Identifying the error connections in the network

Abstract： In this paper, we analyzed the structural properties of complex networks and researched the problem of identifying the error connections in six kinds of network. For these networks, we considered their topology and further analyzed some of the specific characteristics.

Firstly, We study them visually by drawing visual figures of network. After analysis, we found that almost all the networks have small world effect, giant branch and their degree distribution skewed to the right. The biological directed network is not obey the power law and its community is divided obviously. Biological undirected network is almost same as directed network except that it obeys the power law and has the assortativity. Information networks' nodes do not have modularity and are very dispersed. Two networks both obey the power law and disassortativity. For social networks ,the directed network obey the power law. The undirected network is almost the same as the directed network. However,it has no giant branch.

Secondly,We found that the biological directed network have similar characteristics to the the food chain and biological undirected networks are similar to biological organs. For both networks, we were using the in-degree and out-degree and common neighbor similarity to identify error connections. The result is the biological directed network's accuracy is 0.364 and undirected network is 0.226. The information directed network is similar to the Internet. We were using in-degree , out-degree and the sorting of PageRank to get the error connections. Two Information network have the same properties.The result is the Information directed network's accuracy is 0.173 and undirected network is 0.309. For the social directed network, we believe that it and twitter have close attention mode. So we assume the existence of the "big V" node and the "active users" node. By analyzing the algorithm about its' topology,we finally come to a result which the accuracy is 0.679 . For the social undirected network, we think that it has the same mode as twitter's friends add mode. We use the same approach to deal with it and the final result is 0.338.

Keywords: visual analysis，degree distribution，common neighbor similarity

CONTENT

Problems restatement

Problems Background

The network is a powerful tool to describe the structure of a real system—— the social network describes the relationship between human beings, and the World Wide Web describes the hyperlink relationship between web pages. With the development of modern technology, we have accumulated more and more network data, but the data is partially incomplete, inaccurate or sometimes distorted. For example, in the biological network, some early proved existing gene-gene and protein-protein interrelations are overturned by new experiments with higher accuracy.

This topic will address real network problems from biology, information and social networks with data of 6 networks. The scale of these networks is ranging from hundreds of nodes to millions of nodes. Each network connection may be undirected (for example, friend-connection in twitter), or directed (such as people "follow" others in twitter). Based on the original real network, we have added a number of false connections which meet following criteria: (1) the number of the false connections is not more than 10% of the total number of connections; (2) the error connections are picked in a completely random manner.

Problems to be solved

(1) Develop a mathematical model to understand the structure and organization mechanics of the network. The structural characteristics of the different types of networks and the organization principle are not always the same.

(2) Propose an effective method to identify the error connections. Show the completeness of how the structural characteristics are discovered; explain the validity and the accuracy of the mathematical model as well as the accuracy of the algorithm.

Issues Analysis

This study is a problem in modern society, with the accumulation of more and more of the network, how do we deal with increasingly large and complex analysis of network data.

One problem requires us to different network architecture model, respectively, analyze their structure and internal mechanisms. First, we analyze the data to arrive at a different network, such as degree distribution, clustering coefficient, connection average geodesic distance of each vertex and so on. With these data, we can analyze the basic nature of the network. We then use these data, it is established random graph model of each network, by analyzing and comparing the model with the original network, to understand the different structure of each network.

Question two asked us to propose an effective way to identify errors in six different network connection, and show the complete structural features from which to discover and explain the accuracy of the validity and accuracy of mathematical models and algorithms. By the first question we already know the structural properties of these network topology, the network respectively organisms, biological undirected networks, information has to networks, information undirected networks, social networks have the social network itself different undirected the structural features of departure, make a reasonable analysis, some of which will certainly remove the correct link, then apply the similarity of link prediction method based on the establishment of a common neighbor similarity index, pick out the wrong link.

Model Assumptions

1. The error does not affect the real link topological properties of each network.
2. Each network has low specificity and most nodes obey some regularity.

Symbol Description

i，j	node
k	Degree
$< \mathrm{k} >$	Average degree
k^{out}，k^{in}	Out degree、in degree
C	Clustering coefficient
A	Adjacency matrix

Modeling and Solution

The first problem of modeling and solving

the biological network model established and solved

Directed Network ：

Directed network, also known as a directed graph in there to the network, each side has a direction, from one node to another node, this side is called a directed edge.

In a adjacency matrix A of the directed network，Element A_{ij} following meanings;

$$A_{ij} = \begin{cases} 1, \text{If the connection from node j to node I} \\ \qquad 0, \text{other} \end{cases}$$

Undirected Networks：

Undirected network is the most simple network, each connection has no direction in the undirection network. In a adjacency matrix A of the undirected network，Element A_{ij} following meanings;

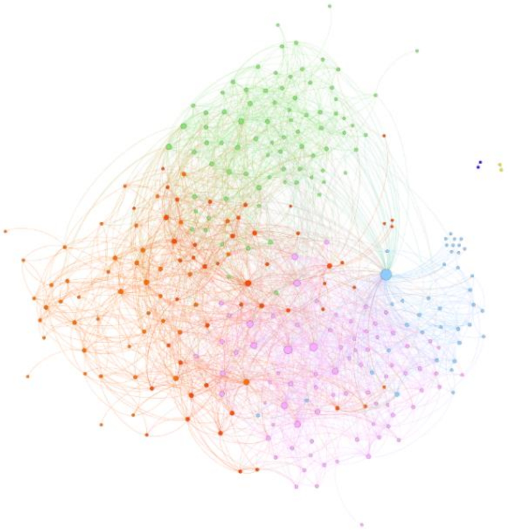
$$\mathrm{A}_{ij} = \begin{cases} 1，\text{If there is an connection between the node i and the node j} \\ 0，\text{If there is no connection between the anode i and the node j} \end{cases}$$

First, we list the data of the biological directed network
Through software gephi obtain the biological directed network the following data

property\ Network Name	BIO-DIRECTED
The number of the nodes	295
The number of the connection	2251
Average Degree	7.631
The largest branch network in the proportion of the total	291/295(98.64%)
Average Path Length	3.589
Diameter	11
Number of shortest paths	66874
Density	0.027
The Number of Community	8
Modularity	0.376
Weak connection assembly	1
Strongly connected assembly	49
Average Clustering Metric	0.143

The table of the data of the biological directed network

We then use the software Gephi to make the Performance Chart of network visualization



the Performance Chart of network visualization of the biological directed network

The nodesof the graph in different colors according to community division, and node size is adjusted according to the degrees. Intuitively drawn from the graph, the biological directed network of community divided obvious.

In addition, we analyze the data and images to observe nature and get several other networks

Small World effect

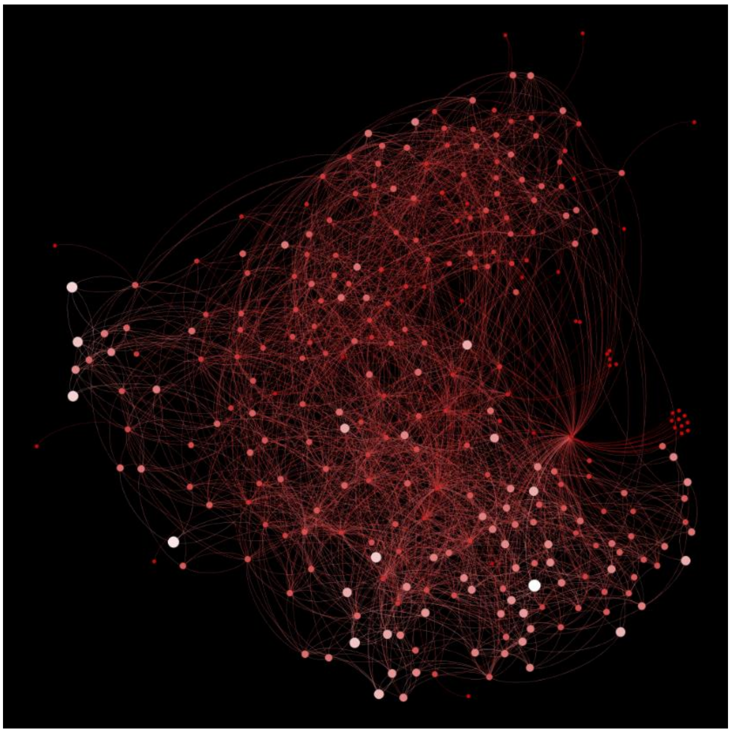
If the average distance between any two points(L) in the network with the increasing number of network nodes(N) in a logarithmic growth,which is L~l n N. And it still has obvious characteristics of the group on the local structure of the network, the network is said to have a small-world effect

Clustering coefficient

Clustering coefficient is a measure of two neighboring nodes also mutually average probability neighbor, in fact, this parameter measures the density of the network in a triangular structure. Considering a given distribution network that the nodes are random, then the clustering coefficient can be expressed as follows.

$$\mathrm{C} = \frac{1}{n}\frac{【< k^2 > -< k > 】^2}{< k >^3}$$

Since the connection of each nodes of the geodesic distance of 3.589 on average, that is in the network and the average map distance between any node is 3.589, while the average clustering coefficient of 0.143, is not very high, but according to the size of each node clustering coefficient intuitive analysis image can be seen the following cases



The chart of the biological directed network according to the clustering coefficient distribution

Local node clustering coefficient is approximately obey community profile.

