Key node mining algorithm for directed weighted air quality network based on propagation characteristics

Chen Song¹, Xiankun Zhang*¹, Xinqian Liu², Dawei Ren², Mei Dong¹

¹College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China
²School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China
zhxkun@tust.edu.cn

Abstract. The decline of air quality seriously affects human life and ecological environment. The blind allocation of governance resources leads to poor improvement. In order to allocate resources reasonably and improve treatment efficiency, a new key nodes mining algorithm for air quality system based on network structure and the characteristics of pollutant transmission is proposed, aiming at resource investing guidance. Firstly, the air quality network is established and its structural characteristics are analyzed. Secondly, according to the diffusion and attenuation mechanism of air pollutants in the network, a bidirectional transmission key node mining algorithm is proposed which takes both the in-links and out-links into consideration. Thirdly, a dynamic independent threshold propagation model in directed weighted network is proposed, and the number of activated nodes is used as evaluation criterion for key node mining results. Finally, experiment is executed on Jing-Jin-Ji PM2.5 air quality network. Experiment results show that the bidirectional transmission key node mining algorithm can get accurate results and good applicability in air quality network.

1. Introduction

Air is essential for animals and plants. At present, air quality related research has become a hot issue in environmental science, physics, computer science and so on. Air quality is affected by many factors, such as economy, meteorology, geography. It is complex, multi-dimensional and variable. Extensive pollutants transmission paths exist in different regions and the concentration distribution is extremely uneven. Different regions play different roles in the network, while some key regions play more important roles in the generation and diffusion process of pollutants. Mining the key nodes in air quality network can help to find out the key areas in massive monitoring data, and tilt the limited governance resources to the local nodes with heavy pollution and strong propagation ability.

Key node mining is a hot topic in network research. The existing methods can be roughly divided into three categories: centrality-based method, node shrinkage and deletion method, eigenvector-based method. Node centrality method [1] is usually simple in calculation and low in time complexity. Among them, the degree centrality algorithm uses the number of adjacent nodes to represent the node importance. The betweenness centrality [2] measures the amount of information carried by nodes in network structure. Closeness centrality [3] is based on the definition of node distance. The second category is based on node deletion [5] and shrinkage [4]. Node deletion method can be analyzed by the number of spanning trees after each deletion. The smaller the number of spanning trees is, the more important the node is. The shrinkage method merges the nodes and their adjacent nodes into a new node. The last
category is based on eigenvector [6]. It treats neighbor nodes differently based on their own information, considering node importance is affected by external information at the same time. Eigenvector method also includes different algorithms, such as, cumulative nomination [7], SALAS [8], LeaderRank [9]. All these methods are based on network structure and node information, and they have achieved good results. However, these methods only rely on the basic topology of the network, lacking the comprehensive consideration of the real world. In this paper we focus on air quality system which is affected by many factors.

Based on the propagation characteristics of air quality network, we propose the BiTr (Bidirectional transition) key node mining algorithm. The air quality network was constructed, then node importance is calculated based on the transfer of the in-link and the side feedback of the out-link. To evaluate the accuracy of the mining algorithm, the ITP (Independent threshold propagation) model is designed based on node influence range. The algorithm comprehensively considers node degree, connection relationship and the characteristics of air quality network. The key node mining results are accurate and targeted.

2. Construction and Analysis of Air Quality Network

In order to characterize air quality propagation mechanism and pollutant interaction under the current large-scale monitoring data, we construct a directed weighted network of air quality. The construction process takes the influential factors of pollutant transmission, actual geographical and meteorological conditions into consideration, providing a basis model for the mining of key nodes.

2.1. Network construction method

Node and link are the two essential elements of a network. The existing air quality data is stored in a disorder form. Abstracting into a visual network structure through the propagation relationship between local regions is of fundamental significance in key node mining work. When constructing the network, local region is regarded as the node in the network, and the propagation relationship of pollutants between local regions is regarded as link in the network.

The calculation of the propagation relationship is based on the diffusion mechanism of air pollutants. Considering the geographical and meteorological factors as well as the correlation of local pollutant concentration, the propagation cost $P_{coij}$ is defined. When the propagation cost is less than the threshold, there is a path between node i and the corresponding node. Propagation cost is defined as follows:

$$P_{coij}(t) = \frac{Dist_{ij} \cdot Hum_{i}(t)}{Per_{ij} \cdot \Delta Win_{ij}(t)}$$

In Eq(1), $P_{coij}(t)$ is the propagation cost from local region i to j at time t; $Dist_{ij}$ is the arc length distance between local region i and j on the surface of the earth sphere; $Hum_{i}(t)$ represents the humidity of local region i at time t at the beginning of propagation; $Per_{ij}$ represents the Pearson correlation between local region i and j; $\Delta Win_{ij}(t)$represents the wind difference between local region i and j at time t. The higher the propagation cost, the more difficult the propagation path will appear between the two nodes. According to $P_{coij}(t)$, propagation path set can be obtained.

