Mining Key Stations by Constructing the Air Quality Spatial-Temporal Propagation Network

GUYU ZHAO, GUOYAN HUANG, HONGDOU HE, JIADONG REN, (Member, IEEE), AND HAITAO HE
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 in part by the National Natural Science Foundation of China under Grant 61772451, and in part by the Graduate Innovative Funding Project of Hebei Province under Grant CXZZBS2020061.

ABSTRACT Deteriorated air quality is a significant hazard to human health and a crucial problem in numerous countries. Existing research has extensively investigated the air quality states and future trends, but neglects the analysis of the time-space relevance behavior and the pollutant propagation relation of air quality. Based on the theory of complex network, this paper characterizes the interaction process between different sites, constructs the air quality propagation network and mines the key stations of the whole region. Firstly, pollutant monitoring data from the Beijing-Tianjin-Hebei region are used for analysis, monitoring stations are abstracted into network nodes, and edge sets are constructed using pollutant paths obtained from spatial and temporal reachability calculation. By accumulating the effect of the paths of the propagation network at different timestamps, the weight of the network during the propagation process is assigned for reflecting the interactive degree of stations. Secondly, the PageRank algorithm evaluates the importance of nodes in the propagation network to find the influential pollution area. Key stations considering both the quantity and quality of propagation relations are then obtained. Experiment results demonstrate the relevance of the node rank distribution and the real pollution condition and verify the reliability of the network and the node ranking algorithm. This work provides a reliable theoretical guidance on the timely prevention and control of air pollution, and gives a basis of the location selection of monitoring stations.

INDEX TERMS Air pollution, pollutant control, complex networks, ranking algorithm, data mining.

I. INTRODUCTION

The deterioration of air quality significantly impacts human health [1], [2], which has arisen an increasing concern in many countries. According to the researches [3]–[5], fine particulate pollutants, such as PM$_{2.5}$, are extremely easy to inhale into the lungs, which poses a serious threat to human health. A report from the World Health Organization (WHO) [6] showed that air pollution caused about 4.2 million deaths each year from stroke, heart disease, lung cancer, chronic respiratory tract or other related diseases. According to the Environmental Performance Index (EPI) of 2018 [7], many of its listed 180 countries suffer from excessive air pollution. Government management and early prevention are the basic functions of environmental protection work, which are not only related to scientific decision-making but also the long-term development of the country [8]–[10]. Therefore, pollution prevention and control need to find out the law of pollutant transmission and the areas with active transmission behavior.

The large scale of air quality data and complex relationship between the dynamic evolution process of pollutants and meteorological, geographical and economic factors, make it difficult to accurately describe the air quality system. Accurate characterization and measurement of spatial-temporal distribution and interaction of air quality have become the key issue in the field of air quality research. The conception of the grid is often utilized in the area of spatial interpolation of air quality for characterizing the spatial relation of different areas. Shepard [11] used the grid division method, assuming that each grid is a unit and the concentration of air pollutants in each grid is uniform and thus carried out two-dimensional spatial interpolation of air quality. Zheng et al. [12] divided the study area into disjoint grids (e.g., 3km × 3km), and...
integrated the time and space dimensions to carry out spatial interpolation of air quality, combined with the influence of adjacent grids. Bai et al. [13] and Agrawal et al. [14] made use of the definition grid to implement PM$_{2.5}$ concentration interpolation. Also, Goodin et al. [15] used an artificially defined grid to establish a three-dimensional urban wind field. Vardoulakis et al. [16] applied the grid to simulate the air quality under the conditions of street canyons. Pisoni et al. [17] increased the spatial flexibility of the “source-receptor” relationship in the air quality model by using the method of layered grids (with small granularity in adjacent grids and large granularity in distant grids). Although applications of the grid that quantifies the spatial relationship contribute to obtaining the spatial effect of air quality to some extent, the above methods are largely influenced by subjective factors. Such methods lack sufficient research on the specific constraints and correlation of air quality inner environment, which affects the quality and efficiency of subsequent analysis.

In recent years, air quality researches have attracted more and more attention to cluster analysis. Jiang et al. [18] used the pollutant concentration for clustering and then BP neural networks for air quality prediction. To promote clustering efficiency for analyzing air quality, Reyes et al. [19] improved the clustering process based on the genetic algorithm. Sefidmazgi et al. [20] detected the change of climate time series according to the bounded variation clustering. Dincer and Akkus [21] forecasted air pollution based on the use of a robust fuzzy clustering algorithm of time series. Besides, to improve the PM$_{2.5}$ forecast, Mahajan et al. [22] presented a Grid-Based clustering method. Zhao et al. [23] conducted regional PM$_{2.5}$ prediction based on the clustering result of geographical distance between stations. The clustering algorithm is adopted to meet the actual demand to some extent, but only considering the similarity of a single time dimension (time series) or space dimension (geological distance). Soh et al. [24] and Wen et al. [25] used the integration of time series and geological distance analysis to select the top-k related stations for air quality prediction. However, these two articles only consider the local effect of spatial and temporal dimensions but do not get the overall relationship of the whole region. Hence, the main issue that needs to be solved for mining the mechanism in the air quality evolution and pollution propagation is properly combining the temporal and spatial relevance between different sites from a global view, which is not settled by former researches.

Since most researches concern the evaluation and prediction of air quality at a single place and lack the analysis of the specific temporal and spatial relevance between different sites, mining the most influential places through the pollutant propagation of a region is an innovative and important topic in air quality area. The complex network model has led to a series of important discoveries [26]–[28] and played a great role in the analysis of the real world. Many complex systems in the real world exist in the form of complex networks or can be transformed into complex networks [29]–[34]. According to the researches of complex network theory, excavating the key nodes in the complex network is an important topic [35], [36]. The key node discovery is very significant for studying the robustness and security of the constructed network [37]–[39]. In terms of the invulnerability of complex networks [40], [41], deliberately attacking the key nodes of scale-free networks will make the whole network paralyzed quickly. Similarly, some locations or orientations play a key role in the propagation of pollutants. Finding these places can effectively and timely prevent and control pollution, and carry out targeted monitoring deployment. Based on the complex network theory, a new effective characterization and investigation method was presented for air quality analysis in our former work [42]. The representation method synthesized the comprehensive effects of time and space from a global perspective and conformed to the characteristics of the complex network. However, the former research mainly learns the regional relations from the community perspective and lacks further excavation of key stations to make the granularity of analysis more precise. On this basis, this paper proposes a spatial-temporal network model of air quality, abstracts the regional air quality distribution and structure into a complex network, and mines the important pollution stations according to the interaction relationship and interaction intensity between nodes.

