ICIC Express Letters
Part B: Applications
Volume **8**, Number **2**, February **2017**

ICIC International ©2017 ISSN 2185-2766
pp. **387–394**

AN APPROACH TO FIND CRITICAL TRANSPORT PATHS OF AIR POLLUTANT BASED ON COMPLEX NETWORK

Wenyuan Liu[1], Xilong Kang[1], Jiaxin Yu[1], Xiaoli Wang[2]
Wenbo Song[2], Yanan Wei[2] and Liang Zhang[2]

[1]School of Information Science and Engineering
Yanshan University
No. 438, Hebei Avenue, Qinhuangdao 066004, P. R. China
wyliu@vip.163.com; jxyu@ysu.edu.cn

[2]Hebei Province Environmental Emergency and Heavy Pollution Weather Forewarning Center
1104 Box, Yaqing Street 30, Yuhua District, Shijiazhuang 050037, P. R. China

Received July 2016; accepted October 2016

Abstract. *The critical transport paths of air pollutants play an important role in optimizing the regional monitoring network and forecasting air quality. A Complex Network-based model of air pollutions is put forward to Find Critical Transport Paths (FCTP) in area. First, nodes in network model proposed are mapped as air quality monitoring stations, establishing edge by air pollutants dispersion. Therefore, the spread process of air pollutants in the monitoring sites of Beijing-Tianjin-Hebei (BTH) can be abstracted as a Complex Network. Second, according to the computation of similarity among paths, fuzzy similarity matrix is established to calculate the similarity among transport paths. Third, an efficient method is presented to cluster paths through structural decomposition of fuzzy similarity matrices. At last, a set of critical transport paths is derived. The experimental results show that the proposed algorithm – FCTP is efficient in mining critical transport path under air pollutants Complex Network.*

Keywords: Complex Network, Critical transport path, Fuzzy similarity metric, Clustering

1. **Introduction.** In recent years, with the rapid economic and social development, the BTH region has become one of China's most economically developed regions; meanwhile, the rapid program of industrial development has resulted in frequent heavy haze pollution. The haze pollution can not only delay the traffic but also induce serious health problems. Many studies have shown that air pollution in urban cites is caused not only by local emission sources but also significantly by regional air pollutant transport from surrounding areas [1,2]. Thus, how to mine the critical transport paths of air pollutants becomes an urgent problem which remains to optimize the air monitoring network to predict air quality in the future.

Therefore, mining the critical transport paths to Beijing from historical air quality, meteorological and geographic data in BTH is important. However, available studies on the critical transport paths of air pollutants have mostly focused on the air quality model. The 5th-generation Mesoscale Model (MM5) [2] and Weather Research Forecasting (WRF) meteorological models [3] analogs meteorology, the Comprehensive Air-quality Model with extensions (CAMx) [4] and Community Multiscale Air Quality (CMAQ) model [2,3,5] were applied to simulating air pollutants concentrations to mining transport paths. All the models should be based on the detailed data of pollution sources to obtain good simulation effect, but collecting fine data is very difficult.

Today, Complex Network is always a research hotspot [6]. Many studies show that the spread of epidemic and rumors [7] between nodes have great progress. However, researching the spread of air pollutants is few based on Complex Network model, because

of fledgling domestic pollutant monitoring. Many researchers study the characteristics of Complex Network. Studies show that the Complex Network phenomena, namely small-world effect and scale-free property. Huang et al. provided a critical execution paths discovery strategy in scale-free software execution network [8]. Because air pollutions are suspended in the air, the spread mechanism of air pollutants is in accordance with dynamic mechanism of Complex Network [9]. There is a new way of thinking on transport path of air pollutants dynamic network and its application.

The contribution of this paper is an FCTP approach that is presented to find critical transport paths in air pollutant transport network. First, air monitoring stations [10] are abstracted into nodes in Complex Network. Second, establish directed edge based on the factors such as meteorological, altitude and horizontal distance [11] by air pollutants dispersion model [12], and then weight of edge is yet gained. Third, a directed-weighted Complex Network of pollutants dissemination in BTH monitoring sites is created via Air Pollutants Transport Complex Network (APTCN) algorithm. Fourth, the paths were obtained from Air Pollutants Transport Complex Network by defined source and sink nodes. Then path similarity matrix is formed by calculating similarity of paths and fuzzy similarity matrix is created to transform path similarity matrix. Finally, the critical transport path set is obtained by structure modeling clustering [13] method to cluster paths in the fuzzy similarity matrix.