The importance of the node

The importance of a node by PageRank value to measure. Mathematically speaking, PageRank is defined formula

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$$

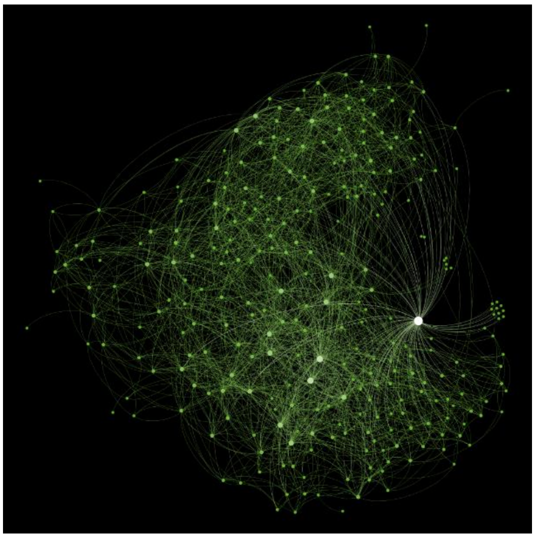
In the form of matrix, the above equation can be expressed as the following in the form of

$$\mathrm{x} = \alpha \mathrm{AD}^{-1} x + \beta 1$$

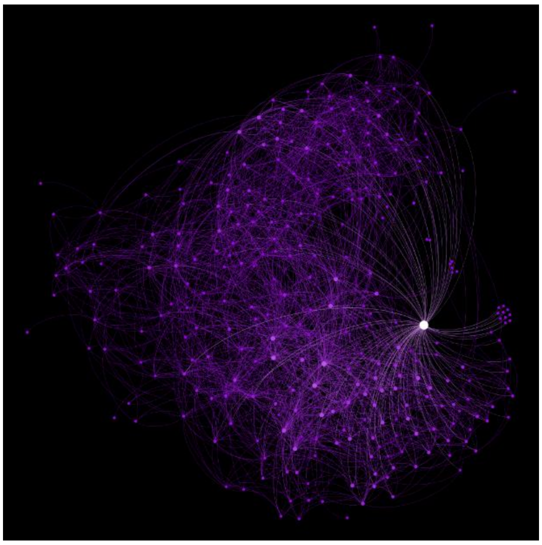
among them, 1 represents a vector（1,1,1，…）,D is a diagonal matrix. Its elements is D=max$\left(k_i^{out}，1\right)$.If $\beta = 1$，sorting out can get the following equation:

$$\mathrm{x} = (\mathrm{I} - \alpha \mathrm{AD}^{-1})^{-1} 1 = D(D - \alpha A)^{-1} 1$$

Calculated and drawn by gephi PageRank distribution nodes and node-based distribution center of the feature vector based.



Node-based PAGERANK distribution.



Node degree distribution of feature vectors based center

From the figure, the figure most important point is relatively not high, but the importance of the existence of a relatively high node 44, its PageRank value is 0.113, eigenvector centrality is 1.

The Distribution of the degree of the network:

Degree is one of the simplest and most important concepts characterize the properties of a single node. Undirected network node is defined as the number of edges and nodes directly connected, with an average degree of all nodes in the network is called the average degree of the network that marked $< \mathrm{k} >$

Given network G adjacency matrix $\mathrm{A} = \left(a_{ij}\right)_{N*N}$

$$k_i = \sum_{j=1}^{N} a_{ij} = \sum_{j=1}^{N} a_{ij}$$

$$< \mathrm{k} >= \frac{1}{N} * \sum_{i=1}^{N} k_i = \frac{1}{N} \sum_{i,\ j=1}^{N} a_{ij}$$

The following relationship between network nodes and network the number of connection M:

$$2\mathrm{M} = \mathrm{N} < \mathrm{K} \geq \sum_{i,\ j=1}^{N} k_i = \sum_{i,\ j=1}^{N} a_{ij}$$

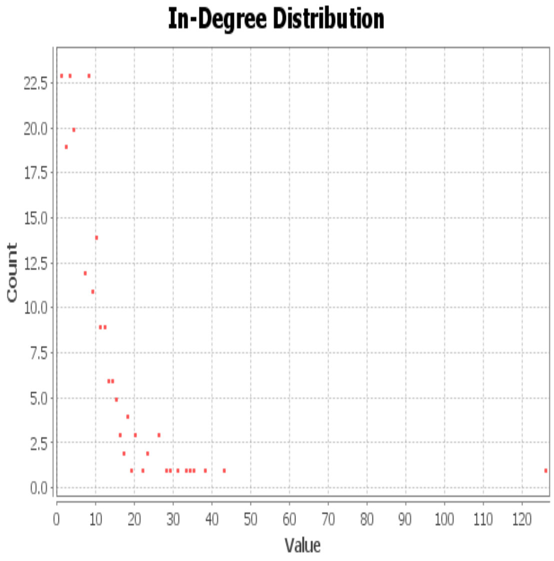
$$\mathrm{M} = 0.5\mathrm{N} < \mathrm{K} >= 0.5 \sum_{i=1}^{N} a_{ij}$$

$$< \mathrm{k} >= \frac{2M}{N}$$

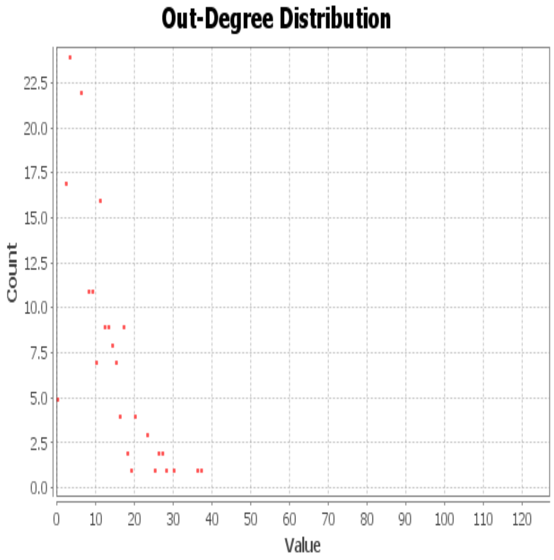
There are degrees of nodes in the network to include out-degree and in-degree。Out-degress is the number that based on the degree of a node as a starting point, starting at the connection of the node that marked k_i^{out}. In-degree of a node is the starting point, the number of arcs terminating edge of the node that marked k_i^{in}.
The in-degree and the out-degree of a node adjacency matrix elements can also be represented:

$$k_i^{in} = \sum_{j=1}^{N} a_{ij}\ ,\ \ k_i^{out} = \sum_{j=1}^{N} a_{ji}$$

Because the biological network is the directed network, and thus we should analyz the in-degress and the out-degree.
By calculating gephi software available



The chart of the in-degree distribution of the biological directed network



The chart of the out-degree distribution of the biological directed network

From figure, the network degree distribution is statistically right side

Figure redrawn using the logarithmic scale(The data of the two axes are logarithmic. While also increasing the column width so that the effect is more apparent), if the distribution roughly follow a straight line, then the degree distribution follows a power-law distribution.

According to mathematical terms, a linear function of the relationship between the logarithm of the degree distribution is true.

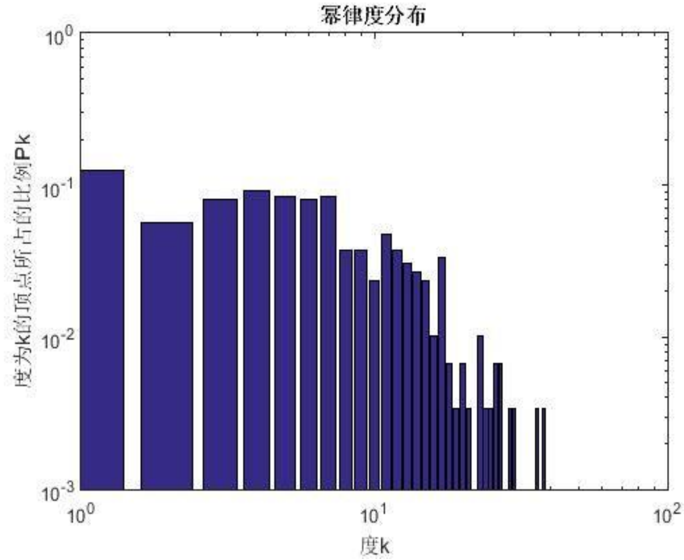
$$\ln p_k = -\alpha \ln k + c$$

A and c both are constants.Here's minus non-mandatory, it can be omitted, but because of the slope of the figure is negative, a negative number would be very convenient if the value is positive, then the figure will be the opposite slope.

Both sides of the equation doing exponentiation, logarithmic relationship can be denoted as follows:

$$p_k = Ck^{-\alpha}$$

Among then, $C = e^c$ is a constant. This form of distribution that changes according to the power of k, called a power law.
Use MATLAB to redraw biological directed network are plans to network distribution curve



the power law degree distribution of the biological directed network

According to empirical chart, you can see the degree distribution of the network do not obey the power law.

Assortativity

The definition of the correlation function is as follows:

$$< jk > - < j > < k > = \sum_{j,k} jk(e_{jk} - q_j q_k)$$

In general, a large-scale network in accordance with the absolute value of the calculated values obtained are also large, but can be normalized to eliminate this effect, allowing the same level with and compare different sizes with different networks. When the network is fully equipped with, $e_{jk} = q_k \delta_{jk}$ is maximum that alse called the remainder of the distribution of the variance:

$$\sigma_q^2 = \sum_k k^2 q_k^2 - \left[\sum_k k q_k\right]^2$$

Thereby obtaining a normalized correlation coefficient, which is equipped with the following factors:

$$\mathrm{r} = \frac{1}{\sigma_q^2} \sum_{j,k} jk(e_{jk} - q_j q_k)$$

obciously, $\mathrm{r} \in [-1,1]$,if $\mathrm{r} > 0$, then the network is **assortativity.If** $\mathrm{r} < 0$,the network is un**assortativity.By MATLAB calculating,** the **assortativity of the** biological directed network is -0.0919. The distribution network is un**assortativity**, but the level is not strong.