The main contributions of this paper are as follows:

1. A temporal and spatial correlation measurement method for air quality is established by defining and calculating the spatial and temporal reachability spectrum.
2. A modeling method of air quality based on the complex network theory is provided to characterize the process of pollution propagation.
3. A reasonable key node mining method based on the PageRank algorithm is proposed on the air quality propagation network to find out the primary active locations of propagation in the region.
4. Experiments employing real data are conducted to demonstrate the reliability of our method and visualize the ranking results. The ranking results are corroborated with real air quality conditions and get an average Hit Rate of 87%.

II. DEFINITIONS

The air quality monitoring system covers a collection of monitoring sites, reporting the AQI or main pollutant concentration. As investigated in [42], the monitoring sites record the concentration by the hour, resulting in a hierarchical organization as shown in Figure 1, of which each layer $M(t_i)$ represents a circumstance of air quality at a given time $t_i$. In the process of air quality analysis, the dynamic interaction between different sites plays an important role in the future air quality state. In this paper, the spatial and temporal correlation and related factors are combined for a unified air quality analysis, and a model based on complex network theory is established to reasonably construct dynamic interaction...
pattern and improve the reliability of the regional analysis. The basic terms used in this section are defined as follows.

**Definition 1 (Reachability):** Reachability indicates whether the influence of a site can reach another one through the process of pollution propagation.

**Definition 2 (Time Reachability Spectrum (TRS)):** Time Reachability Spectrum represents whether the reachability can be established between different sites according to the time relation of a specific time span. Time Reachability captures whether the propagation possibility between sites is existed by calculating the temporal relevance of air quality state, which considers the similarity of the changing trends of certain pollutants’ concentration.

\[
\text{TRS}(t) = \begin{bmatrix}
    a_{11} & \cdots & a_{1N} \\
    \vdots & \ddots & \vdots \\
    a_{N1} & \cdots & a_{NN}
\end{bmatrix}
\]  

(1)

Here, \(\text{TRS}(t)\) stands for the Time Reachability Spectrum at time \(t\), \(a_{ij}\) means the value of \(\text{TRS}(t)\) between node \(v_i\) and node \(v_j\). The value of \(a\) is either 0 or 1. When \(a_{ij} = 0\), node \(v_i\) and node \(v_j\) do not possess reachability at the moment \(t\). Otherwise \(a_{ij} = 1\), the reachability between node \(v_i\) and node \(v_j\) is existed.

**Definition 3 (Space Reachability Spectrum (SRS)):** Space Reachability Spectrum means whether the reachability between sites can be established based on spatial relations. Space Reachability aims to reveal whether two sites have the propagation ability under the current spatially effective conditions, such as distance, wind direction, etc.

\[
\text{SRS}(t) = \begin{bmatrix}
    b_{11} & \cdots & b_{1N} \\
    \vdots & \ddots & \vdots \\
    b_{N1} & \cdots & b_{NN}
\end{bmatrix}
\]  

(2)

Similarly, \(\text{SRS}(t)\) stands for the Space Reachability Spectrum at time \(t\), \(b_{ij}\) means the value of \(\text{SRS}(t)\) between node \(v_i\) and node \(v_j\). \(b_{ij}\) takes the value 0 or 1.

In real circumstances, the requirements of both TRS and SRS are decisive to determine the final Reachability between different sites. Time Reachability proves the strong correlation of the air quality evolution of two given sites, which means the two sites can probably possess a path of pollution propagation. And, Space Reachability is achieved by satisfying the crucial propagation conditions.

**Definition 4 (Time-Space Reachability Spectrum (TSRS)):**

Time-Space Reachability Spectrum represents the results that whether the reachability is established between different sites under the joint action of time and space.

\[
\text{TSRS}(t) = \begin{bmatrix}
    c_{11} & \cdots & c_{1N} \\
    \vdots & \ddots & \vdots \\
    c_{N1} & \cdots & c_{NN}
\end{bmatrix}
\]  

(3)

\(\text{TSRS}(t)\) stands for the Time-Space Reachability Spectrum at time \(t\), \(c_{ij}\) means the value of \(\text{TSRS}(t)\) between node \(v_i\) and node \(v_j\). \(c_{ij}\) takes the value 0 or 1.

**Definition 5 (Spectrum Cumulative Matrix (SCM)):** Spectrum Cumulative Matrix denotes the summation of the spatiotemporal reachability spectrum (TSRS) between the site pairs at each time in a corresponding period.

\[
\text{SCM}(T) = \begin{bmatrix}
    s_{11} & \cdots & s_{1N} \\
    \vdots & \ddots & \vdots \\
    s_{N1} & \cdots & s_{NN}
\end{bmatrix}
\]  

(4)

\(\text{SCM}(T)\) denotes the SCM during the period \(T\). \(s_{ij}\) is the value of \(\text{SCM}(T)\) between node \(v_i\) and node \(v_j\).

**Definition 6 (Air Quality Propagation Network (AQPN)):** Air Quality Propagation Network abstracts the propagation relation of air quality system into a network structure based on the complex network theory, which can represent the state distribution and the dynamic interaction process of air quality.

\[
\text{AQPN} = \{V, E, W\}
\]  

(5)

Here, \(V\) represents the Node set of \(\text{AQPN}\). \(E\) is the Edge Set. \(W\) is the Weight Set of edges.

- **Node Set (V).** Node Set is a set of sites characterized for monitoring pollutant changes.

\[
V = \{v_1, v_2, \ldots, v_i, \ldots, v_N\}
\]  

(6)

\(N\) represents the total number of nodes.

- **Edge Set (E).** Edge Set contains air quality propagation paths between nodes in the Node Set (V).

\[
E = \{e_{ij} | i, j = 1, 2, \ldots, N\}
\]  

(7)

\(e_{ij} = <v_i, v_j>\)

(8)

- **Weight Set (W).** The Weight Set corresponding to the Edge Set is gained by dividing the accumulated weight matrix by the period length.

\[
W = \{w_{ij} | i, j = 1, 2, \ldots, N\}
\]  

(9)

\[w_{ij} = \frac{\text{SCM}}{T}\]

(10)

Here, \(\text{SCM}\) represents the Spectrum Cumulative Matrix and \(T\) represents the period length of the constructed network.