The rest of this paper is organized as follows. Section 2 suggests some assumptions, introduces some definitions and presents Complex Network algorithm. Section 3 presents the structure modeling clustering algorithm, along with the generation of transport paths, the calculation of path similarity and fuzzy similarity matrix of paths. The experiments on datasets are conducted in Section 4. Thus we conclude this paper in Section 5.

2. Pollutant Transport Network Modeling.

Air pollutions transport is governed by many factors and processes. To avoid model being too complex, do the following assumptions.

1) Nodes represent air monitoring stations and edges denote the interactions or relationships between monitoring station and weight on edges indicates if there exists the largest throughput relationship between each pair of station. Then the network can be described by triple $G = \{V, A, W\}$, among which, $V = \{v_1, v_2, \ldots, v_n\}$ is the set of nodes in the network, $A = \{a_{ij}; i, j = 1, 2, \ldots, n\}$ is the set of edges and $W = \{w_{ij}; i, j = 1, 2, \ldots, n\}$ is the weight set of each edge.

2) If edge $a_{ij} \in A$ as well as the horizontal distance between node i and j is more than 50km, then a_{ij} and a_{ji} will be disconnected. The distance of air pollutants diffusion in the horizontal direction is defined due to the following two reasons. First, air pollutants have long-distance transport characteristic, and the transport distance is more than 50km [2]. Second, taking account of the distribution of air monitoring stations in the BTH – there has only been one air monitoring station in each county in which most of them are close to 50km.

3) If edge $a_{ij} \in A$ as well as the altitude between node i and j is more than 100m, a_{ij} will be disconnected. Air pollutants have a significant diffusion effect within 100m in the vertical direction [14].

4) If edge $a_{ij} \in A$ and a_{ij} is not the largest throughput, it will be disconnected. The largest throughput can make certain the air pollutants spread is the main path in the period T.

Definition 2.1. (Weight on edges). *Along with farther distance, greater altitude, smaller density, the difficulty diffusion of pollutants is more [15]. And wind direction decides the direction of diffusion direction of air pollutants [13]. If two nodes i and j ($i, j \in V$ and $i \neq j$), on which its altitude difference is $\Delta H_{ij}(m)$, horizontal distance is*

$\Delta S_{ij}(km)$, and Air Quality Index (AQI) difference is ΔAQI_{ij}, following gas diffusion theory and the above assumptions, w_{ij} can be stated as below at t time period.

$$w_{ij}(t) = \gamma \frac{R_{ij}(t)\,|\Delta AQI_{ij}(t)|}{|\Delta H_{ij}\Delta S_{ij}|} \pm \varepsilon(t) \tag{1}$$

In Formula (1), γ is correction coefficient with values between the interval $(0,1)$, and $\varepsilon(t)$ is the maximum fluctuant value of $w_{ij}(t)$ in a period. Suppose that $\vec{Q_{ij}}$ is direction of node i to node j, $\vec{F(t)}$ is wind of node i, unit (m/s), and $\theta_{ij}(t)$ is included angle between $\vec{Q_{ij}}$ and $\vec{F(t)}$; according to the theory of atmospheric flow mechanics, wind coefficient is defined as the following formula.

$$R_{ij}(t) = \begin{cases} 0 & if\ \pi/2 \le \theta_{ij}(t) < \pi \\ \left|\vec{F}\right| \cos\theta_{ij}(t) & if\ 0 \le \theta_{ij}(t) < \pi/2 \end{cases} \tag{2}$$

Definition 2.2. (Weight on nodes). *It is easy to know the importance of a node by the weight of edge; the more in-degrees there exist in node, the more numbers there are of transmissions, and the more significant it is. Weight on node is expressed as*

$$w_{in}^{v_i} = \sum_{v_j \in V, v_j \to v_i} w_{ji}(t) \tag{3}$$

In Formula (3) $v_i \in V$, the edge a_{ji} exists and the weight also exists. Adjacency matrix of air pollutants spread at time t, $(t \in T)$; assume that T is a period, which is defined as

$$m(t) = (a_{ij})_{N\times N} = \begin{cases} w_{ij}(t) & if\, Max(w_{ij}(t))\ exists \\ 0 & if\ no\ w_{ij}(t)\ exists \end{cases} \tag{4}$$

Accumulating Formula (4) in the period T, it can be represented as $M = \sum_{t\in T} m(t)$.

