Team Control Number

31262

For office use only

T1 _____

T2 _____

T3 _____

T4 _____

Problem Chosen

C

For office use only

F1 _____

F2 _____

F3 _____

F4 _____

2014 Mathematical Contest in Modeling (MCM) Summary Sheet

A Three-dimensional Network Impact Analysis Model
Based on Centralizing, Connecting and Spreading Characteristics
ABSTRACT

Last decade has witnessed a burst in the research of network impact analysis. However, most of the previous research focused on single factor or single algorithm to analyze the impact, which is insufficient for complex networks.

According to our observation and correlation analysis, we propose a characteristic classification method to systematically construct a three-dimension network impact analysis model. We establish the concept of centralizing characteristics, connecting characteristics and spreading characteristics, each of which consists of three sub-characteristics. Sub-characteristics include degree, eigenvector centrality, PageRank algorithm, betweenness centrality, clustering coefficient, node removal method, closeness centrality and two newly-proposed characteristics——spreading breadth index and spreading depth index both obtained from a submodel we design ourselves. Principal Component Analysis (PCA) is applied to obtain three one-dimension characteristic vector respectively. Finally a weighted sum of the three characteristic vectors is obtained to represent the impact measurement result for each node in the network.

Three datasets have been used for testing the rationality of the model and very promising performances have been measured. Additional efforts are made to extract the data, validate our model, visualize the network and discuss various utilities. In this way, we offer a rather comprehensive and reliable solution to this problem.

We strongly recommended our model because of its novel ideas, convincing analysis, exquisite visualization and promising performances.

Keywords

Network Impact Analysis; Centralizing Characteristics; Connecting Characteristics; Spreading Characteristics; Spreading Breadth Index; Spreading Depth Index; PCA.

A Three-dimensional Network Impact Analysis Model
Based on Centralizing, Connecting and Spreading Characteristics

Content

1. INTRODUCTION

In recent years, people are increasingly finding themselves bombarded with information. As with researchers, they have to filter huge mass of existing papers to find the most useful one. This situation calls for a method to help people analyze influence and impact, which correspondently leads to the burst in the study of social networks.

Most of the previous research focused on single factor to analyze the impact, which is insufficient for complex network. Our goal is to give a multi-factor impact measuring model for impact analysis in research network and extend the utility of our model to other areas of society. In addition, efforts is made to propose new analysis index, validate our model, visualize the network and discuss various utilities. In this way, we offer a rather reliable solution to this long-standing problem.

The paper is organized as follows. Section 2 contains a review of previous studies in the field of network impact analysis. Section 3 demonstrates our modeling method in details and systematically describes relevant algorithms, including two impact analysis index we propose——Spreading Breadth Index and Spreading Depth Index. We give result and analysis for the five tasks in section 4. Finally in section 5 we summarize the main contribution of the present paper and discuss the potential weakness and sensitivity.

2. BACKGROUND

Widely accepted and efficiently utilized in the real world, the concept of a network, having formed for a long time, gradually arouses people's interest to study and unearth some abstract characteristics of it. Some researchers, like Girvan, M., & Newman, M. E. (2002), are trying to depict the detailed structure of the network, while others, in spite of the same research direction, use some quantitative parameters to give an estimate of characteristics the network has. Bonacich, P. (2007), Latapy, M. (2008) and Yan, E., & Ding, Y. (2009) are among those people. They used a quantified system to describe the centrality, connection and other characteristics of the network.

Implementing some existing or new algorithms to operate on the network system in order to obtain some meaningful results is also attempted by some researchers, like Batagelj, V. (2003). However, they must make some modification to the existing algorithm to make it fit the research thought and approach in the field of network science.

As some creative and outstanding work completed by some researchers, like Jolliffe, I. (2005), some existing mathematical algorithms and methods become more rational, which allow us to use them to construct a more comprehensive appraising system to evaluate the characteristics of the network. In this article, we concentrate our effort on establishing a complete model to estimate the importance and influence of each vertice in a network.

3. METHODOLOGY

After deliberate study of previous research, we propose a characteristic classification method to systematically construct a three-dimension network impact analysis model. In our model, there are three types of characteristic parameters that are playing critical roles in impact analysis. They are centralizing, connecting, and spreading characteristics.

Centralizing characteristics focus on describing the extent of structural centrality for a vertice in its communities. In this type, there are degree, eigenvector centrality and PageRank, which reveal the feature of a vertice in the network from a similar angle.

Connecting characteristics, including betweenness centrality, clustering coefficient and node removal method, are to indicate how the vertice contribute to guarantee the connection of the network.

Spreading characteristics, such as closeness centrality, spreading breadth index and spreading depth index, are to quantify the efficiency of information dissemination for the whole network with the information starting from a particular vertice. Based on the fact that any vertice in our model is an information source (people or paper) and all any other vertices connected to it may receive the information released by it, we must think up a standard to assess the information spreading ability of each vertice. So we design a brand-new submodel (spreading breadth index and spreading depth index are the characteristic parameters obtained from it), along with an existing standard (closeness centrality), to describe the information spreading ability for each vertice in the network. This kind of feature naturally should be regarded as another aspect for vertice evaluation.

Our data-processing procedure is based on the classification method above. After getting the nine parameters of a vertice through computer programming or *gephi(a visualization software)* calculation, we classify the parameters into three groups according to above analysis. And for each group, we use Principal Component Analysis (PCA) to get a comprehensive result of the three parameters in a group. Now with three parameters obtained from three groups correspondingly after PCA, we use three weight factor to operate on the three parameters to work out a final evaluation result of a vertice.

3.1 Centralizing Characteristics
The first type of characteristics of a network is centralizing characteristics. Actually while taking vertex-influence-evaluation into consideration, the first idea coming to our mind is to check to what extent a vertex is in the center of the network. Centralizing characteristic parameters are defined to quantify such kind of extent

3.1.1 Degree Centrality (D)
Degree centrality, or degree, of a vertice equals to the number of edges a vertex has in common with other neighbor vertices. If there are totally D vertices and E edges in the network, the two sums have the following relation

$$D = 2E \tag{1}$$

Generally, the vertice with a higher degree or more connection edges is more central in structure and has the tendency to possess a greater ability to influence others. Those nodes should have a relatively more important role among all the nodes in the network.

3.1.2 Eigenvector Centrality (C_e)

In a network, if we merely use the degree centrality to describe the extent a vertex is located in the center, the standard could be too one-sided and we may miss some important features of the network.

As a result, eigenvector centrality is defined. In some networks, some vertices with a high degree are connected to lots of low-degree vertices and the eigenvector centrality is to quantify the extent of such situations. This parameter is defined to standardize the centrality of vertices from another angle.

3.1.3 PageRank (PR)
PageRank is initially proposed, over ten years ago, by Page and Brin (1998). This parameter is used to assess the rank or importance of a vertice in a network according to a method of iteration using the following equation:

$$PR(p) = (1 - d)\frac{1}{N} + d\sum_{i=1}^{k} \frac{PR(p_i)}{C(p_i)} \tag{2}$$

where N is the number of vertices in the network, d is a damping factor, and p_i is all other vertices linking to the selected vertice p. After continuous operations of iteration, each point refresh its PR once and once again. Finally all the vertices will have a weight to indicate the importance of it in the whole network.