![Diagram](Diagram.png)

Select the propagation path set according to real circumstance. Firstly, distance is selected as an existence condition for propagation. Only the propagation path within the definite distance interval can be reserved. The reason is that pollutants are not reachable when the distance between two stations is
too far, and a large number of closed loops will appear when the distance is too close. Secondly, wind difference is selected as the existence condition of paths. When wind difference is negative, we consider the path has no propagation basis. After path selection, the air quality propagation network corresponding to an independent time is obtained. Then the networks of each independent time during one cycle are accumulated. The number of path occurrences is taken as weight. Then the directed weighted air quality network of the whole cycle is constructed from the perspective of statistics.

2.2. Analysis of the propagation characteristics

Air quality network has the general characteristics of complex network, but due to the difference between air pollutants and real individual, the calculation and ranking of node importance in this network is different from the traditional algorithms. Specifically, there are two typical characteristics of pollutant transmission: the first one is the multi-directional diffusion of heavily polluted areas. A large number of pollutants emitted from pollution sources lead to a significant concentration increase of local pollutants, which diffuses to the neighbour regions along the concentration difference direction. This action forms multiple out-links in network structure. We do reverse reasoning from network structure to the actual situation, and it can be inferred that the local regions with more out-links mostly are heavily polluted areas, which has a great influence on the air quality in the whole network. The more out-links, the higher its local importance. The second characteristic is the attenuation of pollutant transmission. In time dimension, local pollutants continuously diluted with time. In spatial dimension, pollutants can spread to different areas with long distance. But with the concentration decreasing, its influence ability gradually decreases according to the level. As shown in Fig.2, A is a heavily polluted local region, pollutants are released from sources and then spread to the surrounding regions in multiple directions, forming out-links. Pollutants continue spreading to C after arriving at B. Due to the distance increase and concentration, propagating capacity of pollutants decreases continuously.

![Fig.2 Schematic diagram of pollutant transmission characteristics](image)

3. Mining Key Nodes Based on Propagation Characteristics

Key nodes mining depends on node importance ranking. The location and relationship of local regions are the basic indicators of node importance measurement. In this section, BiTr key node mining method is proposed according to the characteristics of air quality network. The key nodes mined out in this method integrating network structure and the propagation characteristics together.

3.1. BiTr algorithm

BiTr algorithm calculates node importance based on the basic properties of nodes and the interaction process of network links. Nodes with higher importance are selected as key nodes. According to the characteristics, both out-link and in-link are used. According to the attenuation characteristics of the propagation process, the attenuation factor is introduced to decrease the influence level by level. The algorithm is based on the basic structure of the directed weighted network, and the global information of the network is considered in combination with the actual propagation process of pollutants.

The main steps of the algorithm are as follows:

1. Traverse network structure, each node is given the same initial importance (set as 1).
(2) Importance transmission and feedback. In the process of transmission, the importance of each node is accumulated by in-link importance. In-link importance is calculated according to the importance and out-weight of its source node. After one transmission, the importance of node a is updated as:

\[ IM^1_a = \sum_{b \in \text{in}(a)} \frac{1}{\text{deg}(b)} \cdot \frac{w_{ba}}{w_{out}(b)} \]  

(2)

\( \text{in}(a) \) is the source node set of a’s in-links, and \( \text{deg}(b) \) is the degree of node b. In addition, in the air quality network, the more out-links, the more important the node is. Out-link has obvious feedback effect, so we take it as a calculation parameter. Then parameter \( \beta \) is introduced to adjust the feedback weight. After one transmission and feedback, node importance is updated to

\[ IM^1_a = \sum_{b \in \text{in}(a)} \frac{1}{\text{deg}(b)} \cdot \frac{w_{ba}}{w_{out}(b)} + \beta \sum_{c \in \text{out}(a)} \frac{1}{\text{deg}(c)} \cdot \frac{w_{ca}}{w_{in}(c)} \]  

(3)

\( \text{out}(a) \) is the target node set of out-links.

(3) Continue iterating and repeat step (2) until node ranking is stable or the specified number of iterations is reached. At the end of each iteration, check whether node ranking sequence changes or not. When node ranking sequence remains unchanged for three consecutive iterations, it is considered to be stable. After \( t \) iterations, the importance of node a is as follows:

\[ IM^t_a = \sum_{u \in \text{in}(a)} \frac{1}{\text{deg}(b)} \cdot \frac{w_{ba}}{w_{out}(b)} + \beta \sum_{c \in \text{out}(a)} \frac{1}{\text{deg}(c)} \cdot \frac{w_{ca}}{w_{in}(c)} \]  

(4)

(4) Accumulate node importance in each iteration. At the same time, attenuation factor is introduced. The importance of the current moment is not attenuated, while that of each time before decays continuously at the rate of \( \alpha \). BiTr importance is defined as:

\[ \text{BiTr}^t_a = \sum_{i=1}^{t} \alpha^{t-i} \cdot IM^i_a \]  

(5)

BiTr importance comprehensively reflects the basic properties of the input and output of nodes, and accumulates the importance at different times to reduce the randomness and improve the accuracy.