**Definition 7 (Node Rank (NR)):** Node Rank is a ranking list for each node relying on the node importance evaluation algorithm.

\[
\text{NR}(T) = \{(v_1, R_1), \ldots, (v_N, R_N)\}
\]  

(11)
NR(T) represents the ranking list of period T, (vi, R_i) expresses the corresponding rank R_i of node vi.

III. METHODOLOGY

As shown in Figure 2, our main work can be divided into two parts. Firstly, construct the Time-Space Reachability Spectrum (TSRS) of each time. Based on the complex network theory, sites with the monitoring functions are abstracted into network nodes. Through the temporal relation analysis during the process of air quality propagation, the linear correlation of short time sequences of different nodes is calculated to produce the TRS. Besides, the SRS is generated by quantifying the capability of pollution diffusion and spread between nodes. Then, the TSRS is obtained by the intersection operation of the two spectra. According to the TSRS, the adjacency matrix, i.e. Edge Set (E), of each time is constructed. Each element of E represents a path that can propagate pollutants between two nodes. Hence, the topology of the Air Quality Propagation Network (AQPN) is established. Secondly, gain the Node Rank (NR) of every period. By accumulating the number of paths between nodes during a long period, the Weight Set (W) of AQPN corresponding to E is generated. Also, the initial node scores are assigned relying on the pollution level during the period. On this basis, the improved PageRank node evaluation algorithm is utilized to calculate the importance of each node.

A. CONSTRUCTION OF TSRS

As the above analysis, the evolution of the air quality system needs to be detected from the dimensions of time and space. On this basis, the construction of TSRS quantifying the air quality propagation is mainly divided into three parts: (1) construction of TRS, (2) construction of SRS, and (3) construction of TSRS. Since the monitoring station mainly monitors and records the change of the pollutant concentration near the surface, it can adequately reflect the propagation process of pollutants in the whole region. Therefore, the monitoring station of air quality is selected to build the Node Set (V) of AQPN. The following contents describe the construction process of TSRS of current time t.

1) CONSTRUCTION OF TRS

The TRS represents the intensity of inter-node propagation in the air quality system based on the time correlation, and mainly reflects the approximation of the air quality changes of the two nodes. The short time series of each node produced in this part is generated by the sliding window and presented by time sequences of the pollutant concentration. The window size is set as 24 hours. In our methodology, each moment t belongs to a short period sequence (T_j = 24 hours) in the long period (T = a month) calculation process of AQPN. Hence, the short time series contains the information of a whole day and completely manifests the changing trend of a short periodic variation in the month. In the time dimension, the influential results on the air quality states of critical factors such as meteorological conditions and human activities can be reflected in these sequences.

Assuming there exist two short time sequences of node vi and node vj noted as C_i = {c_{i-23}, ..., c_{i-1}, c_i} and C_j = {c'_{i-23}, ..., c'_{i-1}, c'_i}. Pearson correlation is used to calculate the correlation degree of the two short time series, as shown in formula 12. The TRS of time t is then calculated by using the correlation coefficient as formula 13.

\[
Pearson(i,j) = \frac{cov(C_i, C_j)}{\sigma(C_i)\sigma(C_j)} \quad (12)
\]

\[
TRS_{ij} = \begin{cases} 
1 & \text{if } Pearson(i,j) \geq \rho \\
0 & \text{otherwise}
\end{cases} \quad (13)
\]

Here, cov(C_i, C_j) is the covariance of the two concentration time series C_i and C_j, \(\sigma(C_i)\) and \(\sigma(C_j)\) are the standard deviation of C_i and C_j respectively. \(\rho\) is the threshold of TRS. According to the correlation level list of Pearson correlation, the value of \(\rho\) is 0.6.

2) CONSTRUCTION OF SRS

The SRS is mainly calculated based on the spatial propagation capacity of the pollutant between nodes. Meteorological conditions are the main influencing factors of local state and regional pollutant propagation of air quality [1], [11]. Table 1 shows the linear correlation (Pearson correlation coefficient), level correlation (Spearman correlation coefficient), and gray correlation statistical values between each meteorological factor and \(PM_{2.5}\). The monitoring data utilized is generated in the Beijing-Tianjin-Hebei region from May 1, 2014, to April 30, 2015. According to Table 1, fine particles are strongly correlated with wind speed and humidity of meteorological factors.

| Meteorological Factor | Pearson Correlation Coefficient | Spearman Correlation Coefficient | Grey Relevant Coefficient |
|-----------------------|---------------------------------|----------------------------------|---------------------------|
| Temperature           | -0.061735                       | -0.013164                        | 0.558092                  |
| Pressure              | -0.179858                       | -0.033251                        | 0.393982                  |
| Humidity              | 0.332793                        | 0.291710                         | 0.620452                  |
| Wind Speed            | -0.240087                       | -0.193321                        | 0.862703                  |
| Wind Direction        | -0.059670                       | -0.090845                        | 0.476593                  |
| PM                    | 1                                | 1                                | 1                         |

The wind is an important meteorological factor affecting air pollutants in a region. Wind speed affects the diffusion speed, diffusion range, and local pollution state, while wind direction influences the propagation direction of pollutants. When the wind speed is high, the air in the local area will become thinner, and the precipitation effect of pollutants will be reduced. Therefore, the local area will more likely gain good air quality, but the pollutants will more likely expand to other places. When the wind speed is very low, it not only makes it difficult for pollutants to spread but also promotes the accumulation of pollutants, which can easily cause severe
local pollution. The humidity is another vital factor. When the relative humidity is quite high, the absorption of water in the air increases the weight of the fine particles, which leads to the difficulty of settling and increases the concentration of pollutants in the air. This phenomenon aggravates the local pollution situation. Besides, according to the results of Gauss Diffusion and other meteorological analysis, distance also plays a decisive role in pollutant diffusion. During the process of propagation, the pollutants will be settled and degraded to a certain extent with the increase of diffusion distance and therefore is difficult to reach distant sites.

Depending on the above analysis, this article calculates the diffusion capacity of pollutants considering the effects of distance, humidity, and wind.