3.2 Connecting Characteristics

The second type of characteristics of a network is connecting characteristics. For the characteristic parameters in this group, they are defined to indicate the contribution of a vertex to the connection and integrity of the whole network. In other words, they reveal how many times a vertice is located in a key position to make the network connected.

3.2.1 Betweenness Centrality (C_B)

Betweenness centrality is based on the number of shortest paths passing through a vertex. Vertices with a high betweenness play the role of connecting different groups. We use the following equation to define C_B :

$$C_{B(n_i)} = \sum_{j,k \neq i} \frac{g_{jik}}{g_{jk}} \tag{3}$$

In the equation (3), g_{jik} is all geodesics linking node j and node k which pass through node g_{jik} is the geodesic distance between the vertices of j and k.

In social networks, vertices with high betweenness are the brokers and connectors who bring others together (Yin et al., 2006). Being between means that a vertex has the ability to control the flow of information between most others. Individuals with high betweenness are the pivots in the network information flowing. The vertices with highest betweenness also result in the largest increase in typical distance between others when they are removed.

3.2.2 Clustering Coefficient (C_{coe})

A clustering coefficient measures the tendency of nodes in a graph to cluster together. The clustering coefficient of a vertex v (with a degree at least 2) is the probability that any two randomly chosen neighbors of v are linked together. It is computed by dividing the number of triangles containing v by the number of possible edges between its neighbors, i.e. $\binom{d(v)}{2}$, where $d(v)$ denotes the number of neighbors of v. We can then define the clustering coefficient of the whole network as the average of this value for all the vertices (with degree at least 2).

3.2.3 Node Removal Method and Total Loss (TL)

Another method to appraise the extent of connection of a given node in a connected network is node removal method. We can know the connection importance of a vertex by removing it from the network and then estimate the consequential loss. The final result, the value of the loss, can be used to evaluate the importance, with regard to connection characteristic, of the removed vertice.

If a vertice is removed from the network, two kinds of losses could be led to.

Self-Loss (SL) is based on the fact that after removing, all the vertices in the remaining network are not connected to the removed vertex anymore, and we could use the length of the shortest path from the removed vertex to other vertices to quantify the loss of the i-th vertice:

$$SL_i = \sum_{j \in \{1,2...N\}, j \neq i} \frac{1}{d_{ij}} \tag{4}$$

where d_{ij} is the distance between the two vertices with label i and j.

Mutual loss (ML) is the loss of disconnection caused by the removal of a vertex. Assume that the remaining network have K connected components and each component has $N_i (i = 1, 2...K)$ vertices, then there will be totally $\sum_{i=1}^{K} \sum_{j=i+1}^{K} N_i N_j$ pairs of disconnected vertices. We assume that all the pairs form a set S, and we could define ML_i for the i-th vertice, as:

$$ML_i = \sum_{j \in S} \frac{1}{d_j}$$

(5)

where j is an arbitrary pair disconnected vertices in the set S.

Finally we could define the total loss (TL) as:

$$TL_i = SL_i + ML_i$$

(6)

Practically, if we would like to calculate the value of TL_i, we should know the distance matrix, which could be obtained by Floyd Algorithm and each element in which stands for the distance between two vertices, before the removal(D) and after the removal(D'). Then the sum of the reciprocal of the non-zero elements in the first line is SL of the removed vertice. To get the ML, we check each element in D and find all the non-infinity elements, above the diagonal and not in the first line in D. Those elements form a set of T. Then we select out all the elements in T, which have values of infinity in the corresponding place in the matrix D' and we can get ML by calculating the sum of the reciprocal of those elements.

3.3 Spreading Characteristics

The third type of characteristics of a network is spreading characteristic, which we define to describe how information flow, such as academic resource, could be spread in the network between vertices. Establishing the following parameters to describe such characteristics is indispensable to estimate the extent of information spreading so as to evaluate the importance of each vertice to the whole network.

3.3.1 Closeness Centrality (C_c)

Closeness centrality is a sophisticated, however useful way to evaluate the characteristic of a vertice. An institutive fact is that for a vertice P, if most vertices in the whole network all have very large distances from it, it must be less central compared with a vertice Q with most vertices having smaller distances from it. So we could use the following equation to define the close centrality:

$$C_c(n_i) = \sum_{j=1}^{N} \frac{1}{d(n_i, n_j)}$$

(7)

Where $C_c(n_i)$ is the closeness centrality of the vertice and $d(n_i, n_j)$ is the distance, the length of the shortest path, between the two vertices in the network. In the equation, each distance between the two vertices contributes to the closeness centrality separately and determines the extent of centrality together.

Practically speaking, this parameter could be used to estimate whether it is easy or not to spread information from a given vertice to other vertices in the network.

3.3.2 Spreading Breadth Index (B_s) and Half-network Period (T_h)

For a given network with N vertices in total and a given vertice P, we define the spreading breadth index to help describe the information-spreading-efficiency of the network based on the vertice selected.

In the following several paragraphs, we will propose the submodel mentioned at the beginning of the article.

Assume that at time t=0, we release a particle (standing for a piece of information) at vertice P and it passes through exactly one edge per second to reach another vertice. If the particle meets a branch at a vertice with degree m, it will split into (m-1) particles and each of them will choose one of the m edges, except the one they come from, to go on moving. Then we set a timespan T (with the unit of second), and define $n(t)(t = 1,2,\ldots,T)$ to be the total number of

vertices that are being occupied by particles at the time t. Next we can suppose the expression of the spreading breadth index of the vertice P to be:

$$B_s(P) = \sum_{t=1}^{T} a(t) \frac{n(t)}{N-1}$$

(8)

where $(N-1)$ is the total number of vertices in the network except P and $a(t)$ is an attenuation factor of each term.

The attenuation factor $a(t)$ for the definition of the spreading breadth index is necessary. Take the spread of information as an example, if a vertice could let the information reach the same number of other vertices in a shorter time, this vertice is more significant, as a result of which for a vertice P, those other vertices reached by the particles released from P earlier should have a larger weight. So as time goes by, the particles reached later should contribute less to the spreading breadth index and should be multiplied by an attenuation factor to lessen their importance.