Community Properties:

Modularity is a measure commonly used standard quality divided society in recent years, the basic idea is to divide the community after the network with the corresponding null model are compared to measure the quality of the community division.The so-called zero modle corresponding to a network model, refers to the network have the same properties in some ways(such as the same number of connection or the same distribution, etc.) while in other areas completely random random graph model. For a given real network, it is assumed to find a community divided, then the sum of the number of connection in the interior of all communities be calculated as follows:

$$Q_{real} = \frac{1}{2}\sum_{ij} a_{ij}\delta\ （C_i，\ C_j）$$

$A = (a_{ij})$ is the adjacency matrix of the practical network. C_i and C_j represent nodes i and nodes j in the community of the network: If the two nodes belong to the same community,that $\delta = 1$;or $\delta = 0$.

For a same size zero model corresponding to the actual network, if the community is divided by the same way, then the number of connection within communities as the sum of the expected value

$$Q_{null} = \frac{1}{2}p_{ij}\delta\ （C_i，\ C_j）$$

$p_{i\square}$ is the expected value of the number of connections in the zero model nodes i to nodes j.

A module of the network on the definition of the difference between the number of sides the number of sides for the internal network of associations and societies internal accounts corresponding to zero model of the entire network M ratio of the number of sides,

$$Q = \frac{1}{2M}\sum_{ij}\left(a_{ij} - \frac{k_i k_j}{2M}\right)\delta(C_i，\ C_j) = \frac{1}{2M}\sum_{ij} b_{ij}\delta(C_i, C_j)$$

$$b_{ij} = a_{ij} - \frac{k_i k_j}{2M}$$

$B =\ （b_{ij}）_{N*N}$ also called modules Matrix.

By the software gephi calculated,the modularity of the biological directed network is 0.376,the number of the community is 8, the quality of the community divisied is good.

The giant branch

As the size of the nodes of a network is proportional to the growth of branches, called the branch of the giant branch network.There is often a particularly large piece of communication in the network, which contains a considerable proportion of the entire network nodes.

$$\mathrm{u} = （1 - \mathrm{p} + \mathrm{pu}）^{n-1} = 【1 - \frac{c}{n-1}（1-u）】^{n-1}$$

when $\mathrm{n} \to \propto, \mathrm{u} = e^{-c（1-u）}$.

Since u is not part of the giant branch node proportion, it belongs to the branch node huge proportion of $\mathrm{S} = 1 - \mathrm{u}$, Using s eliminate u,we can obtain:

$$\mathrm{S} = 1 - e^{-cS}$$

The formula for any given degree of mean c, when the network size tends to infinity, the huge size of the branch network to increase the proportion. By calculating gephi software, bio-directional network, the ratio of the giant branch accounted for 98.64 percent, which means that the vast majority of the nodes in the network are communicated with each other.

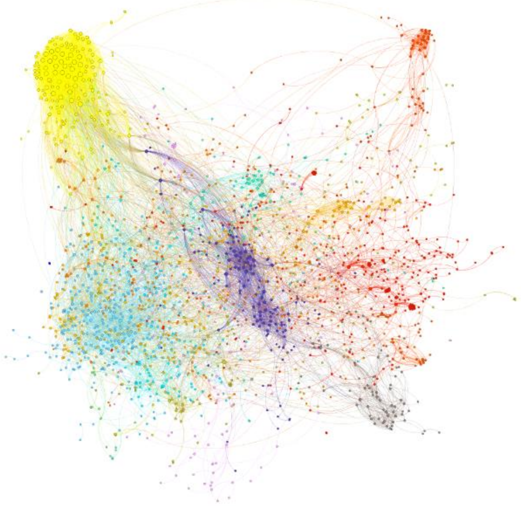
the biological undirected network model established and solved

By software gephi obtain the biological undirected network the following data

The table of the data of the biological undirected network

property\ Network Name	BIO-UDIRECTED
The number of the nodes	2188
The number of the connection	10492
Average Degree	9.59
The largest branch network in the proportion of the total	2180/2188(99.63%)
Average Path Length	4.807
Assortativity	0.4862
Diameter	15
Number of shortest paths	4750228
Density	0.002
The Number of Community	18
Modularity	0.691
Weak connection assembly	1
Average Clustering Metric	0.319
totle triangle	53751

We then use the software Gephi to make the Performance Chart of network visualization

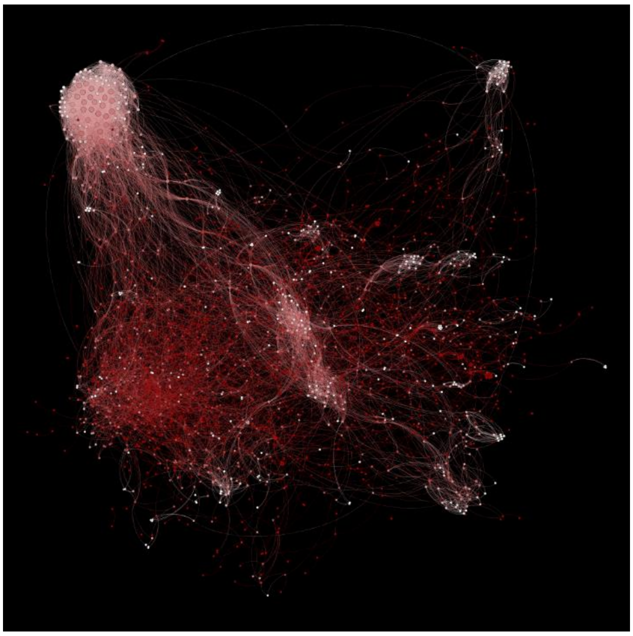


the Performance Chart of network visualization
of the biological directed network

Intuitively drawn from the figure, biological undirected network is obvious community divided.Several other properties of network analysis:

Small World effect

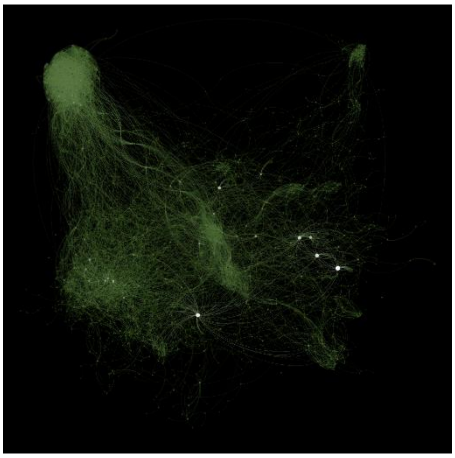
Connecting each node to the average geodesic distance is 4.807, while the average clustering coefficient is 0.319.They are not really high.According to the size of each node clustering coefficient intuitive analysis image can be seen the following cases.



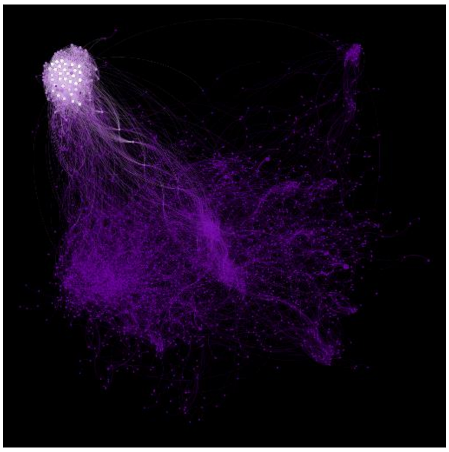
The chart of the biological undirected network according to the clustering coefficient distribution

Local node clustering coefficient is approximately obey community profile

The importance of the node



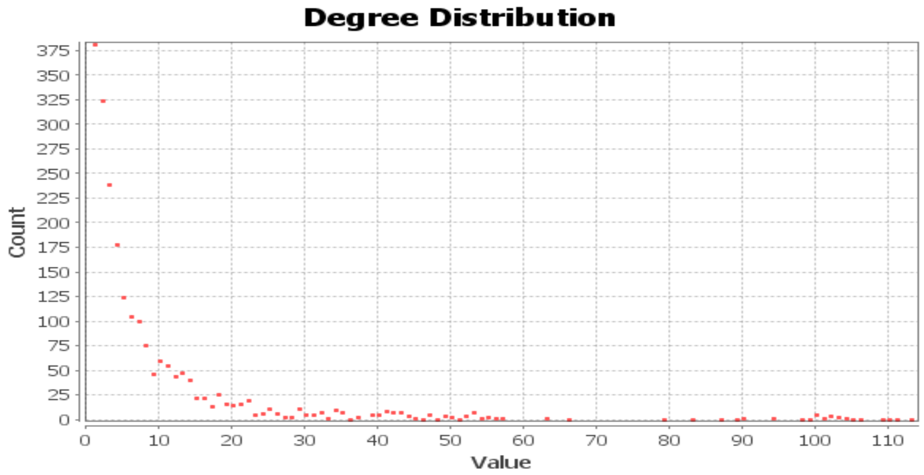
Node-based PAGERANK distribution



Node degree distribution of feature vectors based center

The importance of availability from the figure, based PAGERANK sorted, the figure most points is relatively not high, but there are some relatively high importance of nodes. Feature vector based distribution center, you can see there is an area in the upper left graph, the points are tight, high eigenvector centrality.