The geographic coordinate set $G = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_i$ and $y_i$ are the latitude and longitude of node $v_i$ is used to calculate the spatial distance between site $v_i$ and site $v_j$.

$$d_{ij} = d_{geo}(x_i, y_i, (x_j, y_j))$$ (14)

The effect of wind is calculated by wind direction and speed. As shown in Figure 3, the wind force action between two points is represented by the difference of the components of the respective wind force on the line $l_{ij}$ between the two points.

$$F'_i(t) = F_i(t) \cos \theta_i(t)$$

$$\Delta F_i(t) = \begin{cases} |F'_i(t) - F'_j(t)| & \text{if } 0^\circ < \theta_i(t), \theta_j(t) \leq 90^\circ \\ 0 & \text{otherwise} \end{cases}$$ (15)

Here, $F_i(t)$ is the wind force of node $v_i$, $\theta_i$ is the angle between the wind force on the node $v_i$ and the connection $l_{ij}$ between the two nodes. $\theta_i(t)$ is the angle at time $t$.

The relative humidity of the node at the current moment is used to calculate the influence of humidity.

This section adopts the joint action of the local diffusion possibility and the interaction intensity between sites to calculate the SRS at time $t$. This process uses the Gauss kernel function to calculate the spatial spread capability $S_{pre}$ and the local diffusivity $Loc_i$. The result of SRS is achieved as a sum of the two values, hence the effects of them are considered to be equal. The formulas for calculating the SRS are shown as equation 17 to 19.

$$S_{pre}_{ij} = \begin{cases} 1 & \text{if } S_{pre} \neq 0 \text{ and } S_{pre} + Loc_i \geq \beta \\ 0 & \text{otherwise} \end{cases}$$ (17)

$$Loc_i = \alpha e^{-H_i}$$ (18)

$$SRS_{ij} = \begin{cases} 1 & \text{if } S_{pre} \neq 0 \text{ and } S_{pre} + Loc_i \geq \beta \\ 0 & \text{otherwise} \end{cases}$$ (19)
In the equations, $\Delta F_{ij}(t)$ represents the wind force effect at time $t$ between node $v_i$ and node $v_j$. $d_{ij}$ is the distance between $v_i$ and $v_j$. $H_i$ is the relative humidity of node $v_i$, $c$ is the threshold of the distance. And, $\alpha$ and $\beta$ are the local effect threshold and the spatial reachability threshold separately.

3) CONSTRUCTION OF TSRS
The TSRS is obtained by intersecting the TRS and SRS at time $t$.

$$TSRS_{ij} = \begin{cases} 1 & \text{if TRS}_{ij} = 1 \text{ and } SRS_{ij} = 1, \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

$TSRS_{ij}$ means the TSRS between node $v_i$ and node $v_j$. According to the TSRS of each moment, the propagation paths over the entire region at that moment is established and constitute the topology structure of the AQPN, as shown in Figure 4.

![FIGURE 4. Example diagram of paths construction based on TSRS(t).](image)

B. MINING KEY NODES OF AQPN

Based on the above steps, the AQPN is then constructed dynamically by month. The importance of the node in the AQPN indicates the ability of the node to propagate pollution. Specifically, the pollution influence of nodes can be understood as the capacity to impact the air quality of surrounding nodes after a certain degree of pollutant accumulation. Our model utilizes the PageRank algorithm [43] to estimate the node grade in the air quality network.

PageRank algorithm is an algorithm based on node centrality and is proposed to rank pages of the Web network. The topological structure of AQPN is similar to that of the Web network. The nodes in AQPN are equivalent to the pages in Web network, and the establishment of a path between one node and other nodes is equivalent to the page adding a link to other pages. In the air quality system, sites with higher concentrations of pollutants are more likely to cause pollution diffusion. Besides, nodes in the region with more propagation paths are more likely to transmit pollutants. Therefore, two criteria of nodes evaluation in this paper can be summarized as follows:

**Quantity criterion:** The higher the amount of the linked high-influential nodes of node $v_i$ is, the higher the importance of the node $v_i$ is.

**Quality criterion:** The higher the influence of the linked nodes of node $v_i$, the higher the importance of the node $v_i$ is.

1) CALCULATE THE INITIAL NODE WEIGHT
In the air quality system, the diffusion of the pollutant depends not only on the actual reachability with other places but also on its pollution state. Sites with higher pollutant concentrations tend to spread to the adjacent areas with lower pollutant concentrations. Consequently, the initial weight of the node is assigned according to the pollution level, which is acquired by calculating the average pollutant concentration in the corresponding period. For example, if the level of the average pollutant concentration of node $v_i$ is 3, the initial weight $R^0_i$ of node $v_i$ is 3. Table 2 is the comparison of pollutant levels of fine particle pollutant PM$_{2.5}$.

| PM$_{2.5}$ Concentration(μg/m$^3$) | Level |
|----------------------------------|-------|
| <35                              | 0     |
| 75$\leq$x$<35$                   | 1     |
| 115$\leq$x$<150$                 | 2     |
| 150$\leq$x$<250$                 | 3     |
| $x\geq250$                       | 4     |

2) CALCULATE THE PATH WEIGHT
The path weight of AQPN reflects the probability of pollutant propagation in this path. According to the calculation of TSRS given in the previous section, when the value of the TSRS between two nodes is 1, the two nodes can establish a propagation path at that time. In this case, the TSRS describes the frequency of establishing paths between nodes. Therefore, the SCM is obtained by summing the TSRS of each moment during the period $T$.

$$SCM_{ij} = \sum_{k=1}^{T} TSRS_{ij} \quad (21)$$

$SCM_{ij}$ means the SCM of the path between node $v_i$ and node $v_j$. $T$ is the length of the period.

On this basis, to prevent data overflow, the weight matrix $(W)$ is obtained by dividing the SCM by the period length $T$.