Now we must determine $a(t)$. First we assume that average degree of vertices in the network is D, and we could get the roughly estimated equation:

$$n(t+1) = (D-1)n(t)$$

(9)

because from time t to time $t+1$, each particle will become $(D-1)$ ones to go on passing into the branches except the one it comes from. From equation (9) we can know $n(t)$ has the form:

$$n(t) = n(1)(D-1)^{t-1}$$

(10)

Apply (10) to (8) we get:

$$B_s(P) = \frac{n(1)}{N-1} \sum_{t=1}^{T} a(t)(D-1)^{t-1}$$

(11)

Noticing that T, a changeable timespan, is only for testing, $B_s(p)$ should be independent of T. So we must let the term $(D-1)^{t-1}$ disappear. Assume:

$$a(t) = b(t) \frac{1}{(D-1)^{t-1}}$$

(12)

Apply (12) to (11) we get:

$$B_s(P) = \frac{n(1)}{N-1} \sum_{t=1}^{T} b(t)$$

(13)

To let $B_s(p)$ indispensable of T in (13), we could let $b(t)$ to be the multiple of $\frac{1}{T}$. For convenience we set:

$$b(t) = \frac{1}{T}$$

(14)

and we get from (13) and (14):

$$B_s(P) = \frac{n(1)}{N-1}$$

(15)

Unfortunately, this could not be regarded as the spreading breadth index of P because the equation (9), based on which (15) is obtained, is not accurate and the procedure from equation (9) to equation (15) is just to determine $a(t)$. Apply (14) to (12) we know:

$$a(t) = \frac{1}{T(D-1)^{t-1}}$$

(16)

Apply (16) to (8) we could get the definition of $B_s(P)$:

$$B_s(P) = \sum_{t=1}^{T} \frac{1}{T(D-1)^{t-1}} \frac{n(t)}{N-1} = \frac{1}{T(N-1)} \sum_{t=1}^{T} \frac{n(t)}{(D-1)^{t-1}} \tag{17}$$

where T is the timespan we choose to get B_s(P), N is the number of vertices in the network, D is the average degree of the vertices in the network and $n(t)(t = 1,2,...,T)$ is the total number of vertices that are being occupied by particles at the time t.

Furthermore, we could get the lower bound of the $B_s(P)$. For a network with N vertices, we have:

$$n(t) \geq 1 \tag{18}$$

if T is selected properly. Apply (18) to (17) and we get:

$$B_s(P) = \frac{1}{T(N-1)} \sum_{t=1}^{T} \frac{n(t)}{(D-1)^{t-1}} \geq \frac{1}{T(N-1)} \sum_{t=1}^{T} \frac{1}{(D-1)^{t-1}} \geq \frac{D}{T(N-1)(D-1)} \tag{19}$$

the last sign of inequality is correct because we eliminate all the terms in the $\sum_{t=1}^{T} \frac{1}{(D-1)^{t-1}}$ if $t \geq 3$.

For upper bound, we notice that

$$n(t) \leq N-1 \tag{20}$$

Apply (20) to (17), we have:

$$B_s(P) = \frac{1}{T(N-1)} \sum_{t=1}^{T} \frac{n(t)}{(D-1)^{t-1}} \leq \frac{1}{T(N-1)} \sum_{t=1}^{T} n(t) \leq \frac{1}{T} \tag{21}$$

Plus, the timespan T must be the same for all the vertices while testing $B_s(P)$ for each vertice. One thing we should pay enough attention to is that if the particle reaches the edge vertice of the network, it will stop moving. Another factor we should consider carefully is that the timespan T must be selected properly. For one thing, it should not be too large because the particle will cover every vertice in the network in the end for each vertice selected at the beginning. For another, if the timespan is too small, the particle will have insufficient time to spread and the evaluation of the spreading breadth index could be unconvincing.

Additionally, we define the half-network period to evaluate the spreading characteristic of the network. Also for a network with n vertices, all the condition is exactly like the submodel explained above. Now we define the time $T_h(P)$ to be the time needed for the particle to have reached half of the N-1 vertices ($\sum_{t=1}^{T_h(P)} n(t) = \frac{N-1}{2}$) in the network. However, starting from some vertices, the particle may never reach half of all the vertices. For example, if the vertice is in a small component of the network, finally it could only reach each vertice in the component, but not the whole network. In this occasion, we could define the half-network period as infinity.

It could be obviously noticed that for each given point P, the spreading breadth index $B_s(P)$ and the half-network period $T_h(P)$ for P have the following relations:

$$B_s(P) = \frac{1}{T_h(P)(N-1)} \sum_{t=1}^{T_h(P)} \frac{n(t)}{(D-1)^{t-1}} \geq \frac{1}{T_h(P)(N-1)} \sum_{t=1}^{T_h(P)} \frac{n(t)}{(D-1)^{T_h(P)-1}}$$
$$\geq \frac{(N-1)/2}{T_h(P)(N-1)(D-1)^{T_h(P)-1}} = \frac{1}{2T_h(P)(D-1)^{T_h(P)-1}} \tag{22}$$

and:

$$B_s(P) = \frac{1}{T_h(P)(N-1)} \sum_{t=1}^{T_h(P)} \frac{n(t)}{(D-1)^{t-1}} \leq \frac{1}{T_h(P)(N-1)} \sum_{t=1}^{T_h(P)} n(t) = \frac{(N-1)/2}{T_h(P)(N-1)} = \frac{1}{2T_h(P)}$$

(23)

So if we know $T_h(P)$ of vertice P, we can restrict the range of $B_s(P)$.

It should be noticed that the two variable $B_s(P)$ and $T_h(P)$ are both used to evaluate the spreading-breadth efficiency of a vertice. If a vertice has a larger spreading breadth index and a smaller half-network period, the spreading ability of the network based on this point will be better.

While considering the practical meaning of the two parameters, it is not hard to notice that they could be used to confirm how wide the information flow could be spread through a particular point, which undoubtedly has a significant meaning to the whole network.

3.3.3 Spreading Depth Index (D_s)

After consideration of the breadth characteristic of the network, it is natural to think up an idea to define some parameter to evaluate the depth characteristic of the network. Assume that we still have use the particle model in the section above, that is we release a particle from P at t=0 and let it move in the network. Similarly, we set a timespan T as the time measurement factor, and define the maximum of the distances(the length of the shortest path) from all the particles at time t to the vertice P as $d(t)$. So we can assume, just like the equation to get the spreading breadth index, the form of the spreading depth index of P as:

$$D_s(P) = \sum_{t=1}^{T} m(t) \frac{d(t) - d(t-1)}{D}$$

(24)

where D is the diameter of the network, that is, the maximum of the distances for any two vertices in the network. In (24), $d(t) - d(t-1)$ is the increase of the distance in the t-th second. $m(t)$ is the attenuation factor for distance. It is easily understood that the spreading distance of the information should have a smaller and smaller weight as time increases to guarantee the time efficiency of the spreading. Make the approximation:

$$d(t) - d(t-1) = 1$$

(25)

and use almost the same way as in the process to get spreading breadth index, we have:

$$m(t) = \frac{1}{T}$$

(26)

And finally we define:

$$D_s(P) = \sum_{t=1}^{T} m(t) \frac{d(t) - d(t-1)}{D} = \frac{1}{TD} \sum_{t=1}^{T} d(t) - d(t-1) = \frac{d(T)}{TD}$$

(27)

as the spreading depth index of vertice P.

For the spreading depth index, we can get the upper bound and the lower bound of it:

$$\frac{1}{TD} \leq D_s(P) = \frac{d(T)}{TD} \leq \frac{D}{TD} = \frac{1}{T}$$

(28)

since $d(T)$ is in the range of $[1, D]$.