Definition 2.3. (Source nodes). *The monitoring stations are long-suffering heavy pollution (AQI of daily average more than 200) in Hebei province and Tianjin at a period T.*

Definition 2.4. (Sink nodes). *All monitoring stations in Beijing are chosen as sink nodes.*

Definition 2.5. (The coverage rate of critical transport paths). *The paths before clustering are all the transport paths of air pollutants from different cities to Beijing, and the critical paths are the paths after clustering from different cities to Beijing. Assume the coverage rate of critical transport paths can be defined as CR, which is defined as*

$$CR = \frac{The\ spread\ times\ of\ critical\ paths}{The\ spread\ times\ of\ paths\ before\ clustering} \tag{5}$$

The procedure of Pollutants Transport Complex Network algorithm is described as follows.

Algorithm 1: *Pollutants Transport Complex Network*

Input: *Datasets; T; All stations V;*
Output: *$G_T = \{V, A, W\}$;*
1. Traverse T, $(t \in T)\{$
2. Get source nodes V_t at time t, and traverse V_t $\{$
3. If source node V_{ti} and node V_j, $\Delta S_{ij} < 50km$ && $\Delta H_{ij} < 100m$ && $Max(w_{ij}(t))$ $\{a_{ij} = w_{ij}(t), a_{ij} \in A;\}$
4. Obtain adjacency matrix of air pollutants spread $m(t) = (a_{ij})_{N\times N};\}$
5. Calculate $M_T = (a_{ij})_{N\times N} = \sum_{t\in T} m(t); w_{ij}(T) = a_{ij}, w_{ij}(T) \in W;\}$

3. Fuzzy Clustering.

3.1. **Fuzzy similarity matrix.** Path similarity is displayed in the form of similarity matrix. Then, fuzzy similarity matrix is created by standardizing path similarity matrix, which is the input of fuzzy clustering adopted in the work. Path is obtained based on the network spread of air pollutions by depth-first-search algorithm from source nodes to sink nodes. Many algorithms can achieve fuzzy similarity matrix, for example, quantity product method and nearness degree method. The path similarity model is defined as product of the longest common subsequence path and weight of node, which reflects both of the node spread and the importance of the node itself simultaneously. Assume that the nodes path is defined as $t_i = \{t_{i1}, t_{i2}, \ldots, t_{in}\}$ $(i = 1, 2, \ldots, m)$, n is the number of nodes in path t_i and the sequence in path yet represents the direction of path. m is the number of all paths. Path similarity matrix of all paths is defined as

$$station\_node = \begin{bmatrix} sn_{11} & sn_{12} & \cdots & sn_{1m} \\ sn_{21} & sn_{22} & \cdots & sn_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ sn_{m1} & sn_{m2} & \cdots & sn_{mm} \end{bmatrix} \tag{6}$$

In Formula (6), sn_{ij} is the sum of weight on common nodes multiplying the number of transmission for path $t_i = \{t_{i1}, t_{i2}, \ldots, t_{ip}\}$ and $t_j = \{t_{j1}, t_{j2}, \ldots, t_{jq}\}$. p, q are the number of node in path. The $|MaxLCS(t_i, t_j)|$ means the length of corresponding sequence, if $MaxLCS(t_i, t_j)$ denotes the longest common subsequence. The similarity of path is described as follows.

$$Sim(t_i, t_j) = \sum_{k=1}^{|MaxLCS(t_i,t_j)|} n_k w_{in}^{v_k},\ v_k \in MaxLCS(t_i, t_j) \tag{7}$$

In Formula (7), n_k is the number of transmission to v_k in Complex Network of air pollutants. Afterwards, standardize matrix *station_node* elimination dimension, forming a $[0, 1]$ matrix, which is fuzzy similarity matrix.

