Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 808901, 6 pages
http://dx.doi.org/10.1155/2013/808901

Research Article

Study of Evolution Model of China Education and Research Network

Guoyong Mao[1,2] and Ning Zhang[3]

[1] *Department of Electronic Information and Electric Engineering, Changzhou Institute of Technology, Changzhou 213002, China*
[2] *German Research School for Simulation Science, 52062 Aachen, Germany*
[3] *Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*

Correspondence should be addressed to Guoyong Mao; gymao@mail.shu.edu.cn

Received 2 September 2013; Revised 18 November 2013; Accepted 18 November 2013

Academic Editor: Zidong Wang

By searching the hyperlinks with domain name ".edu.cn" which constitutes the China Education and Research Network, we build a complex directed network containing 366,422 web pages containing 540,755 URLs. These URLs constitute a complex directed network through self-organization. By analyzing the topology of China Education and Research Network, we found that it is different from the common Internet in several aspects. Most of the vertices have incoming links, a few vertices have outgoing links, and very few vertices have both incoming and outgoing links. The vertex distribution has a power-law tail. A large proportion of newly added edges always connect with those pages selected from one subnetwork that they belong to, instead of connecting with the pages selected from the whole network. According to these features, we presented the evolution model of this complex directed network. The results indicate that this model reflects some main characteristics of China Education and Research Network.

1. Introduction

The research on complex networks is developing at a brisk pace, and significant achievements have been made in recent years; among them is the introduction of scale-free network and related models [1–4], as it makes big progress in revealing the characteristics of dynamic evolution of complex networks. Theoretical and empirical research on complex network has been carried out with some important achievements [5–9].

China Education and Research Network (CERNET) was established since 1995. More than 1000 universities and research institutes have been connected to this network so far. It has 36 regional network centers and main nodes, which are distributed among different provinces of China. As of now this network has host machines more than 1,200,000 and has become the second largest internet in China. However, compared with the large number of researches that has been done on the general Internet [10–13], only a few work is on CERNET can be found. From these studies we found that the features of CERNET are different from those of the general Internet, especially in the structure and formation mechanism [14, 15]. Hence, the study on CERNET is quite important.

We have been working on CERNET since 2005 and trying to establish the evolution model of CERNET for analysis and prediction purposes [14–16]. However, due mainly to the large scale of CERNET and lack of computing power, it took quite a long time to adjust the parameters to modify the model at that time. Therefore, the model we got is relatively simple which cannot well reflect the main features of CERNET [16]. For example, the average shortest path length of the simulation model is only about 2.8, far from 8.95 of the real network [17].

In this paper, the CERNET we analyze is a virtual network made up of web pages where ".edu.cn" is included in the addresses of all these pages. In this network, all web pages are nodes, and all the hyperlinks in these pages that link to other pages are the directed edges. This directed complex network has 366,422 nodes and 540,755 edges. We analyze the features of this network and extract the evolution model using empirical methods to reveal the formation mechanism of CERNET.

The remainder of the paper is organized as follows. Topological structure of CERNET is analyzed in Section 2, and the evolution model of CERNET and comparison between the real and simulated networks are described in Section 3, before giving conclusion and future work in Section 4.

2. Topological Structure of CERNET

There are several features that can be used to characterize a network, for example, the degree distribution, the average shortest path length, and the clustering coefficients. Among them the degree distribution is considered to be the most important [2].

From graph theory we know that the number of edges connected to one node is the degree of this node. For directed graph, the outdegree is the number of output edges and the indegree the number of input edges. Using the data we collect, we setup a database of CERNET and get the $P_{\text{out}}(k)$ and $P_{\text{in}}(k)$, where $P_{\text{out}}(k)$ is the probability that one page has k output pages and $P_{\text{in}}(k)$ is the probability that one page has k input pages. The formulas we use to calculate the output and input probability of node i are listed in (1) and (2), respectively, where M_{out} is the maximum outdegree of the network and M_{in} the maximum indegree of the network:

$$P_{\text{out}}(k_i) = \frac{(k_i)}{\sum_{j=1}^{M_{\text{out}}}(k_j)}, \tag{1}$$

$$P_{\text{in}}(k_i) = \frac{(k_i)}{\sum_{j=1}^{M_{\text{in}}}(k_j)}. \tag{2}$$

We plot the double logarithmic curves of $P_{\text{out}}(k)$ and $P_{\text{in}}(k)$ that change as a function of k, as shown in Figures 1 and 2, respectively. Linear-regression analysis is done on the linearized data, as shown in the straight red lines in these figures. From Figure 1 we see that the tail of outdegree distribution of CERNET follows the power law distribution, $P_{\text{out}}(k) \sim k^{-r_{\text{out}}}$, where $r_{\text{out}} = 2.48$. From Figure 2 we see that the indegree distribution generally follows the power law distribution, but the tail is not very smooth, $P_{\text{in}}(k) \sim k^{-r_{\text{in}}}$, where $r_{\text{in}} = 2.40$, which differs greatly with the Poisson distribution predicted using the traditional theory of random graph.