The spreading depth index could describe the distance of spreading information in a certain time, which means that a vertice, or a researcher, with a larger index has a more outstanding ability to disseminate information and contributes

more to the whole network. Actually in the definition equation (27), the term $\dfrac{d(T)}{T}$ could be regarded as the spreading velocity, which is independent of the time T, of the information flow.

3.4 DATA PROCESSING

3.4.1 Principal Component Analysis(PCA)

During the process of data analysis, we obtain many indexes. However, it will be too complex to take all these indexes into consideration. In order to reduce the dimension of the indexes, we consider the utilization of Principal Component Analysis (PCA). The basic use of PCA is as a dimension-reducing technique whose results are used in a descriptive manner, but there are many variations on this central theme (see [1]). Because the 'best' two-(or three-) dimensional representation of a dataset in a least squares sense is given by a plot of the first two-(or three-) principal components, the components provide a 'best' low-dimensional graphical display of the data (see [1]). In general, if we want to reduce a n-dimension characteristic matrix to a q-dimension one, PCA operates on the data as the following steps:

a. Calculate the covariance matrix ($Cov_{n \times n}$) of $A_{m \times n}$, where $A_{m \times n}$ documents the characteristic matrix, m documents the size of dataset, and n documents the total number of indexes.
b. Find the eigenvectors and eigenvalues of the covariance matrix. We can assume that $a_1, a_2, \ldots, a_n$ are the eigenvectors and $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues.
c. Choose $a_{i1}, a_{i2}, \ldots, a_{iq}$, which correspond to the q largest eigenvalues, and we can get matrix $C_{n \times q}$.
d. Let$D_{m \times q} = A_{m \times n} \cdot C_{n \times q}$, and $D_{m \times q}$ is the lower dimension of the matrix.

3.4.2 Data Processing Procedures

As mentioning above, we divide the parameters into three groups according to their correlation. We assume that these three groups represent different aspect of the model, and the parameters in the same group have a higher correlation coefficient.

Hence, after the preliminary-data-process, we obtain three characteristic matrices $A_{m \times 3}$, $B_{m \times 3}$ and $C_{m \times 3}$. Matrix A documents the information of centralizing characteristics, matrix B documents the information of connecting characteristics, and matrix C documents the information of spreading characteristics. m is the size of dataset. After normalizing, all parameters are set to the range [0, 1]. Then we apply the above-mentioned PCA procedure to matrix A, B and C respectively to obtain three one-dimension characteristic vector: $Central$, $Connect$ and $Spread$.

Finally, we obtain a weighted sum of the three characteristic vectors. The sum is named influence measurement (IM):

$$IM = \alpha \cdot Central + \beta \cdot Connect + \gamma \cdot Spread \tag{29}$$

IM could represent the importance and influence of each vertice in a network. In general, α, β, γ can be assigned to the same weight.

4. RESULT AND ANALYSIS

4.1 Task1

4.1.1 Building the co-author network of the Erdos1 authors

To build the co-author network from the file Erdos1, we first eliminate the lines to indirect coauthors of Erdos, that is to say, the lines whose one endpoint is not the direct coauthor of Erdos. Without links to Erdos, there are some isolated authors left. Obviously, these isolated authors can't be of vital importance. Therefore, we eliminate these isolated authors from our dataset.

4.1.2 Analyzing the properties of this network

After extracting the data of co-author network, we import the data into Gephi and get some information from it, showing in TABLE 1.

TABLE 1. Summary statistics for co-author network

	Value
Nodes	474
Edges	1640
Average Degree	6.920
Connected Components	5
Density	0.015
Average Clustering Coefficient	0.282
Modularity	0.493
Number of Triangles	1828
Diameter (Longest Path Length)	10
Average Path Length	3.825

There are 474 vertices and 1640 edges in this network, in which each vertice represents an author and each edge means the cooperation between two authors. The value of the average degree of all the vertices means that in average, each author collaborates with 6.920 authors. The 4th line shows that there are 5 distinctive connected components, between which the authors do not cooperate with each other. The modularity of a partition is a scalar value between −1 and 1 that measures the density of edges inside communities compared with the density between communities. The last line documents the average length of shortest path. The meaning of clustering coefficient have been mentioned above.

Now we have obtained some global property of this model. In order to have a more direct recognition, we visualize some properties in FIGURE 1.
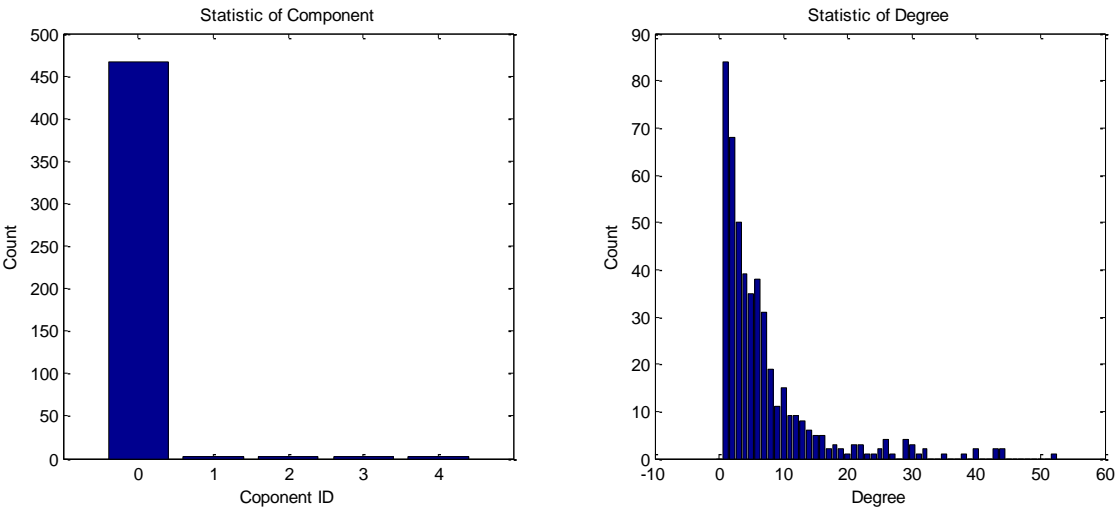

Figure 1. Component and degree distribution

FIGURE 1 shows the distribution of components and degrees. We can easily perceive that almost all vertices are in one component, which means this network has a strong extent of relationship and deserves to be analyzed. The second picture shows the distribution of the degree, which is one of the most important factors in determining the influence of authors. Through this picture, we can know that most vertices have few degrees, which means they have weaker relations with each other.