3) CONDUCT THE PAGERANK ALGORITHM
Based on the initial weight of the node and the weight on the path, the PageRank algorithm is implemented to evaluate the importance degree of each node. The calculation formula of the PageRank algorithm is as follows:

$$R_i = \frac{1 - d}{N} + d \sum_{v_j \in P(v_i)} \frac{R_i}{Q(v_j)} \quad (22)$$

In the equation, $d$ is the damping coefficient, $d = 0.85$. $N$ is the total number of nodes. $P(v_i)$ denotes the node set that link from the node $v_i$. $Q(v_j)$ denotes the node set that point to node $v_j$. $R_i$ is the importance score of node $v_i$. The matrix...
form of the above equation is as follows:

\[
R = \begin{bmatrix} \frac{1 - d}{N} \\ \vdots \\ \frac{1 - d}{N} \end{bmatrix} + d \begin{bmatrix} \sum_{i \in U_1} w_{i1} \\ \vdots \\ \sum_{i \in U_N} w_{iN} \end{bmatrix} R
\]

(23)

Here, \( R = [R_1, \ldots, R_N]^T \). If node \( v_j \) is not point to node \( v_i \), \( q_{ij} = 0 \). For each \( v_i \), \( \sum_{i=1}^{N} q_{ij} = 0 \).

The AQPN constructed in this paper is an undirected weight network. Depending on the theory of undirected network, the out-degree of each node is equal to the in-degree, i.e. \( M(v_i) = L(v_i) \). Based on the concept of weighted network, the transfer matrix \( M \) in this section is defined as follows:

\[
M = \begin{bmatrix} \sum_{i \in U_1} w_{1i} & \cdots & \sum_{i \in U_N} w_{1i} \\ \vdots & \ddots & \vdots \\ \sum_{i \in U_1} w_{Ni} & \cdots & \sum_{i \in U_N} w_{Ni} \end{bmatrix}
\]

(24)

where \( w_{ij} \) is the path weight between node \( v_i \) and node \( v_j \). \( U_i \) represents the total number of nodes that establish paths with node \( v_i \).

PageRank algorithm multiplies the initial weight matrix \( R^0 \) and transition matrix \( M \) iteratively until the grade score of each node reaches the convergence state. According to the final grade score, the node grade list (NR) of the node-set can be obtained.

C. PARAMETERS

This section gives the selection process of each parameter involved in the above-mentioned formula in detail.

Firstly, \( \rho \) is the reachability threshold when calculating the TRS. The threshold value is selected according to the grade of the Pearson correlation coefficient. The comparison table of the Pearson correlation coefficient is shown in Table 3. In this method, \( \rho \) takes the value of 0.6. That is, when there is a strong correlation coefficient between the time series of two nodes, the value of TRS is 1.

TABLE 3. Pearson correlation coefficient comparison table.

| Absolute Value of Pearson Correlation Coefficient | Degree of Correlation |
|-------------------------------------------------|-----------------------|
| 0.8-1.0                                         | Highly Relevant       |
| 0.6-0.8                                         | Strong Relevant       |
| 0.4-0.6                                         | Moderate Relevant     |
| 0.2-0.4                                         | Weak Relevant         |
| 0.0-0.2                                         | Very Weak Relevant or No Relevant |

Secondly, three parameters are involved in the calculation of SRS. Here, \( c, \alpha, \) and \( \beta \) are distance threshold, local effect threshold, and spatial reachability threshold, respectively. The distance threshold \( c \) is selected by constructing the topology network (not AQPN). The network establishes a path only when the distance between two nodes is less than the distance threshold \( c \). The selection criterion is to produce a network with the largest average shortest path and contains all sites in that network as much as possible. As shown in Figure 5, when the distance threshold \( c \) is 55, the average path distance reaches the peak value and only one station is lost. The local effect threshold \( \alpha \) is determined by the Pearson coefficient based on the linear fitting between humidity and pollutant concentration, as shown in Figure 6. \( \alpha \) takes the value of 0.52. The threshold \( \beta \) selects the upper quartile of the spatial reachability value (the sum of the spatial spread capability \( Spre_{ij} \) and the local diffusivity \( Loc_{i} \)) of all nodes at each time. When the value of spatial reachability is greater than the threshold value, the corresponding spectrum (SRS) is 1. Otherwise, it is 0. Figure 7 shows the spatial reachability numerical distribution of the first 50 moments (hours) of May 2014. It can be seen from Figure 7 that the value of \( \beta \) is continuously changing.

V. RESULTS AND DISCUSSION

In this part, the ranking results of node importance and relevant information obtained in our model are analyzed in detail. The experiment is based on the analysis of near-surface monitoring data in the Beijing-Tianjin-Hebei region. The pollutant selected for study in this paper is PM\(_{2.5}\). Hence, the pollutant analyzed in the experiment, if not specified, is PM\(_{2.5}\).
The specific information of the data is given in [11]. The data used in this paper are the actual data monitored and recorded by air quality monitoring stations and weather condition monitoring stations in the Beijing-Tianjin-Hebei region. The data set is detected by the hour, that is to say, every record of the data is produced every hour. According to the specific details of the data set listed in Table 4, it can be seen that the data set covers a wide range area and has certain representativeness. At the same time, the data is very suitable to verify the regional analysis proposed in this paper because of the strong regional interaction near the earth surface recorded in the data set.

### TABLE 4. Some details of datasets.

| Datasets       | Beijing | Tianjin |
|----------------|---------|---------|
| Time span      | 2014/5/1-2015/4/30 | 2014/5/1-2015/4/3 |
| Nearby cities  | 14      | 17      |
| In-city stations | 36      | 27      |
| In-city instances | 278,023 | 189,604 | 106.4 | 104.3 |
| Neighboring Sta. | 233     | 267     |
| # of instances | 1,272,979 | 1,436,651 |
| Aqi            | 17      | 20      |
| In-city sources | 17      | 106,614 |
| Nearby sources | 116,847 | 195     |
| Meteorology    | 177     | 1,108,873 |

The experimental results are mainly analyzed from three aspects: (1) Statistical results of grade scores of nodes; (2) Visualization of the important rank of nodes in combination with geographical conditions; (3) Comparative analysis of node importance level and other relevant information. (4) Evaluation of final results.

### A. STATISTICAL DISTRIBUTION OF RANK SCORES

This section provides statistical insights into the ranking results based on the PageRank algorithm. To verify the rationality of the constructed model and the effectiveness and reliability of the ranking algorithm, the distribution regularity, and the evolution trend of ranking scores are explored. From May 2014 to April 2015, Figure 8 and Figure 9 present the fitting curves of the scores of nodes every month.

