Complex Network Properties of Chinese Natural Science Basic Research

Jianguo Liu, * Yanzhong Dang and Zhongtuo Wang

Institute of System Engineering, Dalian University of Technology
2 Ling Gong Rd., Dalian 116023, Liaoning, P. R. China

Abstract

In this paper, we studied the research areas of Chinese natural science basic research from a point view of complex network. Two research areas are considered to be connected if they appear in one fund proposal. The explicit network of such connections using data from 1999 to 2004 is constructed. The analysis of the real data shows that the degree distribution of the research areas network (RAN) may be better fitted by the exponential distribution. It displays small world effect in which randomly chosen pairs of research areas are typically separated by only a short path of intermediate research areas. The average distance of RAN decreases with time, while the average clustering coefficient increases with time, which indicates that the scientific study would like to be integrated together in terms of the studied areas. The relationship between the clustering coefficient $C(k)$ and the degree $k$ indicates that there is no hierarchical organization in RAN.

Key words: Complex networks, power-law distribution, clustering coefficient, evolution network, Chinese Natural Science Basic Research

PACS: 89.75.Da, 89.75.Fb, 89.75.Hc

1 Introduction

In the past few years there has been a growing interest in the study of complex networks. The boom has two reasons: the existence of interesting applications in several biological, sociological, technological and communications systems and the availability of a large amount of real data [1,2,3,4,5,6,7,8,9,10,11]. Recent works on the mathematics of networks have been driven largely by the

* Corresponding author. Tel. (+86)13050537943.
Email address: liujg004@yahoo.com.cn (Jianguo Liu.).
observed properties of actual networks and the studies on network dynamics [12,13,14,15,16,17,18,19,20,21].

It also makes sense to examine simultaneously data from different kinds of networks. Recent approaches with methodology rooted in statistical physics focus on large networks, searching for universality both in the topology of the real networks and in the dynamics governing their evolution [49]. These combined theoretical and empirical results have opened unsuspected directions for researches and a wealth of applications in many fields ranging from computer science to biology and sociology [3,4,6,50,51]. In this respect, three important results have been crystallized: First, it has been found that the degree distribution contains important information about the nature of the network, for many large networks following the exponential distribution and the power-law distributions. Second, most networks have the so-called small world property [2], which means that the average distance between different pairs of nodes is rather small. Third, real networks display a degree of clustering coefficient higher than expected for random networks [2,4]. Finally, the assortative mixing is studied to answer why social networks are different from other types of networks [53].

The scientific studies can be considered as being organized within a network structure, which has a significant influence on the observed study collective behaviors. The viewpoints of complex networks are of interest in studying scientific study networks, to uncover the structural characteristics of the networks built on RAN. In this paper, the research areas of natural science basic research is studied from the point view of complex network [54]. In the fund management department, such as National Natural Science Foundation of China (NSFC), the research areas are denoted by the code system, which have the tree structure to demonstrate the inclusion relation between the research areas, such as Physics -> statistical physics -> complex network. The leave codes of the code system always represent the research areas more specially. To make the network reflect the reality more accurately, in this paper, the nodes are defined as the leave nodes of the code system. Because the scientists can fill in the fund proposal two codes: the first application code and the second one, then if one requisition paper filled in two different codes one can consider that the research work is cross the two research areas. The proposals filled in only one code are not considered in RAN. By this definition, there are 371, 349, 367, 400, 456, 544 nodes in RAN from 1999 to 2004. Three complementary approaches allow us to obtain a detailed characterization. First, empirical measurements allow us to uncover the topological measures that characterize the network at a given moment, as well as the time evolution of these quantities. Second, the average distance of RAN decreases with time, which means that the distance between any pairs of research areas is getting short. Third, the average clustering coefficient increases with time, which means that the neighbors of one research area would like to be connected with each other.
2 Data Analysis of RAN

In this section, the topology and dynamics of the empirical network are investigated. The parameters that are crucial to the understanding of the topology of RAN are extracted. The analysis of the data could provide the development trend of Chinese natural scientific basic research system.