In order to have a clearer recognition, we draw a figure to show the features of the network. See Figure 2.
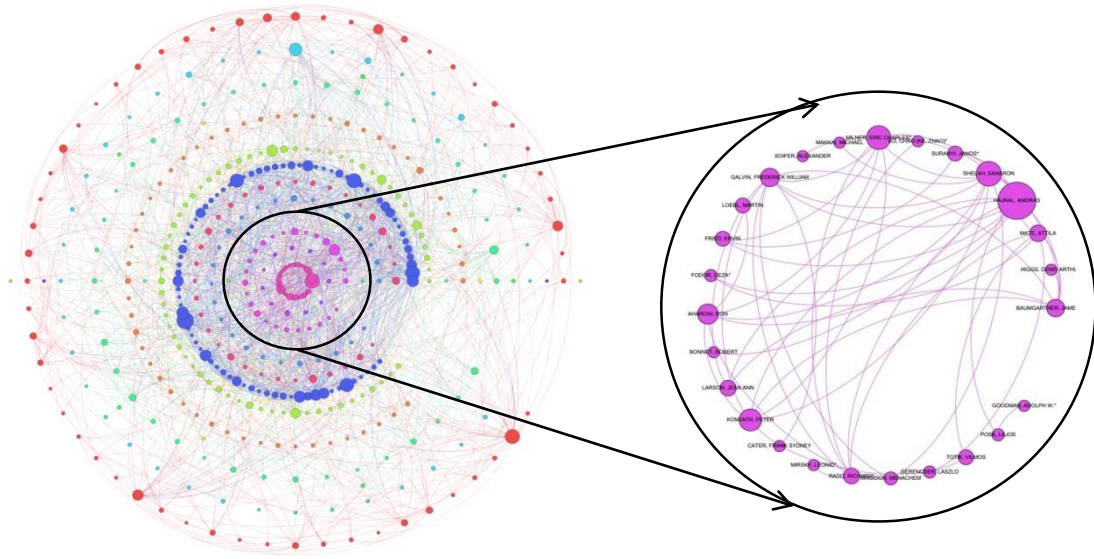
Figure 2.The co-author network

In Figure 2, there are many circles with different colors. We classify these vertices based on their modularity. The size of the vertice is in positive correlation with its degree and the edges between vertices mean that they have cooperation relationship with each other. Each vertice has its own information such as name.

TABLE 2 to TABLE 5 show the top 30 authors based on closeness centrality, betweenness centrality, degree centrality, and PageRank.

TABLE 2. Top 30 authors based on closeness centrality

Rank	Name	Rank	Name	Rank	Name	Rank	Name	Rank	Name
1	HENRIKSEN, M	7	HERZOG, F	13	HUNT, G	19	FELLER, W	25	SEIDEL, W
2	GILLMAN, L	8	BONAR, D	14	SIRAO, T	20	JACKSON, S	26	BLEICHER,M
3	BOES, D	9	CARROLL, F	15	BAGEMIHL, F	21	VIJAYAN, K	27	BABU, G
4	GAAL,S	10	DARLING, D	16	KHARE, SP	22	HWANG,J	28	KUNEN, K
5	SCHERK, P	11	VAN, ER	17	SMITH, B	23	BOVEY, J	29	BUCK, R
6	HERZOG, F	12	WINTNER, AF	18	DARST, R	24	BUKOR, J	30	VOLKMANN,B

TABLE 3. Top 30 authors based on betweenness centrality

Rank	Name	Rank	Name	Rank	Name	Rank	Name	Rank	Name
1	HARARY, F	7	ALON, N	13	RUZSA, I	19	RODL, V	25	LOVASZ, L
2	SOS, V	8	GRAHAM, RL	14	SARKOZY, A	20	SHIELDS, AL	26	ROGERS, CA
3	RUBEL, LA	9	BOLLOBAS, B	15	ODLYZKO,AM	21	TURAN, P	27	FAUDREE, RJ
4	STRAUS, EG	10	PACH, J	16	KLEITMAN, D.	22	SZEKELY,L	28	SHELAH, S
5	POMERANCE, C	11	HAJNAL, A	17	SPENCER, J H	23	CHUNG, FRK	29	WORMALD,NC
6	FUREDI, Z	12	TUZA, Z	18	SCHINZEL,AB	24	SZEKERES, G	30	NESETRIL,J

TABLE 4. Top 30 authors based on Degree centrality

Rank	Name	Rank	Name	Rank	Name	Rank	Name	Rank	Name

1	ALON,N	7	TUZA, Z	13	HAJNAL,A	19	SZEMEREDI,E	25	LUCZAK, T
2	HARARY,F	8	SOS,VT	14	LOVASZ,L	20	CHARTRAND,GT	26	KOSTOCHKA,A
3	GRAHAM,RL	9	SPENCER,JH	15	FAUDREE,RJ	21	STRAUS,EG	27	WEST,DB
4	BOLLOBAS,B	10	PACH,J	16	POMERANCE,CB	22	SARKOZY,A	28	SIMONOVITS,M
5	RODL, V	11	GYARFAS,A	17	KLEITMAN,D	23	BABAI,L	29	RUZSA,I
6	FUREDI,Z	12	CHUNG,FRK	18	NESETRIL,J	24	SCHELP,R	30	WORMALD,NC

TABLE 5. Top 30 authors based on PageRank

Rank	Name	Rank	Name	Rank	Name	Rank	Name	Rank	Name
1	HARARY,F	7	TUZA, Z	13	STRAUS,EG	19	NESETRIL,J	25	ODLYZKO, AM
2	ALON,N	8	POMERANCE,CB	14	CHUNG,FRK	20	FAUDREE,RJ	26	LUCZAK, T
3	GRAHAM,RL	9	FUREDI,Z	15	SARKOZY,A	21	SZEMEREDI,E	27	SCHELP,R
4	BOLLOBAS,B	10	SPENCER,JH	16	KLEITMAN,D	22	CHARTRAND,GT	28	WEST,DB
5	SOS,VT	11	HAJNAL,A	17	GYARFAS,A	23	BABAI,L	29	KOSTOCHKA,A
6	RODL, V	12	PACH,J	18	LOVASZ,L	24	RUZSA,I	30	SHELAH,S

A few authors are highly ranked in all the tables. And these authors can be regarded as one of the most influential authors in the network.

4.2 Task2

In this part, we start to use our data to analyze and determine most important authors. In our model, we have already classified the characteristic parameters into three types (Centralizing Characteristics, Connecting Characteristics, and Spreading Characteristics). We try to extract their main feature with PCA and provide a more comprehensive estimate of the authors by adding a weight factor to each vector. Some variables and their meanings are listed here for convenience：

TABLE 6. Variable and their meaning

Variable	Meaning
IM	Impact measurement
Central	The metric for Centralizing Characteristics
Connect	The metric for Connecting Characteristics
Spread	The metric for Spreading Characteristics
α, β, γ	Weighting coefficient

We assume that Central, Connect, Spread are the vectors extracted from their characteristic matrix with PCA. As for IM, it is defined as:

$$IM = \alpha \cdot Central + \beta \cdot Connect + \gamma \cdot Spread \tag{30}$$

In this passage, we let $\alpha = \beta = \gamma = \frac{1}{3}$. Through comparing IM, we can get their comprehensive rank and compare their influence. TABLE 7 shows the top 10 authors according to each characteristic respectively and provides the list of the top 20 authors by taking all the IM into consideration.

