is a member of ACM.

**GUOYAN HUANG** was born in 1969. He received the Ph.D. degree from Yanshan University, Hebei, China, in 2006, where he is currently a Professor with the College of Information Science and Engineering. His research interests include network collaborative technology and software security. He is the Principal of the National Natural Science Foundation of China under Grant 61772451. He is a Senior Member of the Chinese Computer Society and the ACM.

**HONGDOU HE** was born in 1991. He received the B.S. degree from the College of Information Science and Engineering, Yanshan University, China, in 2014, where he is currently pursuing the Ph.D. degree. He is currently focusing on a project on software security. He is proficient in Java and Python. His research interests include data mining and machine learning. His research has been supported by the National Natural Science Foundation of China under Grant 61472341. He is a member of ACM.

**QIAN WANG** was born in 1987. She received the B.S., M.S., and Ph.D. degrees from the College of Information Science and Engineering, Yanshan University, China, in 2009, 2012, and 2016, respectively.

She was a Visiting Scholar with the University of Hull, from 2015 to 2016. Since 2016, she has been a Lecturer with the School of Information Science and Engineering, Yanshan University. Her research interests include data mining, complex network, and software security.

• • •