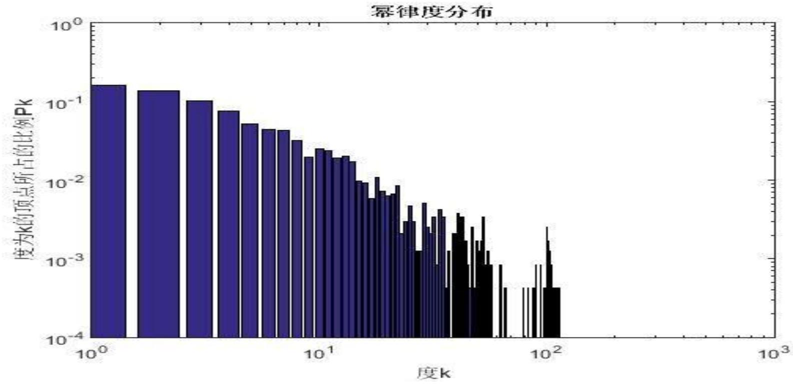
The Distribution of the degree of the network:



The chart of the degree distribution of the biological undirected network

From figure, the network degree distribution is statistically right side

Draw a power-law distribution



The degree distribution of network approximately obey a power law distribution.

Assortativity

By MATLAB calculating, the assortativity of the biological undirected network is 0.4862. The distribution network is assortativity, and the level is strong.

Community Properties:

By the software gephi calculated,the modularity of the biological undirected network is 0.691,the number of the community is 18, the quality of the community divisied is good.

The giant branch

In bio-undirectional network, the ratio of the giant branch accounted for 99.63 percent, which means that the vast majority of the nodes in the network are communicated with each other.

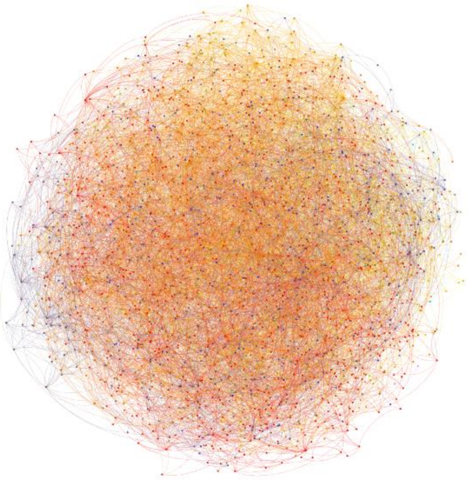
The information network model established and solved

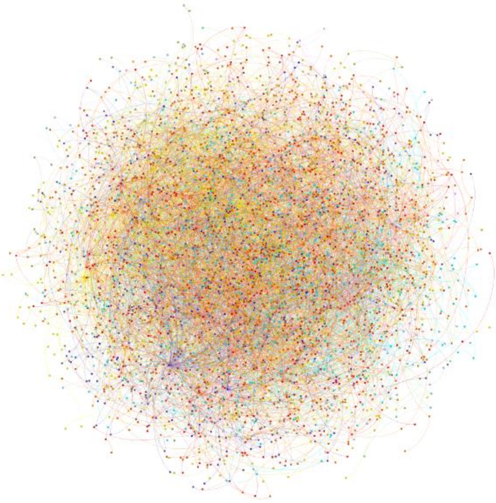
By software gephi obtain the information network the following data

The table of the data of the information network

property\ Network Name	INFO-DIRECTED	INFO-UNDIRECTED
The number of the nodes	2593	4556
The number of the connection	9091	5789
Average Degree	3.506	2.541
The largest branch network in the proportion of the total	2540/2593(97.96%)	4473/4556（98.18%）
Average Path Length	8.547	6.323
Assortativity	-0.0978	-0.1286
Diameter	26	16
Number of shortest paths	1040110	20003388
Density	0.001	0.001
The Number of　Community	40	75
Modularity	0.594	0.873
Weak connection assembly	24	36
strong connection assembly	2303	
Average Clustering Metric	0.072	0.03
totle triangle		696

We then use the software Gephi to make the Performance Chart of network visualization



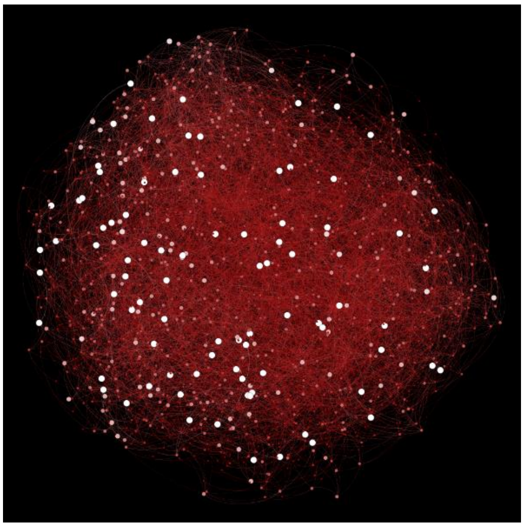


The information directed network　　　　the information undirected network

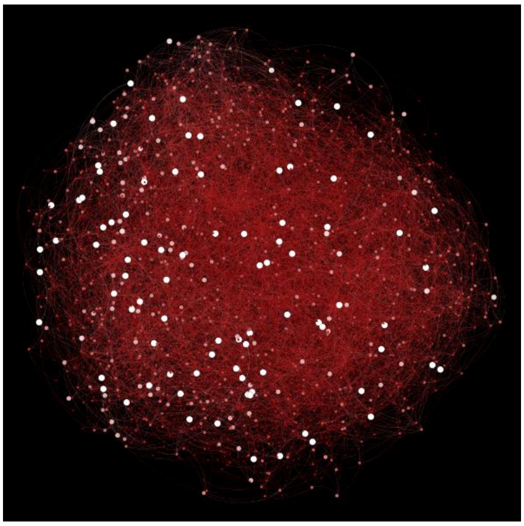
Intuitively drawn from the figure, the image was like an explosion, and uniform color distribution, dividing the two communities not obvious.
Several other properties of network analysis:

Small World effect

Connecting each node to the average geodesic distance are 3.506 and 2.541, while the average clustering coefficient are 0.072 and 0.03.They are not really high.According to the size of each node clustering coefficient intuitive analysis image can be seen the following cases.



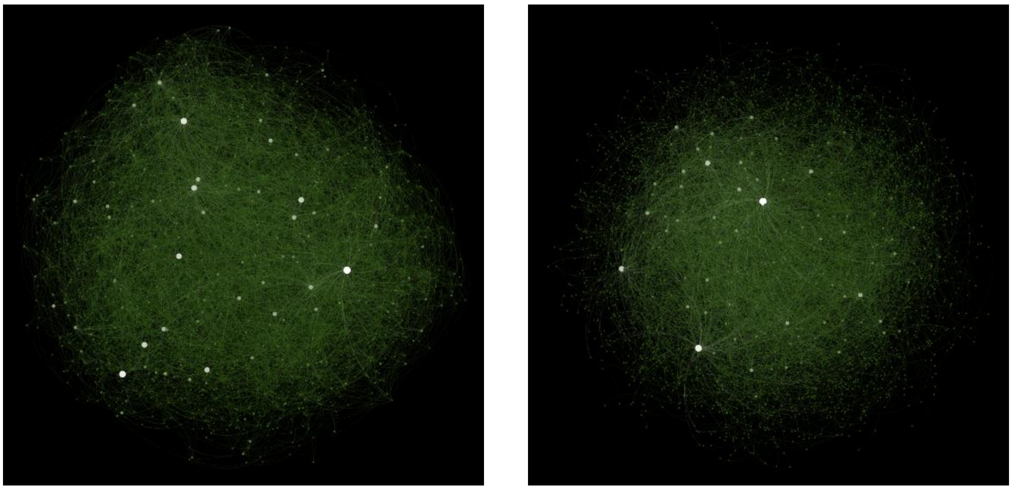
The chart of the information directed network according to the clustering coefficient distribution



The chart of the information undirected network according to the clustering coefficient distribution

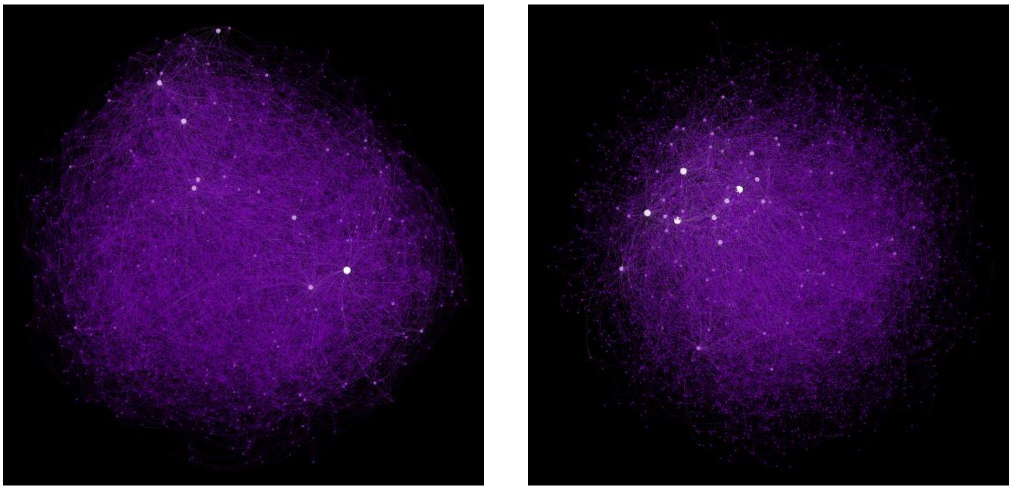
Their nodes'local clustering coefficient are more dispersed, uniform