TABLE 7. Top 10 authors based on central, connect and spread

	Central		Connect		Spread		Final IM		Final IM
Rank	*Name*	*Rank*	*Name*	*Rank*	*Name*	*Rank*	*Name*	*Rank*	*Name*
1	**ALON,N**	1	BECK, I	1	FUREDI,Z	1	**ALON,N**	11	SZEMEREDI,E
2	GRAHAM,RL	2	BEJLEGAARD,N	2	**ALON,N**	2	RODL, V	12	FAUDREE,RJ
3	RODL, V	3	BERGER,MA	3	GRAHAM,RL	3	GRAHAM,RL	13	LOVASZ,L
4	BOLLOBAS,B	4	BLECKSMITH,RF	4	BOLLOBAS,B	4	BOLLOBAS,B	14	PACH,J
5	HARARY,F	5	BOALS,AJ	5	SOS,VT'	5	FUREDI,Z	15	CHUNG,FRK

6	FUREDI,Z	6	BONAR,DD	6	TUZA, Z	6	TUZA, Z	16	NESETRIL,J
7	TUZA, Z	7	BUCK, RC	7	STRAUS,EG	7	HARARY,F	17	SIMONOVITS, M
8	SOS,VT'	8	BUKOR, J	8	RODL, V	8	SPENCER,JH	18	KOSTOCHKA,A
9	SPENCER,JH	9	CARROLL,FW	9	LOVASZ,L	9	GYARFAS,A	19	SCHELP,R
10	PACH,J	10	CATES,ML	10	SPENCER,JH	10	SOS,VT	20	HAJNAL,A

From TABLE 7, we obtain the final comprehensive rank of these authors. ALON, NOGA M. is the most influential author among them. RODL, VOJTECH ranks 2nd and GRAHAM, RONALD LEWIS ranks 3rd. In fact, we can find their name appear frequently from TABLE 2 to TABLE 5. In some extent, this fact proves the rationality of our model. Figure 7 provides the comparison between Top3 authors. We can see differences do exist among these three characteristics. The nodes with high centralizing characteristics and spreading characteristics can have low connecting characteristics, and vice versa, which to some extent validates our hypothesis of the three classification.
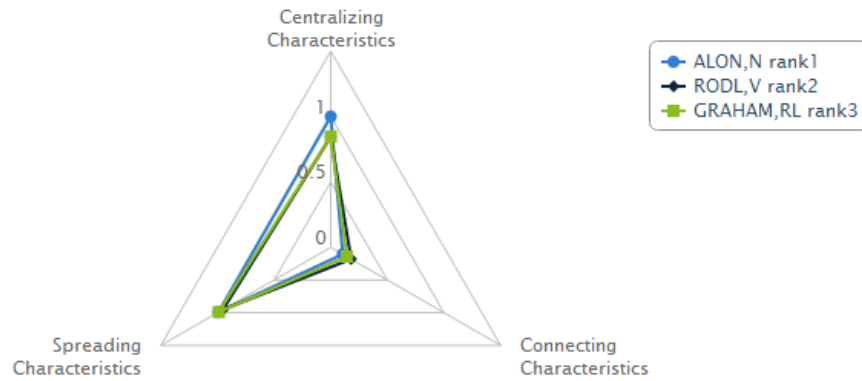

Figure 7. Comparison of top three author in the final rank

4.3 Task3

4.3.1 Choosing a dataset
We use the sixteen papers given to establish the model of network and evaluate the influence and the importance of each vertice. Actually the sixteen papers are selected carefully to guarantee sufficient citation-relations between vertices, or papers, to test the rationality of the model.

4.3.2 Developing the co-author (citation) model
We select the sixteen papers as the vertices of the network to establish our citation model. For each paper, we assign a number from one to sixteen to it for convenience (shown in TABLE 8). In the citation network, for instance, if the paper A cited the paper B, then we add a directed line segment pointing from B to A to indicate this citation relation, as well as the information flow direction. If we assign number m to A and number n to B ($m, n \in \{1, 2 \dots 16\}$), then in the citation matrix C (16*16), C(n, m) will be set to one correspondingly. FIGURE 3 shows the structure of our model.

TABLE 8. Articles' Number

Number	Article Name	Number	Article Name
1	On Random Graphs.	9	Scientific collaboration networks: II.
2	Statistical mechanics of complex networks.	10	The structure of scientific collaboration networks.
3	Power and Centrality: A family of measures.	11	The structure and function of complex networks.

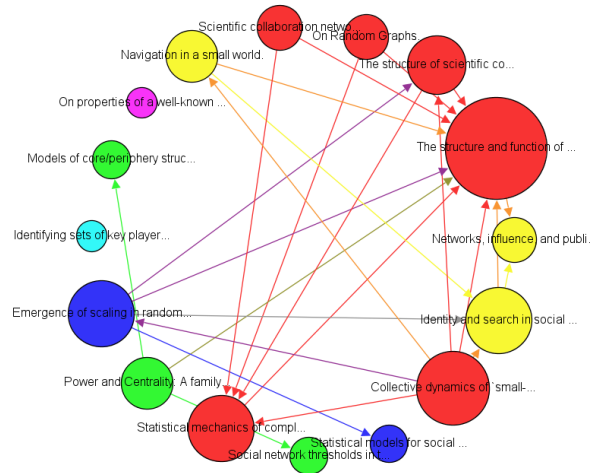| 4 | Emergence of scaling in random networks. | 12 | Networks, influence, and public opinion formation. |
5	Identifying sets of key players in a network.	13	Identity and search in social networks.
6	Models of core/periphery structures. Social Networks.	14	Collective dynamics of `small-world' networks.
7	On properties of a well-known graph, or, What is your Ramsey number?	15	Statistical models for social networks.
8	Navigation in a small world.	16	Social network thresholds in the diffusion of innovations.

Figure 3. The citation network of 16 papers

4.3.3 Paper influence measurement

TABLE 9. Articles' rank based on central, connect and spread

Central		Connect		Spread		IM	
Rank	*No.*	*Rank*	*No.*	*Rank*	*No.*	*Rank*	*No.*
1	14	1	1	1	12	1	12
2	4	2	8	2	2	2	8
3	11	3	9	3	15	3	14
4	3	4	12	4	11	4	10
5	13	5	10	5	13	5	2
6	10	6	14	6	10	6	4
7	8	7	13	7	4	7	13
8	2	8	2	8	6	8	11
9	1	9	4	9	8	9	1
10	9	10	11	10	16	10	9
11	12	11	3	11	1	11	15
12	6	12	5	12	3	12	6
13	15	13	6	13	5	13	16
14	16	14	7	14	7	14	3
15	5	15	15	15	9	15	5
16	7	16	16	16	14	16	7

In TABLE 9, we list the rank of the sixteen papers according to their centralizing, connecting, spreading and comprehensive contributions to the whole network. So we could find that the paper with number 12 (*Networks, influence, and public opinion formation.*) is the most influential one in the network science.