2.1 Degree distribution follows the exponential form

The degree distribution $P(k)$ presents the probability that a randomly selected node has $k$ links, which has been studied extensively for various networks. Networks for which $P(k)$ has a power-law tail, are known as scale-free networks [51,55]. On the other hand, classical network models, including the Erdős-Rényi model [56,57] and the Watts and Strogatz model [1] have an exponentially decaying $P(k)$ and are collectively known as exponential networks. Another observed degree distribution form of real-life networks[6,58], named Stretched exponential distribution (SED) [59] is of the form $P(x)dx = \mu(x^{\mu-1}/x_0^\mu)e^{-(x/x_0)^\mu}dx$ and its accumulative distribution is $P(x) = \exp(-(x/x_0)^\mu)$, which can be stated as $\ln P(x) \sim x^\mu$. Obviously, SED degenerates to an exponential distribution when $\mu$ is close to 1 and to a power...
law when $\mu$ is close to 0. When $\mu$ is between 0 and 1, the degree distribution is between a power law and an exponential function. The data on a single-logarithmic plane show that the degree distribution of RAN can well fitted of an exponential form, see Fig. 1 and Fig. 2.

It seems difficult to find a common function to fit the empirical network because of the fluctuations. However, the accumulative distribution of RAN has less fluctuations and is more stable. The distribution of all the data from 1999 to 2004 can be fitted by an exponential distribution, given by $P(k) = \exp(0.13k)$. The hub nodes of RAN and their degrees from 1999 to 2004 are demonstrated in Table 1.

<table>
<thead>
<tr>
<th>Year</th>
<th>Hub nodes</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>Computer-aided design</td>
<td>28</td>
</tr>
<tr>
<td>2000</td>
<td>Intelligent information processing(IIP)</td>
<td>31</td>
</tr>
<tr>
<td>2001</td>
<td>IIP</td>
<td>37</td>
</tr>
<tr>
<td>2002</td>
<td>Pattern cognition, Artificial intelligence(PC &amp; AI)</td>
<td>33</td>
</tr>
<tr>
<td>2003</td>
<td>PC &amp; AI</td>
<td>52</td>
</tr>
<tr>
<td>2004</td>
<td>PC &amp; AI</td>
<td>63</td>
</tr>
</tbody>
</table>
Fig. 3. The network diagram of 2004. The marked points are the research areas of Network theory and Artificial Intelligence.

2.2 Average distance decreases with time

The ability of two nodes, $i$ and $j$, to communicate with each other depends on the length of the shortest path $d_{ij}$ between them. The average of $d_{ij}$ over all pairs named average distance $D = \frac{1}{N(N-1)} \sum_{ij} d_{ij}$.

Fig. 4. Average distance of RAN. The separation is computed on the data up to the indicated year.

Fig. 4 shows that $D$ decreases with time, which is highly surprising because most network models so far predict that the average distance should increases with system size. The decreasing trend observed by us could have different origins. It is possible that the increasing interconnectivity decreases the average distance. In analogy with the social networks, one can say that there are
3.8 degrees of separation between the research areas.

One can note the continuous declining of the average distance and the more connected nature of RAN fields expressed by a smaller separation. The monotone declining indicates that the scientists tend to study in crossing areas.

2.3 Average Clustering coefficient increases with time

The clustering coefficient $C_i$ can be defined as follows [1]: pick a node $i$ that has links to $k_i$ nodes in the system. If these $k_i$ nodes form a fully connected clique, there are $k_i(k_i - 1)/2$ links between them, but in reality we find much fewer. Denote the number of links that connect the selected $k_i$ nodes to each other by $E_i$. The clustering coefficient for node $i$ is then $C_i = 2E_i/[k_i(k_i - 1)]$. The average clustering coefficient for the whole network is obtained by averaging $C_i$ over all nodes in the system $C = \frac{1}{N} \sum_{i=1}^{N} C_i$. In simple terms, the clustering coefficient of RAN indicates the probability that two different nodes connect to the same node. The data of RAN indicates average clustering coefficient decrease with time, which can be seen from Fig. 5.

![Fig. 5. Average clustering coefficient of RAN, determined for the data up to the indicated year.](image)

2.4 Average degree increases

The number of nodes of RAN increases with time because of the arrival of new research areas. The growth of total number of links comes from two parts. One is the connections generated by new research areas with old ones, the other is the new connections between old nodes. A quantity characterizing the network’s interconnections is the average degree $\langle k \rangle$, denoting the average
number of links per node. The time dependence on \( \langle k \rangle \) of RAN is shown in Fig. 6.