3.2. ITP model

Various key node mining methods lead to different corresponding mining results. However, in key node mining field, it lacks generally accepted criteria, especially in directed weighted networks. IC[10] and LT model[11] are often used as evaluation criteria. However, these two models can only be applied in undirected and unweighted networks. In order to measure the key nodes mining efficiency in air quality directed weighted network, we design the ITP model.

In this model, nodes are stipulated to have only three states: active-enabled, active-disabled and inactive. The active-enabled nodes will remain active, and it tries to activate its neighbor nodes at the next moment after being activated, turning to the active disabled state after one activation process. When the activation value is greater than threshold, this neighbor node is activated successfully and turns to the active-enabled state. The activation process of ITP model is as follows:

(1) Algorithm initialization. Set the activation enabling node set \( \text{sou} \), activation node set \( \text{act} \), activation time \( t \) and threshold. In the initial state, \( \text{act} = \text{sou} \). In this paper, \( \text{source} \) is set as the key nodes set mined out by BiTr algorithm.

(2) Node activation. Traverse all active nodes. For node \( a \), it will try to activate its first-order neighbor \( \text{OutN}(a) \). The activation value of node \( a \) to its adjacent node \( b \) is \( \text{ActVal}_{ab} \). When \( \text{ActVal}_{ab} \) is greater than the threshold, node \( b \) will be activated. Add node \( b \) to the \( \text{sou} \) and \( \text{act} \) set. When node a activates its neighbor node, it is deleted from \( \text{sou} \) set. In order to better simulate the randomness of connection establishment in reality, a random number \( \text{rand} \) between 0 and 1 is introduced. The equation of activation value is shown in Eq6.

\[ \text{ActVal}_{ab} = \frac{w_{er}}{\sum_{j \in \text{NetOut}(s)} w_{st}} \cdot \text{rand} \]  

(6)
In Eq(6), $NeiOut(s)$ is the set of out neighbor nodes of s. The random value makes the model accord with the propagation situation in reality, and its value range is $[0,1]$.

3) Iterative propagation. When the specified time $t$ is not reached and there is an inactive node, repeat step (2).

4) Activation ends. The nodes in set $act$ are all active nodes. $act$ set is the influence range of the initial node set.

Fig.3 shows the specific process of ITP model. In the initial state $sou=act=\{5\}$, at time $t=1$, node 5 tries to active its out neighbor nodes $\{4,6,8\}$ one by one, influenced by out-link weight, node 4 and 6 is active. $act$ becomes $\{4,5,6\}$ and node5 transforms to active-disabled. When $t=2$, $sou=\{4,6\}$, then traverse the out neighbor nodes $\{1,2,7,9\}$ corresponding to node 4 and 6, with 2 nodes being activated. At last $act=\{1,4,5,6,9\}$ and this is the influence range of node 5 in 2 steps.

4. Experiment
We select the PM2.5 pollutant propagation relationship in Jing-Jin-Ji region to establish air quality network. Jing-Jin-Ji region spans about 500 kilometers from east to west and 950 kilometers from north to south. In network construction, the average distance between monitoring stations is taken as the farthest possible propagation distance between local regions, and the shortest propagation distance is set as 20km. The detailed structure of PM2.5 air quality directed weighted network is shown in Fig.4.
In Fig. 4, link weight is represented by its corresponding arrow size. The larger the arrow, the greater the weight. It is easy to get the relationship between the input and output of pollutants from Fig. 4. From the network structure diagram, we can see that the link weights between nodes differ a lot, and the link density also varies, which reflects the heterogeneity of network structure. This is the theoretical basis for node importance ranking and key node mining. BiTr algorithm is used to mine the key nodes in the network. Closeness and PageRank algorithm are selected as the comparative algorithm. The effectiveness of the algorithm is evaluated through the analysis of Top-10 key nodes and ITP model.