In FIGURE 3, we demonstrate the citation relation of the sixteen papers. Each circle stands for a paper and the size of the circle is in positive correlation with the degree of the vertice. The directed line segment, as mentioned above, illustrate the citation relation. For instance, if a segment points from 4 to 11, it means paper 11 cites the paper 4. One interesting thing we observe from the figure is that the circle 11, with the largest size, has a low rank (8) in the last column of TABLE 9, which means paper 11 has a low IM(influential measurement). However, circle 12, with a quite small size, has the highest rank. How could such phenomenon happen?
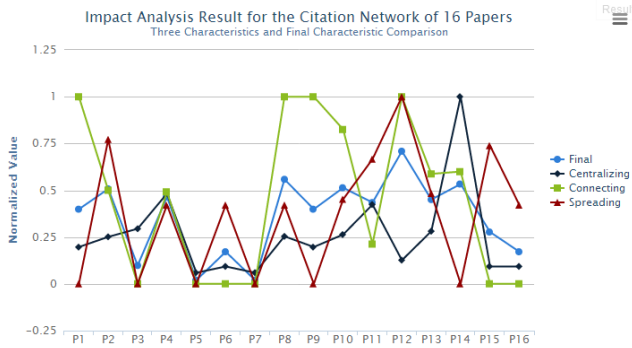

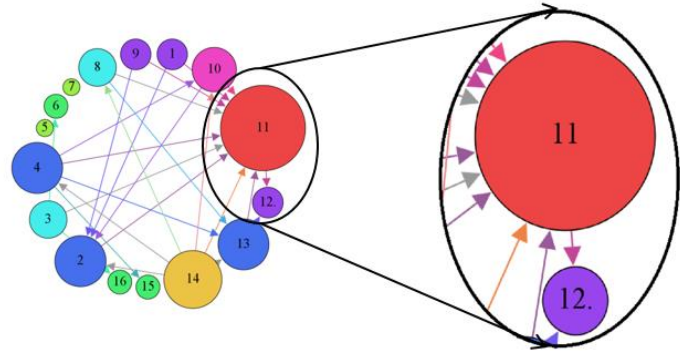Figure 4. Influence analysis result of 16 papers

Figure 5. A detail of citation network

For convenience, we use FIGURE 4, which demonstrates the normalized result of the characteristics of the sixteen papers, to show the superficial paradox. Paper 11, with a relatively high spreading and centralizing contribution, has a quite terrible connecting characteristic at the same time. So in the final assessment result after the weighting process, paper 11 is not as important as we expect at first. Then let's consider paper 12. It has a low centralizing characteristic, but its other two parameters are quite high, which makes paper 12 the most influential ones among the sixteen. If we further observe FIGURE 5, we could find from the detailed drawing that the paper 11 could only be connected to other papers through paper 12. In other words, the information of paper 11 must be with the help of paper 12 to spread in the network. So the paper 12 is actually quite vital for the whole network.

The instance above indicates that we could not simply evaluate the influence of a vertice by its degree and other direct features. Some potential significance must be considered to assess the impact of the vertice more comprehensively.

4.3.4 Utilities in other areas

For an individual researcher in the field of network, we can also apply this model to establish the whole system. Conspicuously, each vertice of the model stands for a researcher. As for the edges in the network, it means that two researchers have cooperative relation, which can also mean they have programs completed together, besides cooperative papers. Next we could use some weight coefficients to work on the cooperative factors mentioned above and get the final weight of each edge. No problem here for the structure of the model, and we can then implement the process in our model to work out the influence and the importance of each vertice. The last question is the data we need. Actually, to implement the model, for any two researchers in the network, we need at least the number of cooperative papers, the number and scale of the cooperative projects and the time, as well as the frequency, of co-citation of the two people's research result.

For a department in a university, we can use this model to estimate the influence and role of each department. Naturally, we use vertices to represent departments in a given universities. Similarly, just as the standard we set for the individual researcher network mentioned above, we will match tow vertices if the two departments have cooperative programs or co-completed papers by the researchers from the two departments. Next we use some weight coefficients to get the final weight of each edge. The final network is expected to like:
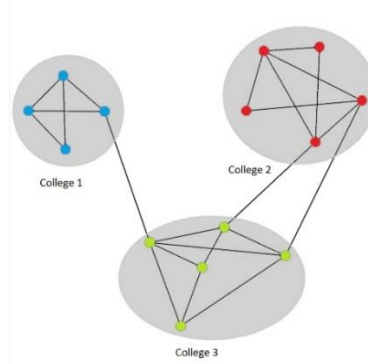
Figure 6 The department network

The vertices(departments) are separated into several communities(colleges), and there are more edges within a community than between two communities in comparison, which could be easily understood since the department within a college should naturally be more correlated. The remaining problem is the data. We can use all the data discussed above in the researcher model since there are a great many professors and researchers in a department and each vertice is actually a large set of professors and researchers. Furthermore, we could select some other data, such as the time of sport activities or public benefit activities organized by the two departments, to make the model more rational and convincing. After all, the academic level is not the only factor to be considered while appraising the role and the influence of a department within a university.

4.4 Task4

4.4.1 Data extraction and network establishment

To further test our model, we extract a relatively larger dataset from a website. Precisely, we first select a hot list of movies in China on the internet (*http://movie.douban.com/top250*), then parse its HTML and crawl the list of movie subject ID. Next we enter the comment page of a particular subject (The Shawshank Redemption for instance) to parsing the HTML to obtain the raw data of user id and user comment score of users who have commented this movie.

Then let us consider how to establish the network model. We regard each user (we have 971 users in total) as a vertice in the network, and for every two users, we calculate the weight of the edge between them. The integral part of the weight is the number of movies both of them have commented, and the decimal part of the weight is a value between 0 and 1 to indicate the similarity of their comments for this movie. If the weight is nonzero, we draw an edge between the two users and assign the calculated value to the edge as its weight. Finally, we get 259179 edges in the network in total.

4.4.2 User influential analysis

We implement our algorithm on a set of the user of Douban Movie. In order to analyze the activity on this site, we try to find the most "influential" user on the site. We try to find the connection of the users by seeing whether they comment on the same movie. In this network, we find 971 users and regard every user as a node. And if they comment on the same movie, there will be an edge between them. We implement our algorithm on this set, and get TABLE 10.

TABLE 10. Top10 of the user of Douban based on central, connecting, spreading and final characteristics

Central		Connect		Spread		IM	
Rank	*No.*	*Rank*	*No.*	*Rank*	*No.*	*Rank*	*No.*
1	856	1	329	1	9	1	9
2	713	2	636	2	11	2	47
3	620	3	504	3	4	3	10
4	736	4	913	4	20	4	65
5	163	5	836	5	8	5	11
6	967	6	509	6	47	6	21

7	107	7	760	7	28	7	29
8	352	8	141	8	16	8	75
9	306	9	399	9	14	9	32
10	654	10	37	10	65	10	28