![Fig. 6. Average number of links per node for RAN. Results are computed on the data up to the given year.](image)

2.5 
Assortative Coefficient

The assortative coefficient \( A \) measures the tendency of a network to connect vertices with the same or different degrees [52,53]. If \( A > 0 \), the network is said to be assortative, non-assortative when \( A = 0 \) and disassortative when \( A < 0 \). \( A \) is defined as follows

\[
A = \frac{1}{\sigma_q^2} \sum_j \sum_k (e_{jk} - q_j q_k),
\]

(1)

<table>
<thead>
<tr>
<th>Year</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>( v )</td>
<td>371</td>
<td>349</td>
<td>367</td>
<td>400</td>
<td>456</td>
<td>544</td>
</tr>
<tr>
<td>( \bar{k} )</td>
<td>4.264</td>
<td>4.768</td>
<td>4.986</td>
<td>5.235</td>
<td>6.197</td>
<td>6.379</td>
</tr>
<tr>
<td>( C )</td>
<td>0.097</td>
<td>0.113</td>
<td>0.107</td>
<td>0.121</td>
<td>0.154</td>
<td>0.197</td>
</tr>
<tr>
<td>( D )</td>
<td>5.367</td>
<td>4.806</td>
<td>4.525</td>
<td>4.273</td>
<td>4.060</td>
<td>3.765</td>
</tr>
<tr>
<td>isolate node</td>
<td>18</td>
<td>29</td>
<td>19</td>
<td>14</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>( A )</td>
<td>-0.0187</td>
<td>0.0033</td>
<td>0.0796</td>
<td>-0.0084</td>
<td>-9.7050e-004</td>
<td>-0.0899</td>
</tr>
</tbody>
</table>
where $e_{jk}$ is the probability that a randomly chosen edge has vertices with degree $j$ and $k$ at either end, $q_k = \sum_j e_{jk}$ and $\sigma_q^2 = \sum_k k^2 q_k - (\sum_k k q_k)^2$. The possible values of $A$ lies in the interval $-1 \leq A \leq 1$. Fig. 7 shows that the assortative coefficient begin to decline from 2001 and it decreases to the negative value from 2002. This indicates that RAN tend to be disassortative network.

![Fig. 7. Assortative coefficient of RAN. Results are computed on the data up to the given year.](image)

2.6 Dependence between clustering coefficient and degree

This intrinsic hierarchy can be characterized in a quantitative manner using the recent findings of Dorogovtsev et. al [45] and Ravasz et. al [46]. The hierarchical organization, implying that small groups of nodes organize in a hierarchical manner into increasingly large groups. In RAN, the relationships between clustering coefficient and the degree of clustering characterizing from 1999 to 2004 are demonstrated in Fig. 8.

From Fig. 8, one can see that there are no dependence between clustering coefficient and $k$ from 1999 to 2004, which means that there is no hierarchical organization in RAN.

3 Conclusion and Discussion

In this paper, we have constructed a research areas network and pointed out that such a network falls into the exponential distribution and small-world
networks categories. With the observation on the section 2, RAN tend to be a disassortative network. The average distance of RAN decreases with time, while the clustering coefficient increases with time. This indicates that the research areas are tend to be crossed studied by the scientists. The dependence between clustering coefficient and $k$ indicates that there is no hierarchical organization in RAN from 1999 to 2004.

The characteristics of RAN may lies in the following reasons.

1. The code system is changeless in a period of time, but the scientific study develops very fast, there are new research subjects emerge everyday. So some scientists can not find the suitable code which could describe their research content but only fill in the proposal two corelational codes.

2. The research subjects become more complexity, which can not be solved by one or two fields of knowledge. So the contents actuate the scientists who studied them to fill in their proposal two codes.

We hope that the work presented here may stimulate further research on this subject. Some opened questions are, for instance, whether the results obtained by RAN held for different countries? What is the evolving mechanism of RAN?

Acknowledgement

The authors are grateful to Dr. Qiang Guo, Bing Wang, Zhong-Zhi Zhang and Tao Zhou for their valuable comments and suggestions, which have led to a better presentation of this paper. This work has been supported by the National Science Foundation of China under Grant Nos. 70431001 and 70271046.
References

[54] Data were gathered from the fund management department of China.