According to the statistical results, ranking scores possess a unified distribution law and the trend of the changing curve can fit into exponential functions. Among the fitting results, the exponential value is between 0.0025-0.0033 and the intercept on the y-axis is negative ranging from -0.983 to -0.980. In respect of the fitting curve, the changing rates of scores at the top or bottom of the ranking list are large, and the changes in the scores of the middle-ranking tend to be flat. In other words, a few nodes of the AQPN have a larger or smaller grade, while most nodes are in the middle places. This result is beneficial to screen the nodes at the top or the bottom of the ranking, and these nodes are more important for revealing the propagation of pollutants. Furthermore, the relative stability of the change curve can be applied to the prediction of node ranks. Besides, there exists a certain similarity of the fitting curves gained from adjacent months. The differences of the index value and the intercept value are smaller in adjacent months and larger in distant months.

### B. RANK DISTRIBUTION VISUALIZATION

In this section, network characterization and ranking results are illustrated and summarized alongside geographical conditions for further demonstrating the rationality of the proposed theory.

Figure 10 shows the topology of the AQPN each month from May 2014 to January 2015 in the Beijing-Tianjin-Hebei region. In these pictures, nodes are presented with red color, and edges between nodes are denoted as yellow lines. For each period, the edges reflect the intuitive propagation path of pollutants and are assigned with corresponding weight. It can be seen that the city of Chengde has an isolated node, which is decided by threshold c. Particularly, one site in Zhangjiakou city set up an accessibility path in May and July, but not establish the path in June. This phenomenon shows that the propagation path varies continuously according to the time and illustrates the dynamic attribute of AQPN.

Figure 11 and Figure 12 is the topographic map of node importance distribution each month from May 2014 to April 2015. The areas colored by gray-yellow in these pictures represent high plateaus and mountains, and the areas colored by green represent low plains. The triangle in the figures represents the geographical location of the node, and the size of the triangle represents the average pollution level of the node in that month. Besides, the depth of red in the triangle indicates the rank level of nodes. The higher the level, the deeper the red.

There are mainly five regions with high node grades, which distribute in the cities of Beijing, Tianjin, Shijiazhuang, Baoding, and Tangshan. These places are in low-lying positions. The relatively higher Zhangjiakou and Chengde areas
are relatively weak in both pollution degree and node grade. This phenomenon is in line with the actual situation. The Beijing-Tianjin-Hebei region is surrounded by mountains on three sides and faces the sea on the east. Nature has the effect of dust particles accumulating toward the lowlands. Pollutants coming from the northwest slow down along the downward of mountains and tend to build up at the foot of the mountains. Meanwhile, fine particulates can combine with water vapor from the sea to form smog. In addition, the prevailing wind direction in north China is westerly or north-westerly. Blow low-speed west wind or northwest wind along with the greater humidity of the weather is easy to produce haze. Therefore, the pollution level and node level of Beijing, Shijiazhuang and Baoding areas close to the mountains and Tangshan and Tianjin areas close to the sea are generally higher.

It can be seen that the nodes with high ranks are mostly distributed in the urban center. In these areas, not only that the traffic and industrial development are more common, but that the vehicle and combustion emissions are more serious. Industrial and related coal combustion emissions are the main cause of PM$_{2.5}$. Also, these urban centers have more monitoring sites that can better monitor changes in pollutants. As a result, the rank scores of city centers are relatively higher.

Another conclusion is that the pollution importance of nodes possesses a continuity changing with time. Concerning the Shijiazhuang region, there appears an upward trend in November and December of 2014. Afterwards, the importance of nodes in this region ranks high in January and February. Besides, the Beijing region continues to present high ranking nodes in February, March and April 2015.

Furtherly, the pollution level and the importance level of nodes tend to be consistent. Larger triangles usually possess a deeper red color. However, some inconsistent cases should be noted. For example, in October 2014, the pollutant level in Chengde was high, but the importance of its nodes is not high. The Beijing area had a low level of pollution concentration in June 2014, but its nodes rank high in importance. These phenomena are in line with that the important node is obtained by combining the local pollutant level and its accessibility to the surrounding nodes.

Due to the consistency with air quality circumstances shown in the results, the AQPN and its evaluation method are presented to possess high rationality and research value.

**C. DETAILS AT THE NODE LEVEL**

To further test the performance of the ranking algorithm, this section carries out a comparison of the detail relevant statistical information. The statistical lists comprised of the top-10 nodes of each month from May 2014 to April 2015 are given from Table 5 to Table 16. Statistical items in these lists include Rank, Station Number, City, Score, Node Degree,
FIGURE 9. Node rank fitting curve from September 2014 to April 2015.
FIGURE 10. Air quality propagation network topology from May 2014 to January 2015.
FIGURE 11. Air quality node rank distribution from May 2014 to January 2015.
and PM$_{2.5}$ Pollution Ratio. The City is the administrative region where the site is located. Since the AQPN established in this paper is an undirected weighted network, the Node Degree means the number of paths the node connected. The PM$_{2.5}$ Pollution Ratio represents the proportion of the number of days, where the PM$_{2.5}$ pollutant concentration exceeds 115 g/m$^3$ (pollution level greater than 3), in the total cycle length (a month).

As a whole, the node ranks reflect the continuity in time and the propagation in space. In terms of the city, Shijiazhuang is used as an example. It ranks the first for the months from May to August in 2014, while the overall ranking of nodes in this region shows a downward trend from September to December, which reveals a time continuity. As for the granularity of the station, the 11011 station ranks the first from May to July 2014, and the third in August. After that, 11011 does not appear in the top-10 list of the month of September, showing a downward trend. Spatially, adjacent sites tend to be similarly ranked. For example, in May 2014, there are three sites in the Shijiazhuang city, which implies the importance of nodes in this region is relatively high. Starting from July 2014, the Beijing region has shown more nodes at the top-10 list of the node ranks. From the perspective of regional statistics, internal nodes of Shijiazhuang, Xinji, and Baoding show a certain degree of ranking approximation due to their proximity.

The comparison between grade score and node degree shows that the node degree of the top-10 ranking node is basically above 10. In particular, the node degree in December is above 8. This result is caused by the presence of the most important node in December in Baoding, where
it has fewer monitoring stations and relatively few established paths. In contrast, there are more monitoring stations in Beijing, and the degree of the top ten nodes in the region is approximately 30. Shijiazhuang area, as the provincial capital city, is next, with the node degree of the top ten importance ranking nodes around 20. The top-ranking nodes also have relatively high levels of PM$_{2.5}$ pollution. The proportions of PM$_{2.5}$ Pollution Ratio in...