We can see that their final rank differs in many ways from the first two ranks. However, it has some similarity with the rank of spreading characteristic. It means that spreading characteristic influence the most in this network. Meanwhile, we plot the distribution of 971 Douban users as follow:
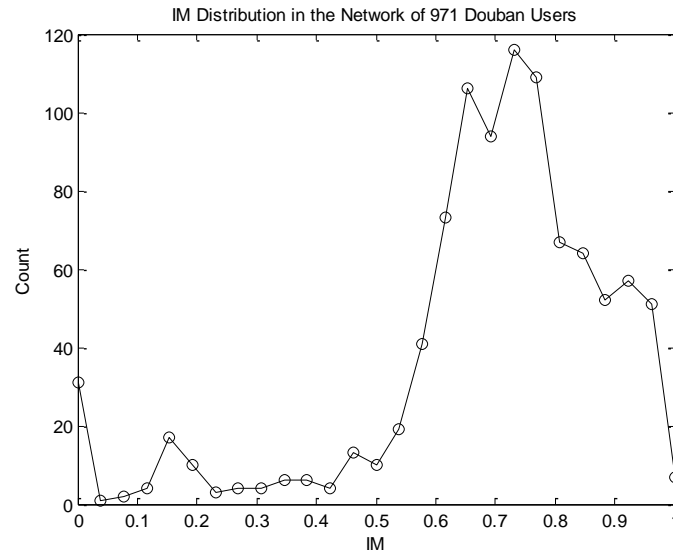

Figure 8. Impact Measurement distribution of 971 Douban users

In the figure, we set 0.05 as step length and calculate the number of the user in this interval. Then we get this figure. We can see that this figure have some similarities with Gaussian distribution. To some extent, the curve fit the fact.

In this task, we can analyze the activeness of users. For many website, it is very crucial to pay attention to this group and consider their advice more carefully. Therefore, it is meaningful to analyze influential" users for them.

4.5 Task5

4.5.1 Understanding and science of modeling influence and impact within networks

To model the network is not a simple task since there is no principle to quantify and simplify either the co-author network or the paper network. Precisely, we must create some self-defined things and combine them together to describe the whole network, despite the fact we also add some classical models to our final framework.

For a network, there already exist some well-defined parameters to indicate its characteristics. However, we additionally define several new parameters ourselves to further reveal the structure of the network in consideration of the fact that there is some distinctive characteristics of the network system to be modelled. Furthermore, we also use the thought of classification in this model. Precisely, we divide the several characteristics into three groups in order to process them separately and finally synthesize the three results together to get a comprehensive evaluation of each vertice in the network system.

4.5.2 Utilities of modeling influence and impact within networks

The main utility of our model is to evaluate the importance of each vertice in a given network. In fact, according to the type and the specific characteristics of the network given, the model could be modified by adding or eliminating some parameters, which could help reveal the property of the network more accurately, to make it more suitable to the question.

Practically, this model could be applied to not only individuals, but also organizations and even nations to make some estimate of the importance. For instance, in the whole science system nowadays, how could we estimate the importance of each discipline? If we regard each vertice in the network as one discipline and two vertices are matched by a line if there are some intersectional fields between the two disciplines. Plus, we can regulate that the weight of the lines stands for the interdisciplinary extent of two disciplines, which can be obtained, for instance, by investigate the number of academic papers related to intersectional field. In this way, the final result of the modeling process could help appraise the significance of each discipline nowadays. Another example of use of this model is to evaluate the influence of the university all over the world, which could be further used as one factor for university ranking. Vertices of the model could be regarded as the universities and the line between two universities mean that the two universities have cooperative programs or projects. The weights of the lines stand for the extent the two universities cooperate. Then we could use our model to evaluate the influence of each universities. However, the result we obtain could only indicate one aspect of the universities, which actually reveals the impact and interrelation between each other among the whole network of universities. Honestly speaking, our model is not feasible for all situations due to the exquisite classification. However, our main idea fits for most of similar models. If we want to take more factors into consideration, we should make a new classification to get a better performance.

5. DISCUSSION AND CONCLUSION

5.1 Strengths
Comprehensive Evaluation

In our model, we set three types of characteristic parameter, with three ones in each type, to attain a panorama of the network structure. The comprehensive evaluation from different angles enables us to analyze and appraise the network with a more tenable and convincing statistical foundation.

Self-defined Evaluation Standards

Among the nine parameters in our model, two of them are completely defined by ourselves. Despite the fact that in the academia, there still exists many standards to delineate the outline of a network model, our self-defined standards, which are designed according to the distinctive peculiarities of our model, can give us a more objective and convincing assessment of the importance of the vertices in the network.

Classification of Characteristics

We hold the opinion that of the nine description parameters, some are correlated and should be considered together, which triggers the idea of the three-type-classification-method. After the process of PCA, every three parameters in a type could be simplified to one parameter. Finally, with the three parameters obtained from the three big groups, we are able to appraise the importance of the vertices in the network from three aspects.

Rational Visualization

To assess our model more exquisitely, we visualize the result data and transform them into many graphs, which could help use have a more intuitive perception of the evaluation result. One advantage of this operation is that we could roughly estimate the rationality of the model conveniently through direct observation of the graphs.

5.2 Weakness and Sensitivity
Dynamic Features

Like many other static models, our model lacks a consideration of the time factor. Precisely, the structure of the network may change with the time and this will lead to a change of the result of the model. It is conspicuous that in the real world, the influence of a researcher, or a paper, could never remain unchanged and this naturally give us a new angle to modify our model.

Limited amount of data

To further test our model, we use some data obtained from a movie website. Totally we consider almost one thousand people to verify the rationality of our model. However, like many other models of network, our model is designed to deal with large network and data set, and the scale of the data tested could be larger to better test the advantages and disadvantages of our model.

Imprecise evaluation of weight

In task 2, we simply let $\alpha=\beta=\gamma=1/3$. However we only know that $\alpha+\beta+\gamma=1$. Our imprecise evaluation may lead to an inaccurate result. In reality, we should have a dataset to train the value of weight constantly. Then the model will have a more convincing and accurate result.

5.3　Contribution

Most of the previous research focused on single factor to analyze the impact, which is insufficient for complex network. To thoroughly and comprehensively describe the features of a network, we systematically construct a three-dimension network impact analysis model, combining nine characteristic parameters, existing ones and self-defined ones, together and dividing them into three groups. Based on the hypothesis that the parameters in the same group are relatively highly correlated, we use PCA to simplify them and get three characteristic parameters totally. We use three weight factors to operate on the three values and obtain a final evaluation value for each node. We collect some raw data and use three datasets to test our model and verify the reasonability of it. Besides, we discuss and further extends the utilities of our model to many other areas of society. At the end of the paper, we objectively assess our model and analyze the advantages as well as disadvantages of it.

6.　ACKNOWLEDGMENTS

7.　REFERENCES

[1] Jolliffe, I. (2005). *Principal component analysis*. John Wiley & Sons, Ltd.

[2] Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, *60*(10), 2107-2118.

[3] Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-world'networks. *nature*, *393*(6684), 440-442.

[4] Latapy, M. (2008). Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical Computer Science*, *407*(1), 458-473.

[5] Bonacich, P. (2007). Some unique properties of eigenvector centrality. *Social Networks*, *29*(4), 555-564.

[6] Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, *60*(10), 2107-2118.

[7] Chen, Y., Hu, A. Q., & Hu, J. (2004). A method for finding the most vital node in communication networks. *High Technology Letters*, 14(1), 21-2.

[8] Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30(1), 107-117.

[9] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.