(1) Top-10 key node mining results

Top-10 key nodes are the most important 10 nodes in the whole network, which play a key role in maintaining the overall network structure and transmitting information. Table 1 lists the key nodes in mined by BiTr and the comparative algorithms.

| Rank | Station number | Over standard rate | Station number | Over standard rate | Station number | Over standard rate |
|------|----------------|-------------------|----------------|-------------------|----------------|-------------------|
| 1    | 1017           | 31.35             | 22007          | 24.93             | 1023           | 24.74             |
| 2    | 1018           | 31.76             | 17018          | 23.32             | 1019           | 20.02             |
| 3    | 1005           | 30.34             | 22005          | 34.10             | 6013           | 19.93             |
| 4    | 1025           | 20.41             | 22009          | 26.07             | 1017           | 31.35             |
| 5    | 17009          | 28.72             | 17009          | 28.72             | 1005           | 30.34             |
| 6    | 17007          | 30.71             | 21006          | 27.47             | 6026           | 29.12             |
| 7    | 1023           | 24.74             | 6025           | 22.04             | 6001           | 26.68             |
| 8    | 6001           | 26.68             | 17012          | 24.18             | 21016          | 23.11             |
| 9    | 14009          | 20.66             | 1036           | 33.05             | 1018           | 31.76             |
| 10   | 17018          | 23.32             | 17026          | 25.09             | 13016          | 25.53             |

It can be seen that the mining results of BiTr and PageRank algorithms are relatively close while the closeness algorithm is quite different from the above two methods. In addition, analyzing the pollutant exceeding rate of nodes, we take the station whose exceeding rate is higher than 30% as severe polluted sites. The number of seriously polluted sites mined by the three algorithms are 4, 2 and 3 respectively. BiTr algorithm mine out the most severely polluted sites.

(2) Activated node number in ITP model

![Fig. 5 Comparison of key nodes activation ability in different algorithms](image)

The number of activated nodes in ITP model is the evaluation criteria of key nodes mining result, and the Top-10 key nodes are taken as the initial active-enable node set. In order to reduce the
randomness, each algorithm runs 1000 times and the average value is taken. The number of active nodes is shown in Fig.5. BiTr algorithm always active more nodes than closeness and PageRank algorithms in the set time.

According to the results of the two experiments, the BiTr algorithm proposed in this paper mines out key nodes with high pollution degree and strong propagation power. These nodes occupy a more important position in the network. This algorithm accords with the actual situation of air quality system, and the mining results are accurate. Experiment (1) and (2) show that the mining results of BiTr algorithm are in line with the actual distribution of pollutants. The key stations have high proportion of pollutants concentration, having strong transmission ability as heavy polluted region.

5. Summary
Air pollutants are easily diffusing and widely distributed. Mining key nodes in air quality network can effectively improve the efficiency of pollutant treatment. In this paper, based on the construction of air quality network, the out-link is taken as part of the importance measurement index, and node importance with attenuation is accumulated. The results of the mined Top-10 nodes and the number of activated nodes in ITP algorithm prove that BiTr algorithm is effective and suitable for air quality network. In future research, we should build a more reasonable network structure, add pollutant concentration as node attribute to measure node importance better.

Acknowledgments
This work was supported by the Tianjin Natural Science Foundation (19JCYBJC15300).

References
[1] D Chen, L Lv, M S Shang, et al. Identifying Influential Nodes in Complex Networks[J]. Physica A Statistical Mechanics & Its Applications, 2012, 391(4): 1777-1787.
[2] S X Zhao , R Rousseau , F Y Ye. H-Degree as A Basic Measure in Weighted Networks[J]. Journal of Informetrics, 2011, 5(4): 668-677.
[3] L C Freeman. A Set of Measures of Centrality Based on Betweenness[J]. Sociometry, 1977, 40( 1): 35-41.
[4] C Song, Huang G, Zhang B, et al. A node influence ranking algorithm based on probability walking model[J]. International Journal of Modern Physics B, 2019, 33(13):1950132.
[5] X L Ren, L Y Lv. Review of node importance ranking methods[J]. Chinese Science Bulletin, 2014(13): 1175-1197(in Chinese).
[6] P F Bonacich. Factoring and weighting approaches to status scores and clique identification[J]. Journal of Mathematical Sociology, 1972, 2(1):113-120.
[7] K Zhang, P Li, B Zhu, et al. Evaluation Method for Node Importance in Directed-weighted Complex Networks Based on PageRank [J]. Journal of Nanjing University of Aeronautics & Astronautics, 2013, 45(3): 429-434.
[8] R Lempel, S Moran. The stochastic approach for link-structure analysis ( SALSA) and the TKC effect[J]. Computer Networks, 2000, 33: 387-401.
[9] L Y Lv, Y C Zhang, C H Yeung, et al. Leaders in Social Networks, the Delicious Case[J]. Plos One, 2011, 6(6): e21202.
[10] A Barrat , M Weigt . On the Properties of Small-world Network Models[J]. European Physical Journal B Condensed Matter & Complex Systems, 2000, 13(3): 547-560.
[11] S Hong, H Yang, G Li, et al. Analysis of propagation dynamics in complex dynamical network based on disturbance propagation model[J]. International Journal of Modern Physics B, 2014, 28(22):763-770.