### October and December 2014, January and February 2015 are above 50%. However, it can be seen that some stations with a lower PM$_{2.5}$ Pollution Ratio rank higher places in the lists. The reason for this result is that these stations have many paths established with surrounding sites, and its surrounding nodes are of high importance. For example, station 1007 with lower PM$_{2.5}$ Pollution Ratio is connected with station 1012 and station 1014 with higher PM$_{2.5}$ Pollution Ratios in April 2015 and ranks the first. According to the principle of PageRank, a node gains a high rank if it is connected to other high-rank nodes.

The above results demonstrate that our modeling and evaluation methodology conforms the temporal and spatial dynamism, the complex network characteristic, and the actual pollution state.

### D. EVALUATION OF RANKING RESULTS

Since there lacks criteria for the evaluation of air quality stations in former researches. Here, the conception of PM$_{2.5}$ Pollution Ratio (PMPR) is used for the evaluation of the ranking results. The median of the PMPR of all nodes is represented by $\gamma$. Considering this as the threshold, we calculate the Hit Rate (HR, A ratio represents that the PMPR value of a node is large than $\gamma$ and the node is also presented in the top-10 list by the ranking algorithm) and False Alarm Rate (FAR, A ratio represents that the value of PMPR of a node below the threshold and the node is falsely shown in the top-10 list by the ranking algorithm) to evaluate the effectiveness of our methodology (Table 17).

As shown in Table 17, except for November 2014, other months all show the values of HR above 50%. Especially, the HR values of June 2014 and February 2015 are 100%, which means that the stations presented in the top-10 list of these two months are all meet the requirement. Besides, three months get 90%.

To further demonstrate the accuracy of our work, we also compare it to three baselines, noted as PR, DC, CC in Table 18. The modeling network of baselines is a simpler version of our methodology and is constructed by calculating the temporal relevance and the spatial constraint of nodes. Temporal relevance is computed by the Pearson Coefficient of monthly time series of different nodes. When the
TABLE 17. HR and FR of the ranking results of different months.

| Time      | Median | Stations (PMPR ≥ y) | HR  | FAR |
|-----------|--------|---------------------|-----|-----|
| May 2014  | 0.130435 | 6                   | 60% | 40% |
| June 2014 | 0.148148 | 10                  | 100%| 0%  |
| July 2014 | 0.258065 | 9                   | 90% | 10% |
| August 2014 | 0.182981 | 6                   | 60% | 10% |
| September 2014 | 0.206897 | 5                   | 50% | 50% |
| October 2014 | 0.419355 | 9                   | 90% | 10% |
| November 2014 | 0.4     | 3                   | 30% | 70% |
| December 2014 | 0.451613 | 8                   | 80% | 20% |
| January 2015 | 0.483871 | 7                   | 70% | 30% |
| February 2015 | 0.357143 | 10                  | 100%| 0%  |
| March 2015 | 0.290323 | 9                   | 90% | 10% |
| April 2015 | 0.172619 | 5                   | 50% | 50% |

TABLE 18. HR of different ranking methods of different months.

| Time      | HR (PR) | HR (DC) | HR (CC) | HR (our) |
|-----------|---------|---------|---------|----------|
| May 2014  | 40%     | 40%     | 40%     | 60%      |
| June 2014 | 80%     | 70%     | 60%     | 100%     |
| July 2014 | 10%     | 90%     | 90%     | 90%      |
| August 2014 | 50%     | 50%     | 50%     | 60%      |
| September 2014 | 40%     | 10%     | 20%     | 50%      |
| October 2014 | 90%     | 90%     | 90%     | 90%      |
| November 2014 | 70%     | 60%     | 70%     | 30%      |
| December 2014 | 70%     | 50%     | 50%     | 80%      |
| January 2015 | 50%     | 40%     | 70%     | 70%      |
| February 2015 | 90%     | 80%     | 40%     | 100%     |
| March 2015 | 70%     | 60%     | 80%     | 90%      |
| April 2015 | 50%     | 50%     | 50%     | 50%      |

Then, the aforementioned baseline PR represents that the PageRank algorithm is applied to the simple version network to evaluate the importance of nodes. Similarly, DC and CC respectively imply the utilization of Degree Centrality [45] and Closeness Centrality [46] on the simple version network for node evaluation. Table 18 presents the results of Hit Rate of the three baselines and our methodology. As shown in Table 18, except for November 2014, other months’ evaluation results based on our methodology show better or equal performance compared to the baselines. These results show that our method of combining the complex network conducted for modeling and the PageRank algorithm used for ranking is both reliable and scalable. Especially, our work takes the TRSR to calculate the temporal and spatial relationships between nodes, which incorporates the temporal relevance of each day and the spatial constraint of meteorological factors and distance. Hence, it accounts for more information for node evaluation. Additionally, our modeling network (AQPN) is constructed based on the TRSR accumulation of a month and simulates the propagation of pollutants to some extent. Under this situation, the PageRank algorithm with both quantity and quality criteria is a good choice for evaluating the node importance.

V. CONCLUSION
The air quality system is dynamic, dependent, and complex. It is critical to accurately analyze air quality and timely prevent and control pollution. Thus, scientifically characterizing the internal structure of air quality distribution and correlation and revealing the dynamic evolution of air quality is a significant issue. In our work, the spatial and temporal reachability spectrum (TRSR) is firstly calculated to quantify the interaction intensity between sites. Secondly, the monitoring station is abstracted into the complex network node, and the connection between nodes is built up according to the TRSR. After that, the edge weight is allocated by the periodic spectrum superposition effect to obtain the AQPN. The network model can scientifically represent the dependence of regional interconnection and interaction of air quality. Finally, the PageRank algorithm is employed to adapt to the structural characteristics of the AQPN, and the key nodes that have a significant influence on the surrounding nodes in the network are mined.

As for the theoretical contribution, this project reveals regional pollutant propagation, performs regional influence measurement in the air quality system, and proposes a more scientific and reasonable air quality analysis method. Therefore, this work provides theoretical guidance and technical support for further air quality analysis, such as prediction, which is more in line with practical scenarios. In practical application, key nodes of air quality contribute to the effective control of pollution by taking special prevention and control measures and benefits the improvement of the overall air quality.