3.2. **Fuzzy clustering of path.** Fuzzy clustering method based on fuzzy similarity matrix is used commonly, for example, Fuzzy C-means clustering. Structure modeling clustering [16] method is employed to cluster path, due to the fact that the spread of air pollutions is structured. This system sets threshold α for matrix *station_node*, *s_n_s* denotes initial fuzzy similarity matrix, and then decomposition to it.

$$s\_n\_s[i][j] = \begin{cases} 1, & s\_n\_s[i][j] \geq \alpha \\ 0, & s\_n\_s[i][j] < \alpha \end{cases} \tag{8}$$

So, the element of matrix is 0 or 1. Subclass and its relationship are obtained from diagonal matrix or lower triangular matrix by transforming fuzzy similarity matrix. All of the above operations on matrices are completed in order to cluster the related paths together. The tightness of path is described as follows.

$$Close\_path(t_x, t_y) = \sum_{i=1}^{m} s\_n\_s[x][i] * s\_n\_s[y][i] \tag{9}$$

Finally, the result makes the maximum value of $\sum_{x=1}^{m} \sum_{y=1}^{m} Close\_path(t_x, t_y)$ by transformation matrix. Then, the adjacency matrix $s\_n\_s[m][m]$ is divided into partitioned matrix by changing what contains the same type of path. Following is the algorithm of path clustering.

Algorithm 2: *Structure Modeling Clustering*

Input: *Path similarity matrix;*
Output: *Path clusters;*
1. Calculating similarity matrix station_node;
2. The longest common subsequence $P_1, P_2, \ldots, P_n$ *of path i and path j, weight is* $\omega_1, \omega_2, \ldots, \omega_n$*;*
3. $station\_node = P_1 * \omega_1 + P_2 * \omega_2 + \cdots + P_n * \omega_n$*;*
4. The value is 1 of diagonal element by normalization and standardization;
5. Calculating reachability matrix, then structure decomposition;
6. Selecting diagonal submatrix as a class element.

4. Experiment Analysis.

4.1. **Datasets.** In the evaluation, the following three real datasets were used, where the three sources are available in Hebei Province Environmental Emergency and Heavy Pollution Weather Forewarning Center (HPEEPWFC). In total, air quality instances have been collected from June 1, 2014 to June 1, 2015.

1) Air quality data: Figure 2(b) presents the geographical distribution of 251 stations in BTH, where each dot for a station, and each larger stand for a source station. Each instance consists of the concentration of six air pollutants: NO_2, SO_2, O_3, CO, $PM_{2.5}$ and PM_{10} that are converted into corresponding daily average AQI for each air pollutant according to the Chinese AQI standard.

2) Meteorological data: Each meteorological instance consists of temperature, humidity, wind speed and direction every day. The time scope is the same with air quality data.

3) Distance and elevation: Distance of both air quality stations can be computed by latitude and longitude coordinates. Elevation can be gained by Google earth.

As shown in Table 1, the sources are generated by calculating the excessive days (AQI $>$ 200) every station in cities of Hebei province and Tianjin.

Table 1. The description of source nodes

Number	City	Station name	Excessive days
1	Shijiazhuang	S1	117
2	Baoding	S2	144
3	Handan	S3	118
4	Xingtai	S4	104
5	Hengshui	S5	99
6	Langfang	S6	84
7	Tangshan	S7	75
8	Zhangjiakou	S8	21
9	Qinhuangdao	S9	81
10	Chengde	S10	36
11	Cangzhou	S11	90
12	Tianjin	S12	88

4.2. **Results.** Figure 1 implies the degree distribution to draw conclusion of scale-free property that spread of air pollutions complex networks own.

Since weight on edge has been introduced in the paper, the result is relatively good for parameter γ is set to be 0.1 and parameter α is set to be 0.5 by analysis of simulations. Learnt from the analysis of Table 2, the number of paths before clustering is 491, which turns to be 47 transport paths after clustering algorithm. The rate of the test paths