The importance of the node



The information directed network the information undirected network

Node-based PAGERANK distribution

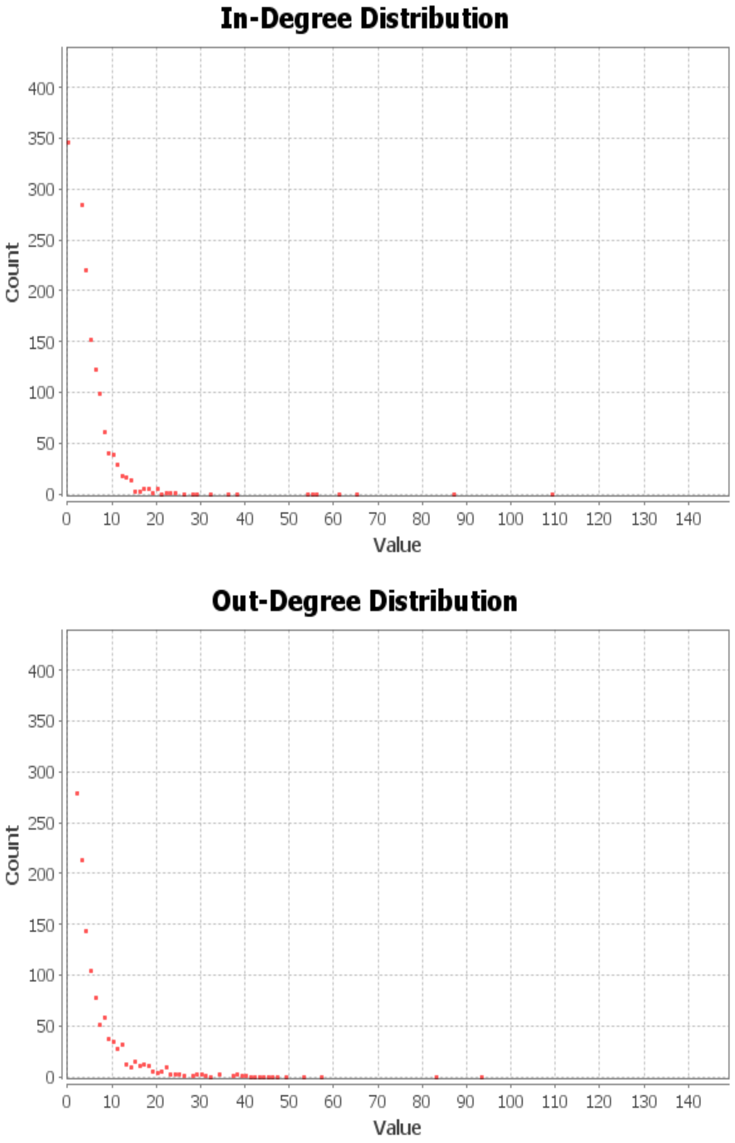


The information directed network the information undirected network

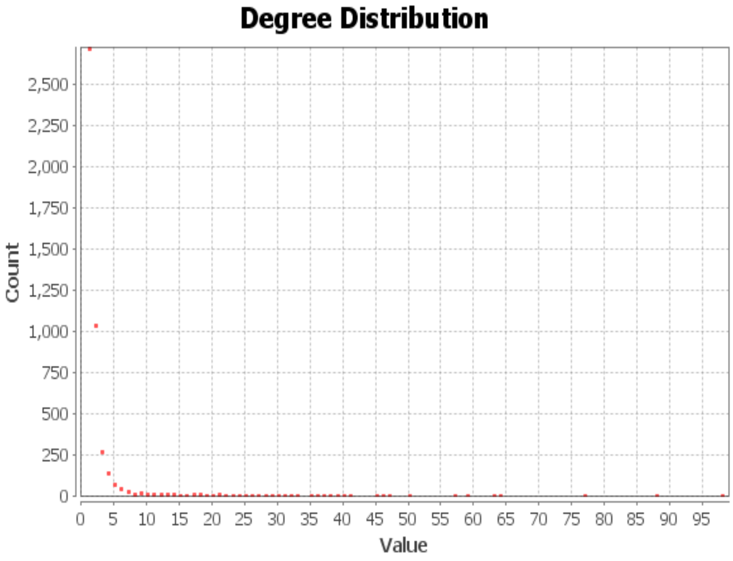
Node degree distribution of feature vectors based center

Obtained from the figure, based on the sort of PAGERANK distribution and feature vectors center, you can see the importance of the figure is not high relative to most of the points, but there are some relatively high importance uniformly distributed nodes in the network.

The Distribution of the degree of the network:

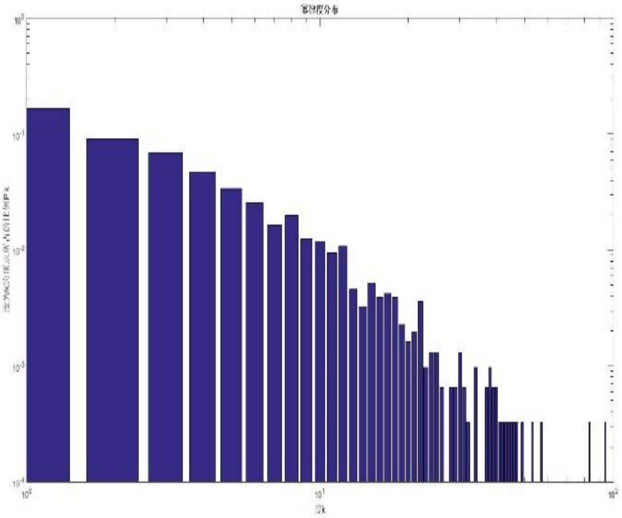


The chart of the degree distribution of the information directed network

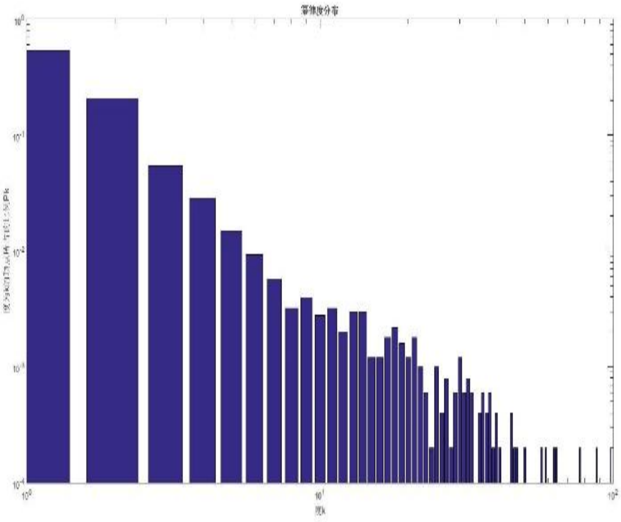


The chart of the degree distribution of the information undirected network

From figure, the network degree distribution is statistically right side. Draw a power-law distribution



The information directed network



the information undirected network

The degree distribution of network approximately obey a power law distribution.

Assortativity

The **assortativity of the** information directed network is -0.0978,and the **assortativity of the** information undirected network is -0.1286. The distribution networks both are un**assortativity**, but the level is weak.

Community Properties:

By the software gephi calculated,the modularity of the information directed network is 0.594,and the modularity of the information undirected network is 0.873, the number of the community are 40 and 75, the quality of the community divisied is good

The giant branch

In informationl network, the ratio of the giant branch accounted for 97.96 percent and 98.18 percent , which means that the vast majority of the nodes in the network are communicated with each other.

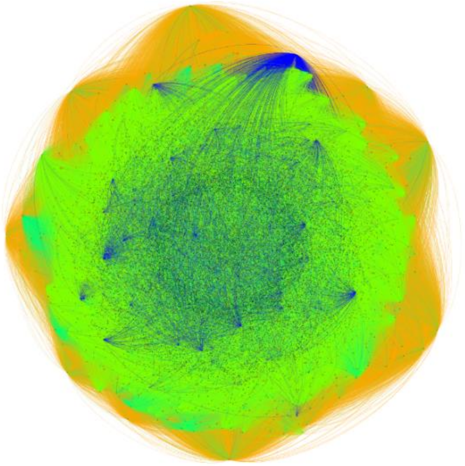
the social network model established and solved

By software gephi obtain the social network the following data

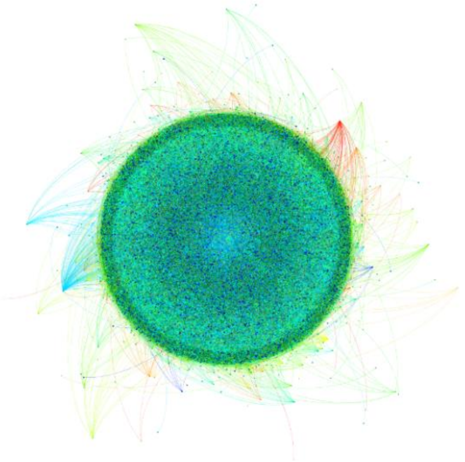
The table of the data of the social network

property\ Network Name	SOCIAL-DIRECTED	SOCIAL-UNDIRECTED
The number of the nodes	25442	50400
The number of the nodes	1506390	44269
Average Degree	59209	1.757
The largest branch network in the proportion of the total	25440/25442(99.99%)	2319/50400（4.6%）
Average Path Length	3.308	11.435
Diameter	10	30
Number of shortest paths	616566726	6713096
Density	0.002	0
The Number of　Community	6	6564
Modularity	0.392	0.999
Weak connection assembly		6556
Average Clustering Metric	0.096	0.051
totle triangle		344