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

REFERENCES
[1] M. L. Bell, R. Goldberg, C. Hogrefe, P. L. Kinney, K. Knowlton, B. Lynn, J. Rosenthal, C. Rosenzweig, and J. A. Patz, “Climate change, ambient ozone, and health in 50 US cities,” Climatic Change, vol. 82, nos. 1–2, pp. 61–76, Mar. 2007.
[2] M. Lippmann, “Health effects of ozone a critical review,” Japca, vol. 39, no. 5, pp. 672–695, 1989.
[3] A. Nel, “Air pollution-related illness: Effects of particles,” *Science*, vol. 308, no. 5723, pp. 804–806, 2005.

[4] C. A. Pope, J. C. Hansen, R. Kuprow, M. D. Sanders, M. N. Anderson, and D. J. Eatough, “Vascular function and short-term exposure to fine particulate air pollution,” *J. Air Waste Manage. Assoc.*, vol. 61, no. 8, pp. 858–863, Aug. 2011.

[5] J. Lelièvre, I. S. Evans, M. Fnaïs, D. Giannadaki, and A. Pozzer, “The contribution of outdoor air pollution sources to premature mortality on a global scale,” *Nature*, vol. 525, no. 7569, p. 367, 2015.

[6] WHO. (2018). Ambient Air Pollution—A Major Threat to Health and Climate. [Online]. Available: http://www.who.int/airpollution/ambient/en/

[7] EPL. (2018). Global Metrics for the Environment-The Environmental Performance Index Ranks Countries Performance on High-priority Environmental Issues. [Online]. Available: https://epi.envirotecenter.yale.edu/results-overview

[8] B. Fu, K. Kurisu, K. Hanaki, and Y. Che, “Influential factors of public intention to improve the air quality in China,” *J. Cleaner Prod.*, vol. 209, pp. 595–607, Feb. 2019.

[9] R. A. Hernandez, “Prevention and control of air pollution in China: A research agenda for science and technology studies,” *Surveys Perspect. Integrating Environ. Soc.*, vol. 8, no. 1, 2015.

[10] Y. Jin, H. Andersson, and S. Zhang, “Air pollution control policies in China: A retrospective and prospects,” *Int. J. Environ. Res. Public Health*, vol. 13, no. 12, p. 1219, Dec. 2016.

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

[12] Y. Zheng, F. Liu, and H.-P. Hsieh, “U-air: When urban air quality inference meets big data,” in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2013, pp. 1436–1444.

[13] Y. Bai, L. Wu, K. Qin, Y. Zhang, Y. Shen, and Y. Zhou, “A geographically and temporally weighted regression model for ground-level PM2.5 estimation from satellite-derived 500 m resolution AOD,” *Remote Sens.*, vol. 8, no. 3, p. 262, Mar. 2016.

[14] 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.

[15] W. R. Goodin, G. J. Mcrae, and J. H. Seinfeld, “An objective analysis technique for constructing three-dimensional urban-scale wind fields,” *J. Appl. Meteorol.*, vol. 19, no. 1, pp. 98–108, 1980.

[16] 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. 153–182, Jan. 2003.

[17] G. A. Mendes, L. R. da Silva, and H. J. Herrmann, “Traffic gridlock on complex networks,” *Phys. A, Stat. Mech. Appl.*, vol. 391, pp. 1777–1787, Feb. 2012.

[18] M. E. J. Newman, “The structure and function of complex networks,” *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.

[19] C. Gao, D. Wei, Y. Hu, S. Mahadevan, and Y. Deng, “A modified evidential methodology of identifying influential nodes in weighted networks,” *Phys. A, Stat. Mech. Appl.*, vol. 392, no. 21, pp. 5490–5500, Nov. 2013.

[20] D. Wei, X. Deng, X. Zhang, Y. Deng, and S. Mahadevan, “Identifying influential nodes in weighted networks based on evidence theory,” *Phys. A, Stat. Mech. Appl.*, vol. 392, no. 10, pp. 2564–2575, May 2013.

[21] R. Albert, H. Jeong, and A.-L. Barabási, “Error and attack tolerance of complex networks,” *Nature*, vol. 406, no. 6794, pp. 378–382, Jul. 2000.

[22] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin, “Breakdown of the Internet under intentional attack,” *Phys. Rev. Lett.*, vol. 86, no. 16, p. 3682, 2001.

[23] G. Zhao, G. Huang, H. He, and Q. Wang, “Innovative spatial-temporal network modeling and analysis method of air quality,” *IEEE Access*, vol. 7, pp. 26241–26254, 2019.

[24] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 107–117, 1998.

[25] Y. Qi, Q. Li, H. Karimian, and D. Liu, “A hybrid model for spatial-temporal forecasting of PM2.5 based on graph convolutional neural network and long short-term memory,” *Sci. Total Environ.*, vol. 664, pp. 1–10, May 2019.

[26] R. Vallaubbajosyula, D. Chakravarti, S. Lutfiati, A. Ray, and A. Raval, “Identifying hubs in protein interaction networks,” *PLoS ONE*, vol. 4, no. 4, p. e5344, Apr. 2009.

[27] S. Wuclity and P. F. Stadler, “Centers of complex networks,” *J. Theor. Biol.*, vol. 223, no. 1, pp. 45–53, Jul. 2003.

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 also focusing on the project on data mining with air quality, which has been supported by the National Natural Science Foundation of China. She is also doing research in 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. He is currently a Professor with the College of Information Science and Engineering, Yanshan University. He is also the Principal of the National Natural Science Foundation of China. His research interests include network collaborative technology and software security. He is a Senior Member of Chinese Computer Society and ACM.

JIADONG REN (Member, IEEE) received the B.S. and M.S. degrees from the Northeast Heavy Machinery Institute, in 1989 and 1994, respectively, and the Ph.D. degree from the Harbin Institute of Technology, in 1999.

He is currently a Professor with the School of Information Science and Engineering, Yanshan University, China. His research interests include data mining, complex networks, and software security. He is a Senior Member of the Chinese Computer Society and a member of the IEEE SMC Society and 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 pursuing the Ph.D. degree. He is also focusing on a project on software security. He is also proficient in Java and Python. His research interests include data mining and machine learning. He is a member of ACM. His research has been supported by the National Natural Science Foundation of China.

HAITAO HE was born in January 1968. She received the Ph.D. degree in mechanical design and manufacturing from Yanshan University, China. She is currently a Professor with the School of Information Science and Engineering, Yanshan University. Her research interests include data mining, network information security, and artificial intelligence.

***