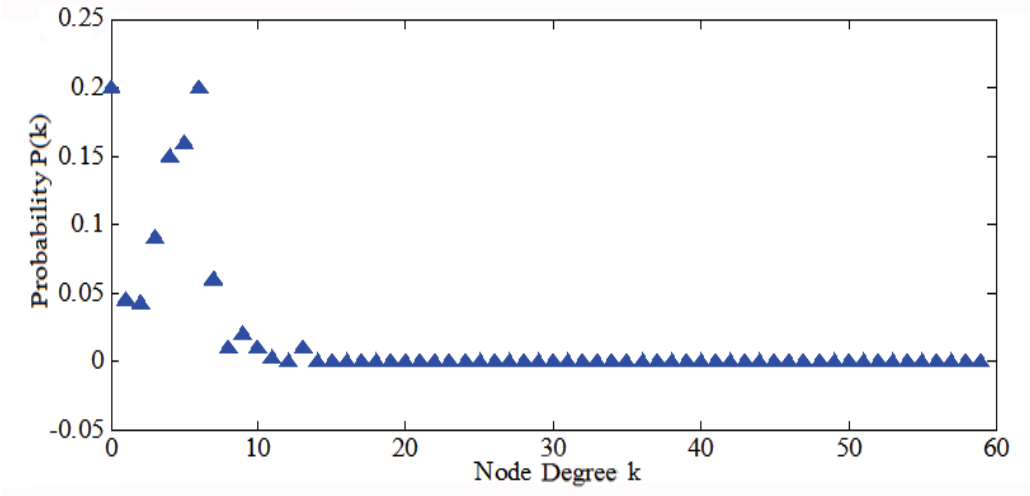
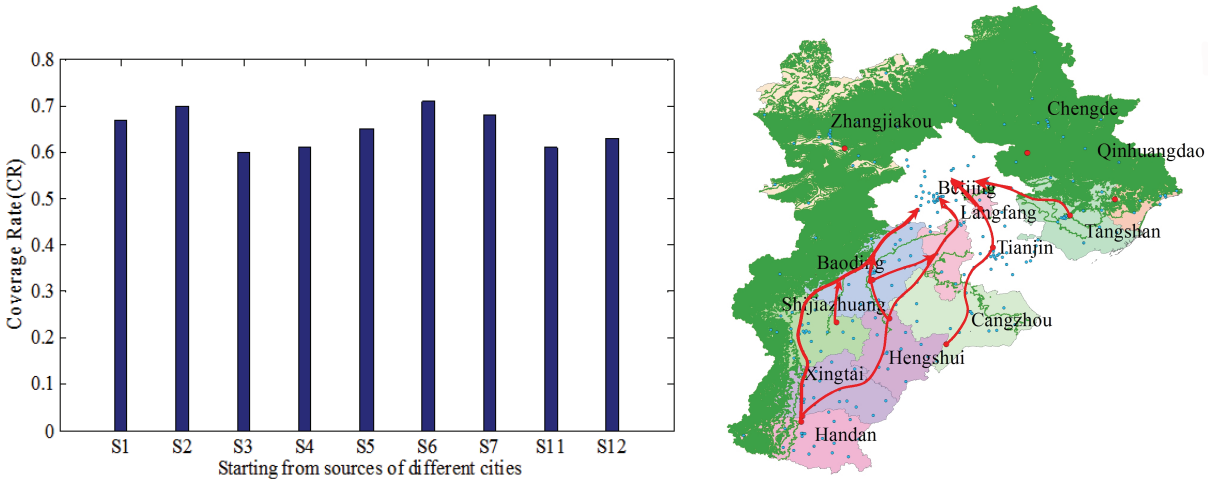


FIGURE 1. Degree distribution of air pollutants transport network in the plot

TABLE 2. Result of transport paths found in air pollutants transport network

Starting from different cities sources nodes	Number of paths before clustering	Number of clusters	Number of critical transport paths	Station coverage rate (%)	Path reduction rate (%)
S1	85	14	14	63	83.53
S2	100	9	9	66	91.00
S3	60	6	6	60	90.00
S4	61	5	5	67	91.80
S5	50	3	3	60	94.00
S6	47	3	3	30	93.62
S7	25	2	2	50	92.00
S8	0	0	0	0	0.00
S9	0	0	0	0	0.00
S10	0	0	0	0	0.00
S11	31	3	3	43	90.32
S12	32	2	2	37	93.75
Total	491	47	47	39.7	90.34



(a) Coverage rate of critical transport paths (b) The longest critical transport paths

FIGURE 2. The description of the critical transport paths

reduces by 90.34%. Zhangjiakou, Chengde, and Qinhuangdao do not show transport path of air pollutants to Beijing.

According to Definition 2.5, Figure 2(a) is trend chart of the coverage rate of critical transport paths (CR), and the spread times of air pollutants on the paths from different cities to Beijing which are calculated by heavy haze pollution process in BTH. Seen from this, the coverage rates indicate that critical transport paths have a good representation. Therefore, the proposed algorithm – FCTP is efficient in mining critical transport path under air pollutants Complex Network.