We then use the software Gephi to make the Performance Chart of network visualization



the social directed network

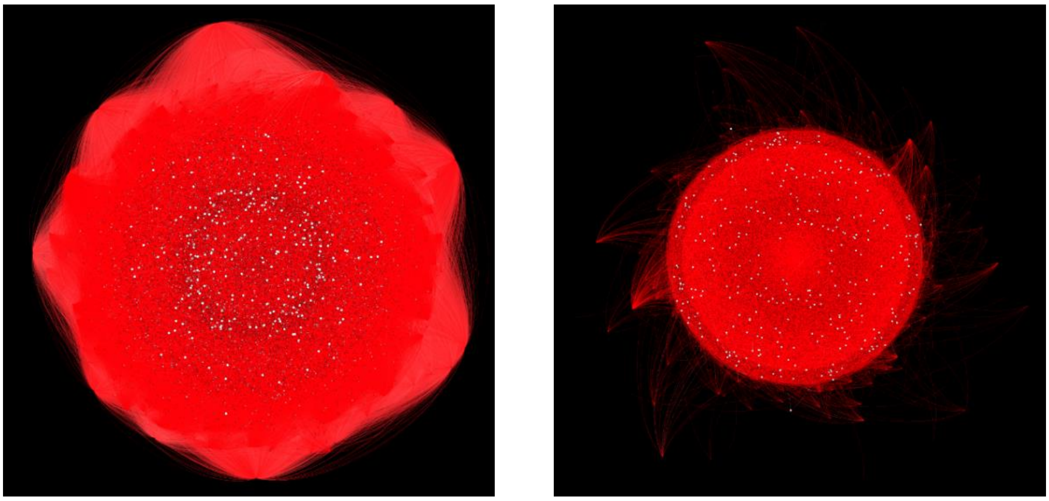


the social undirected network

Intuitively drawn by the map, social network community divide is not obvious, the analysis of several other properties of the network

Small World effect

Connecting each node to the average geodesic distance are 3.308 and 11.425, while the average clustering coefficient are 0.096 and 0.051.They are not really high.According to the size of each node clustering coefficient intuitive analysis image can be seen the following cases.

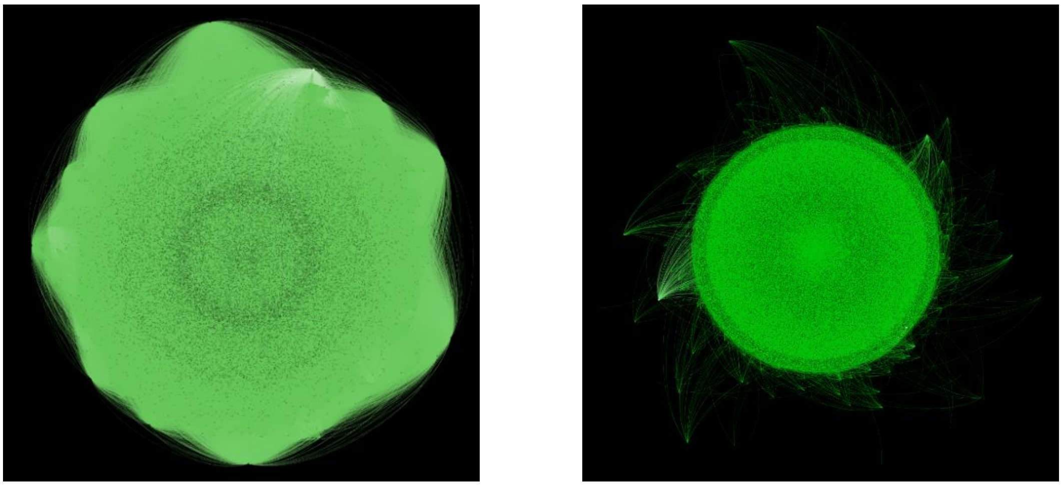


the social directed network the social undirected network

The chart of the social network according to the clustering coefficient distribution

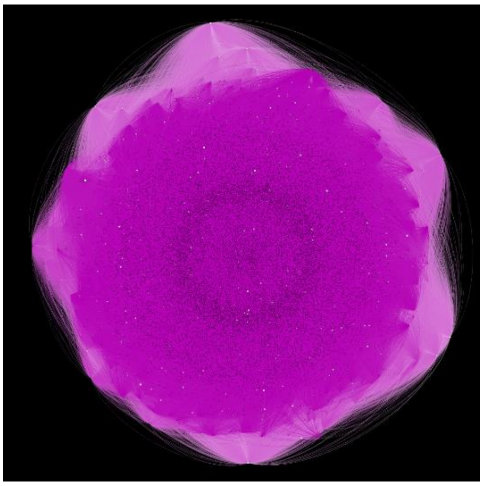
Their nodes'local clustering coefficient are more dispersed, uniform

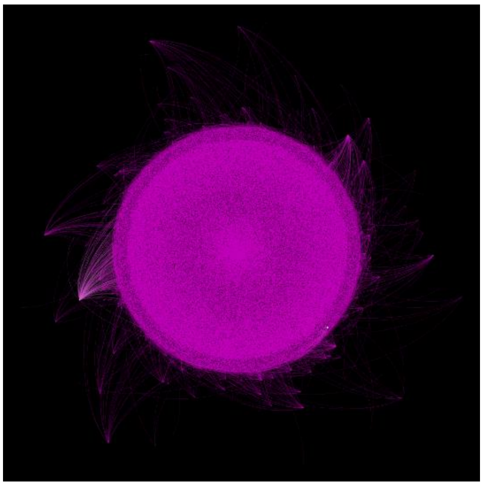
The importance of the node



the social directed network the social undirected network

Node-based PAGERANK distribution



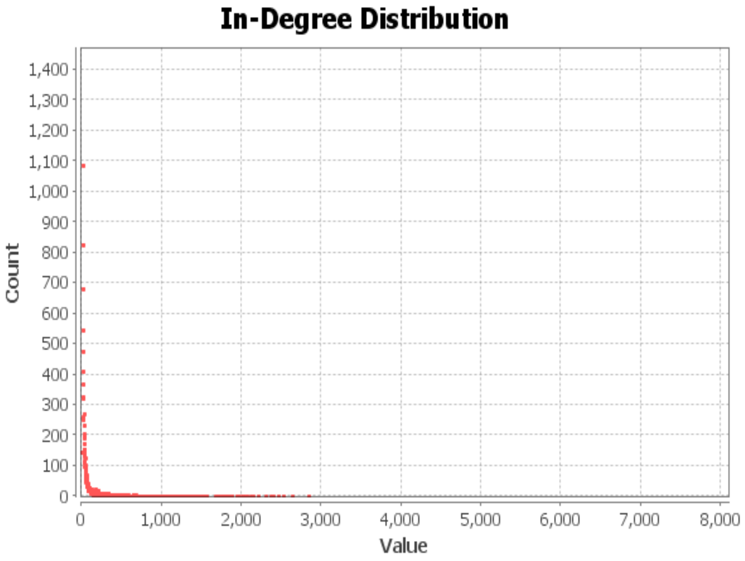


the social directed network the social undirected network

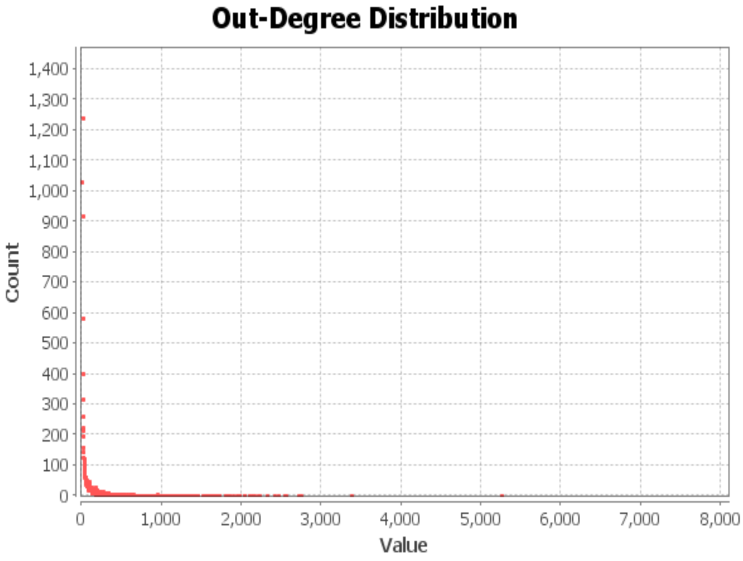
Node degree distribution of feature vectors based center

Obtained from the figure, based on the sort of PAGERANK distribution and feature vectors center, you can see the importance of the figure is not high relative to most of the points, but there are some relatively high importance uniformly distributed nodes in the network.