Figure 2(b) shows the longest critical transport paths of air pollutants, where each line (exist overlap) stands for a path, and the upper and left dark areas stand for the mountains. It can be seen from the path that the main path is from the southwest transport on the lee of the Taihang Mountains to Beijing, and a transport path is from the southern Cangzhou, Tianjin and Langfang to Beijing. Pollutants transport by Yan Mountains from Tangshan to Beijing in the northeast. It can be seen of consistency in pollutants critical transport paths and regional geography and atmospheric flow.

5. **Conclusions.** In conclusion, an FCTP algorithm based on spread of air pollutions network was proposed. A weighted Complex Network is constructed for dynamic process of air pollutants dispersion. And the sites of path are found according to the number of source and sink nodes. The structure modeling clustering algorithm is adopted to cluster paths on path similarity matrices for the first time. This process is in favor for correct path division under the condition of certain scale of transport network, which indirectly proves that the path similarity defined previously is practical. In the end, the longest path was selected in each cluster as the critical transport path. Then critical transport path set is formed, which shows consistency in pollutants critical transport paths and regional geography, and atmospheric flow. The FCTP algorithm is efficient in mining critical transport path under air pollutants Complex Network. Follow-up studies may focus on how to mine key nodes by critical transport paths in the network of spread of air pollutions, which is beneficial to optimize the regional monitoring network to predict air quality in the future.

Acknowledgments. This work is supported by National Science and Technology Infrastructure Program (2014BAC23B00, 2014BAC23B01).

REFERENCES

[1] M. C. Bove et al., An integrated PM2.5 source apportionment study: Positive matrix factorisation vs. the chemical transport model CAMx, *Atmospheric Environment*, vol.94, pp.274-286, 2014.

[2] F. Wang et al., Identification of regional atmospheric PM10 transport pathways using HYSPLIT, MM5-CMAQ and synoptic pressure pattern analysis, *Environmental Modelling & Software*, vol.25, pp.927-934, 2010.

[3] S. Yu et al., Comparative evaluation of the impact of WRF-NMM and WRF-ARW meteorology on CMAQ simulations for O_3 and related species during the 2006 TexAQS/GoMACCS campaign, *Atmospheric Pollution Research*, vol.3, pp.149-162, 2012.

[4] H.-J. In et al., Impact of transboundary transport of carbonaceous aerosols on the regional air quality in the United States: A case study of the South American wildland fire of May 1998, *Journal of Geophysical Research-Atmospheres*, vol.112, 2007.

[5] S. Vardoulakis et al., Modelling air quality in street canyons: A review, *Atmospheric Environment*, vol.37, pp.155-182, 2003.

[6] S. Boccaletti et al., Complex networks: Structure and dynamics, *Physics Reports*, vol.424, pp.175-308, 2006.

[7] R. Pastor-Satorras and A. Vespignani, Epidemic dynamics and endemic states in complex networks, *Physical Review E Statistical Nonlinear & Soft Matter Physics*, vol.63, no.2, p.066117, 2001.

[8] G. Y. Huang et al., An algorithm to find critical execution paths of software based on complex network, *International Journal of Modern Physics C*, vol.26, no.9, 2015.

[9] Y. Moreno et al., Dynamics of rumor spreading in complex networks, *Physical Review E Statistical Nonlinear & Soft Matter Physics*, vol.69, 2004.

[10] Z. Yu, F. Liu and H.-P. Hsieh, U-Air: When urban air quality inference meets big data, *Proc. of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013.

[11] X. Li et al., Recent progress in CFD modelling of wind field and pollutant transport in street canyons, *Atmospheric Environment*, vol.40, pp.5640-5658, 2006.

[12] M. Sharan et al., A mathematical model for the dispersion of air pollutants in low wind conditions, *Atmospheric Environment*, vol.30, no.8, pp.1209-1220, 1996.

[13] X. Xu et al., SCAN: A structural clustering algorithm for networks, *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, CA, USA, 2007.

[14] F. Gebali and M. F. Yassin, Data mining model for atmospheric pollutants from elevated point sources, *Proc. of the 2009 Conference on Information Science, Technology and Applications*, Kuwait, 2009.

[15] J. Li et al., Assessing the effects of trans-boundary aerosol transport between various city clusters on regional haze episodes in spring over East China, *Tellus Series B – Chemical and Physical Meteorology*, vol.65, 2013.

[16] J. Huang et al., SHRINK: A structural clustering algorithm for detecting hierarchical communities in networks, *Proc. of the 19th ACM International Conference on Information and Knowledge Management*, Toronto, ON, Canada, 2010.