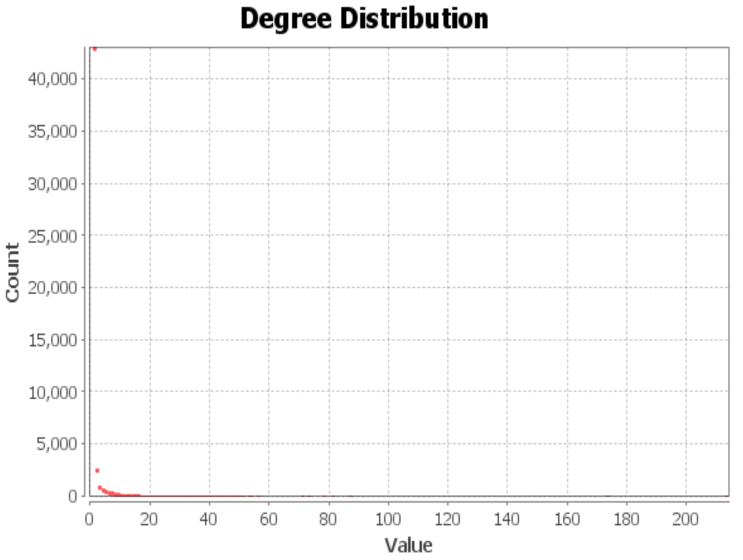
The Distribution of the degree of the network:



The Figure of in-degree distribution of the social directed network



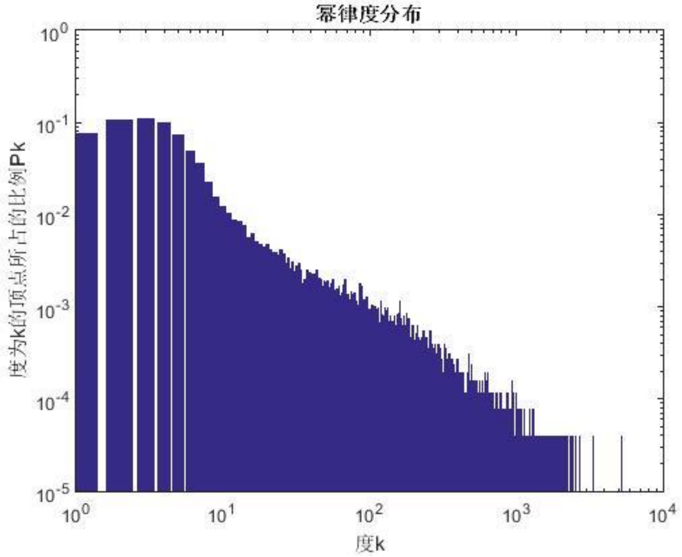
The Figure ofout-degree distribution of the social directed network



The Figure ofout-degree distribution of the social undirected network

From figure, the network degree distribution is statistically right side.

Draw a power-law distribution



The degree distribution of network approximately obey a power law distribution

Assortativity

The assortativity of the social directed network is 0.4862. And the level is strong.

Community Properties:

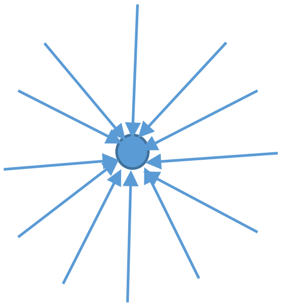
By the software gephi calculated,the modularity of the social directed network is 0.392,and the modularity of the social undirected network is 0.999, the number of the community are 6 and 6564, the quality of the community divisied is good. And the social undirected network has high modularity, divided obvious.

The second problem of modeling and solving

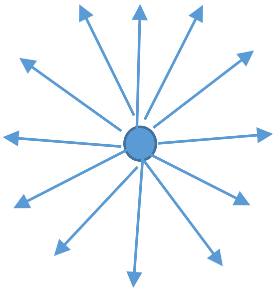
Social directed network analysis:

1. From the data given in the Annex shows, there is a great number of nodes figures, far greater than the number of nodes , in other words,nodes of social directed networks is not continuous. Therefore, we suspect that in the initial state, a large number of nodes’ degree is zero presence in social networks.Due to the data while the node’s degree is zero does not exist, we infer that there is a larger number of broken links in the nodes whose degree is one.
2. We continue to suspect that many in the form TWITTER "Big V" of the existence of a network node. This node status is there are a lot of penetration, and these are connected with the degree of penetration of smaller nodes. This type of connection in a directed social network is an objective reality.Therefore,we

believe that suspicion of the connection between big V and the node whose degree is low is low, it can be excluded. In addition, there is a class of "active users"which have high outdegree and low indegree. Likewise, their suspicion is low, we can be excluded.

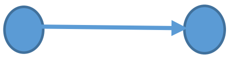


"Big V" node

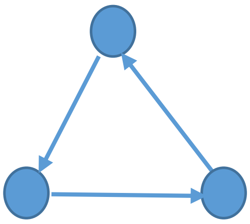


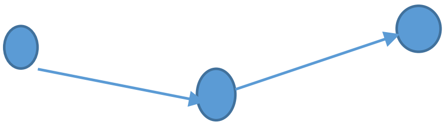
"Active Users" node

3. After the exclusion of such a link, we analyze some of the following special link mode.

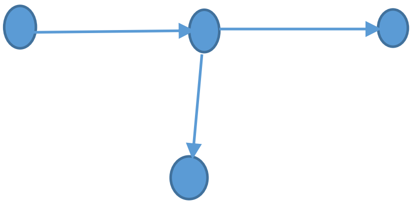


The node which degree is one





The node which degree is two



The node which degree is three

All remaining connections are a combination of more than a few connections. And wherein the first degree for two Linked Mode Because of the common neighbor, a lower degree of suspicion.

Based on the above four properties, we get the following judgment ideas:

1. Make the distribution analysis to calculate the value of the outdegree and the indegree;
2. According to the degree distribution

3. Excluding the node of the hign outdegree and indegree;
4. Place the remaining node by eigenvector centrality sorting will feature a greater degree vector node is also excluded.
5. Repeat the above steps 1,2,3,4 until a 2108 side that believe these sides is wrong link suspect the biggest edge.

The above analysis algorithm can be programmed to give the wrong link and may submit the correct rate of 0.679.

social undirected network analysis:

1. We suspect that social network edge undirected nature is similar to Twitter in the form of mutual add friends. That A sent a request to add B, if B agrees that the two parties to each other become friends. Due to extensive network of friendships, e-pals are difficult to form a large circle of relations, which may also explain why the giant branch does not exist in the network.
2. Similarly, the network also has "active users", particularly the nodes of high degree. Since the link is randomly distributed error for such a small quantity of extremely large point, which produces low probability of error edges, thus the suspection of the edge which is connected with the nodes is low;
3. In the circle of e-pals, there are still real-life circle of friends. Thus, the existence of a common neighbor nodes has lower suspicion.

Based on the above three properties, we still have to judge according to the method of undirected social network.

1. according to the degree distribution,calculate the degree of the nodes;
2. excluding the nodes of the high degree;
3. calculate the common neighbor value of all nodes ,the excluding the nodes which have high common neighbor value;
4. The remaining nodes sort by eigenvector centrality, then excluding the nodes which have high eigenvector centrality;
5. Repeat the above steps 1,2,3,4 until get a number of links to the wrong side to a given number
6. The above analysis algorithm can be programmed to give the wrong link and may submit the correct rate of 0.338.

biological directed network analysis:

1. For directed bio-network, we suspect it has the nature of reality in the food chain; Analyzed by an intuitive graph, according to the degree of all the nodes,they are divided into three categories: the underlying biology, the intermediate layer herbivores class, middle class and the top layer of animal predators.The largest number of underlying biology, which links the most outdegree to the top of the animals and other creatures in this layer, then the indegree of the layer of other organisms, if one of the indegree to emerge into the upper creature was broken

links. Intermediate layer animals were divided into predators and herbivorous animals. Carnivores also point this layer with the underlying biological organisms, and herbivores animals not only point to the bottom point of the layer of biological.The top of the animal only has indegree without outdegree.in other word,it can eat the most animals and

have few predators,such as humanity. If it produces a outdegree, compared with suspected links to dramatically wrong.

2. According to the first question referred, the high modularity of biological network module. Therefore, we suspect, population and level of network module division and related organisms.

ALGORITHM:

1. calculate the outdegree and indegree of nodes;
2. Press the above principles will be divided into three categories nodes;
3. For benthic organisms, pick out a link to the upper organisms;
4. For the middle organisms were selected links to the upper biological and principles are not consistent with its own characteristics;
5. For top Biochemicals, selected out of its links;
6. The above link merge sort that get suspicious of broken links Above ideological use Gephi software.

Biological undirected network analysis:

For undirected bio-network ,because of high modularity, We suspect it has characteristics of biological tissue. At the same kinds of tissues and organs, there is a high degree of individual contact between cells, and between different organs, there is a small number of cells link.

Thus for the network,we use the common neighbor Similarity measure. The common link in all neighbor similarity score sort, that is more likely to score less common neighbor fewer erroneous links. Based on the similarity of common neighbor indicators divided into the following categories:

CN indicators

Based on the similarity of the simplest indicators of local information is the common neighbor similarity index. CN indicators similarity can be called structural equivalence, and two nodes if there are many common neighbor nodes, then both nodes is similar.more similar nodes, the greater the probability of the connection between them.each node obtained by the similarity sorts. Low similar nodes can be approximated that have bigger the probability of error.

Link prediction application CN indicators basic assumption is, if a node has a possibility to many common neighbors, then the link between them is large,.For example, in social networks, if two people do not know each other, but they have a lot of common neighbor, then maybe strangers will be very likely to become friends.

CN indicators defined as follows: for the network node v_x, define their common neighbor set is τ（x）, then two nodes v_x and v_j similarity is defined as the number of their common neighbors,which is

$$S_{xy} = |\tau_{(x)} \cap \tau_{(y)}|$$

Among them ,in the right of equation represent the set. Obviously the number of their common neighbor is equal to the number of paths which length is two between two nodes,

$$s_{xy} = （A^2）_{xy}$$

The base of common neighbor consider the influence about degree distribution of nodes at both ends , from different angles and in different ways to produce the following several similar indicators:

Cosine similarity:

Also known as the Salton Index, which is defined as:

$$s_{xy} = \frac{|\tau（x）\cap \tau（y）|}{\sqrt{k_x k_y}}$$

Jaccard index:

$$s_{xy} = \frac{|\tau（x）\cap \tau（y）|}{|\tau（x）\cap \tau（y）|}$$

Svenson Index:

$$s_{xy} = \frac{2 * |\tau（x）\cap \tau（y）|}{k_x + k_y}$$

Degree nodes favorable indicators:

$$s_{xy} = \frac{|\tau（x）\cap \tau（y）|}{\min\{k_x, k_y\}}$$

Because of the smaller denominator only node determines, then this definition we can see more easily with a high similarity between the conservative node with other nodes.

Degree nodes ignore index:

Its definition is similar to generosity node favorable indicators ,but taking the maximum degree of nodes at both ends,

$$s_{xy} = \frac{|\tau（x）\cap \tau（y）|}{\max\{k_x, k_y\}}$$

LHN-I index:

$$s_{xy} = \frac{|\tau（x）\cap \tau（y）|}{k_x k_y}$$

Denominator $k_{x,}$ k_y is proportional to node v_x and node v_y Common neighbor expectations

$$\mathrm{E}(|\tau（x）\cap \tau（y）|)$$

For this issue, we chose Svenson index.

By the presence of MATLAB programming for all sides to sort and outcome indicators. Correct rate of 0.226

Information directed network analysis

1. For information directed network , we suspect it is similar to the reality of the Internet model. In the Internet, there is a small number and a great amount of core sites, such as GOOGLE, BAIDU other search engines. In this network, there is a similar node, it has high degree, and great importance . For these nodes, the error rate in the side on which the connection in terms of probability is not high. Therefore, we can exclude such nodes.
2. In the information network ,it has low modularity,so Its common neighbor and clustering coefficient of each node can not be used as criteria.
3. Due to the special nature of the information, the independent existence of the information does not exist.We believe that in the network the edges which is connected with nodes that degree is one must be correct;

Based on these three principles, we get the following judgment ideas:

1. calculate the degree of all nodes;
2. excluding the nodes whose degree is one;
3. excluding the nodes the nodes which have high outdegree and indegree;
4. The remaining node are sorted according to PAGERANK;
5. Repeat the above steps 1,2,3,4 until get a number of links to the wrong side to a given number

Submit results to determine the correct rate of 0.173.

Information undirected networks analysis

1. For information undirected network, we can not guess their status in practice, its characteristics were analyzed directly. The network does not have the same module features.
2. The network node is very dispersed, but the degree distribution is skewed to the right, it has a small number nodes which have high degree;

   According to the above principles that we use to the same ideas with the

information directed networks :
1. calculate the degree of all nodes;
2. excluding the nodes whose degree is one;
3. excluding the nodes the nodes which have high outdegree and indegree;
4. The remaining node are sorted according to PAGERANK;
5. Repeat the above steps 1,2,3,4 until get a number of links to the wrong side to a given number
Submit results to determine the correct rate of 0.309.

Model Checking

We split the various networks,according to different modules,then identify broken links.Such as society directed networks , we identify the nodes which have low degree and which have high degree respectively.The result show that ,there is less error ratio in the higher degree node set,in contrast,there is more error ratio in the low degree node set.So, We believe that the model is basically correct.

Model Evaluation

Advantage:

The model of first question: from the perspective of the topological structural properties,such as degree distribution, maximum branch,assortativity and disassortativity, analysis of some common properties of different networks have.
The second question get the nature of the network structure according to the first question,ananlyze different network, draw the appropriate algorithm, program and get results.
Each algorithm is based on the different nature of each network structure, highly relevant, and the results are more accurate.

Disadvantage:

The second question of the universality of the algorithm is not very high, can not be applied to other complex networks, the accuracy of the results needs to be improved.

Model Promotion

The six models in the first question are based on the annux in the topic,have particularity. So we tried to establish random Distribution of an arbitrary network, by fitting the actual Distribution network, simulate in line with the characteristics of each network can be adjusted to the number of nodes of the network. Almost all of the topological structural properties can be shown by this random network of complex networks, analyze the characteristics of different universal networks. But after realization we found its poor performance characteristics of the network effect. This block content needs to be further studied.

In the second question, if we know a sufficient number of samples and correct the error sample, we can use the neural network trained to recognize the error link.

References

Linyuan Lv、Tao Zhou，link prediction，Higher Education Press，2013.

M.E.J.Newman，Networks:An Introduction，Publishing House of Electronics Industry，2014.

Xiaofan Wang、Xiang Li、Guanrong Chen，Network Science：An Introduction，Higher Education Press，2012.

Roger guimer、marta sales-pardo,missing and spurious interactions and the reconstruction of complex networks,

Appendix

```
clear;clc;close
A=load('InfoUD.mat');
P=100;
B=[];
B(:,1)=A.node1;
B(:,2)=A.node2;
if ~all(all(B(:,1:2)));
    B(:,1:2)=B(:,1:2)+1;
end
num=max(max(B));
C=zeros(num);
n=length(B);
for i=1:n
    C(B(i,1),B(i,2))=C(B(i,1),B(i,2))+1;
end
C=C+C';
R=get_degree_correlation(C);
[M,N_DeD,N_predict,DeD,aver_DeD]=Degree_Distribution(C,P);
N_predict=floor(N_predict);
j=sum(N_predict);
D=[];
for k=1:P+1
    D=[D (k-1)*ones(1,N_predict(k))];
end
function [ out ] = get_degree(A,k)
row = A(k,:);
out=size(find(row==1),2);
end
function [M,N_DeD,N_predict,DeD,aver_DeD]=Degree_Distribution(A,P)
N=size(A,2);
DeD=zeros(1,N);
for i=1:N
      DeD(i)=sum(A(i,:));
end
aver_DeD=mean(DeD);

if sum(DeD)==0
    disp(' 该网络只是由一些孤立点组成');
    return;
else
    figure;
    bar([1:N],DeD);
    xlabel('节点编号n');
    ylabel(' ,各节点度数K');
    title('网络中各节点度数大小K的分布图');
end

figure;
M=max(DeD);
predict=0:P;
for i=1:M+1;
    N_DeD(i)=length(find(DeD==i-1));
end
P_DeD=zeros(1,M+1);
P_DeD(:)=N_DeD(:);
bar([0:M],P_DeD,'r');
xlabel('节点的度K');
ylabel('度为K的节点个数');
title('网络中的节点度个数分布图  ');
hold on
N_predict=interp1([0:M],N_DeD,predict,'spline');
plot(predict,N_predict);
hold off
figure;
PK_DeD=zeros(1,M+1);
PK_DeD(:)=N_DeD(:)./sum(N_DeD);
bar([0:M],PK_DeD);
set(gca,'yscale','log','xscale','log');
xlabel('度k');
ylabel('度为k的顶点所占比例');
title('幂律度分布')
function [ r ] = get_degree_correlation( A)
B = triu(A);
M = size(find(B==1),1);
sum1=0;
sum2=0;
sum3=0;
A1 = find(B==1);
length = size(A1,1);
for i=1:length

    [x y]=ind2sub(size(B),A1(i));
    sum1 = sum1+get_degree(A,x)*get_degree(A,y);
    sum2 = sum2+get_degree(A,x)+get_degree(A,y);
    sum3 = sum3+get_degree(A,x)^2+get_degree(A,y)^2;
```

```
end

x1 = sum1/M-(sum2/(2*M))^2;
y1 = sum3/(2*M)-(sum2/(2*M))^2;
r=x1/y1;
end
clear;clc;close
A=load('InfoUD.mat');
P=100;
B=[];
B(:,1)=[A.node1;A.node2];
B(:,2)=[A.node2;A.node1];
load('InfoUD_DeD.mat')
B1=B(:,1);
num0=unique(B1);
mini=min(num0);
maxi=max(num0);

check=mini:maxi;
len=length(check);
i=1;
leak_num=0;
leak=NaN*ones(len);

while i == len
    if num0(i)==check(i)
        i=i+1;

    else
        que_num=num0(i)-check(i);
        std_num=leak_num;
        final_num=que_num+leak_num;
        leak(std_num+1:final_num)=i:i+que_num-1;
        i=i+que_num;

    end

end

B2=B(:,2);
index=1:len;
reform_data=NaN*ones(len,len);
leak_std=1;
```

```
for j=index
    if j==leak(leak_std)
        leak_std=leak_std+1;
        continue;
    else
        judge_sign = (B1 == check(j));
        term=sum(judge_sign);
        reform_data(1:term,j)=B2(judge_sign);
    end

end
L=zeros(len);
S_xy=zeros(len);
AV_DeD=zeros(len);
for i=index
    for j=index
       Lx=reform_data(:,i);
       Ly=reform_data(:,j);
       Lx=Lx(~isnan(Lx));
       Ly=Ly(~isnan(Ly));
       L(i,j)=length((intersect(Lx,Ly)));
       AV_DeD(i,j)=DeD(i)+DeD(j);
       S_xy(i,j)=2*L(i,j)/(DeD(i)+DeD(j));
    end
end
clear;clc;
A=load('S_xy_BU.mat');
UA=load('BioD.mat');
UVA=load('AV_DeD_BioUD.mat');
len1=length(UA.node1);
%C=load('C.mat');
C=zeros(len1,4);
%len1=length(C.C);
D=zeros(len1,4);
C(:,1)=UA.node1;
C(:,2)=UA.node2;
len=length(A.S_xy);
index=1:len;
B=zeros(sum(index),4);
i=1;
k=1;
while i<len+1
     B(k:k+len-i,1)=i*ones(len+1-i,1);
    B(k:k+len-i,2)=i:len;
```

```
        B(k:k+len-i,3)=A.S_xy(i,i:len);
        B(k:k+len-i,4)=UVA.AV_DeD(i,i:len);
        k=k+1+len-i;
        i=i+1;
end
B(:,1:2)=B(:,1:2)-1;
[B1 B2]=find(isnan(B));
B(B1,:)=[];
len2=length(B);
```