We make statistical analysis of these data and get the accumulated frequency of degree and the corresponding ratio of the degree to total degree in CERNET, as shown in Table 1. From Table 1 we can see that a large amount of pages have small connections, a few pages have a medium number of connections, while a tiny minority of notable pages have a large number of connections. This phenomenon is similar to the research result made by Albert et al. [1].

This virtual network of CERNET is made up of subsets of web pages of different universities. The number of web pages of each subset is determined by the corresponding universities; the addition and deletion of pages totally depended on the university that these pages belong to. However, we find that though the number of pages is different for different universities they do share some similar features. For example, the proportion of pages that have output links to the total number of pages is less than 25% in every university, while the proportion of pages that have input links to the total number of pages is usually bigger than 85%. Only a very small number of pages have both output links and input links. Hence, if each university is treated as a subnetwork, then in each network most nodes only have input edges, a few nodes only have output edges, and the number of nodes with both

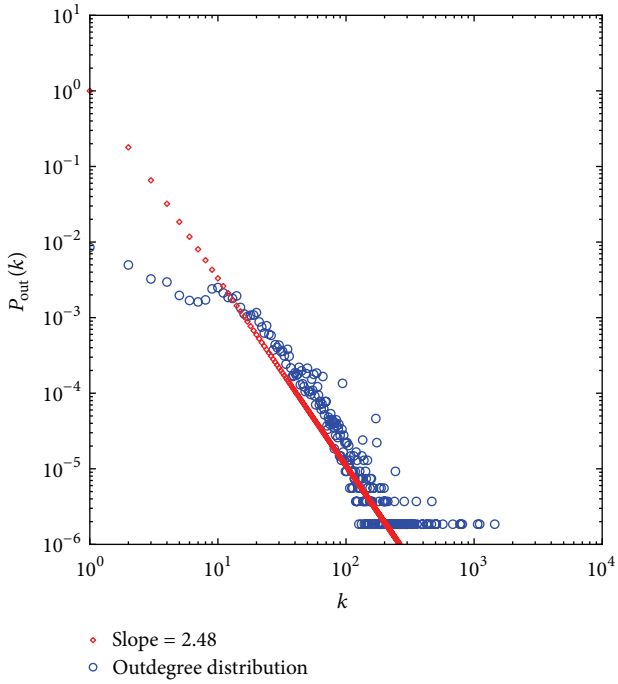

Figure 1: Distribution of outdegree of real data.

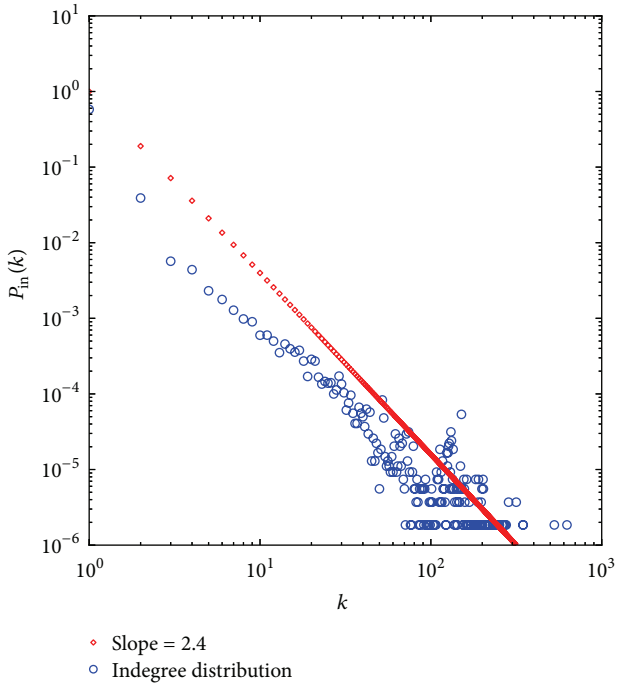

Figure 2: Distribution of indegree of real data.

Table 1: The accumulated frequency and percentage of degree in CERNET.

Outdegree	Accumulated frequency	Ratio to total outdegree (%)	Indegree	Accumulated frequency	Ratio to total indegree (%)
1~50	30463	93.7	1~5	341535	98.0
51~300	32478	6.2	6~199	348477	1.9
301~1449	32510	0.1	201~626	348491	0.1

Table 2: Degree and link features in some universities.

Name of university	Number of web pages	Number of outdegrees	Number of indegrees	Ratio of pages with output edge to total number of pages (%)	Ratio of pages with input edge to total number of pages (%)
CIM	9576	14532	13515	0.15	0.87
SHUFE	9151	13546	12620	0.12	0.93
ZJU	11537	20586	18943	0.085	0.973
CQU	864	1403	1326	0.075	0.987
NBU	1985	2995	2785	0.159	0.908
SHU	15035	19443	18522	0.096	0.939
SUDA	13643	20197	18051	0.071	0.964
CUMT	9371	18089	12293	0.134	0.914
SHISU	10733	13734	12941	0.089	0.933
ECUN	13707	19890	17379	0.068	0.959
SHSMU	3663	6335	5730	0.114	0.916
CUN	6120	7020	6762	0.18	0.84

input edges and output edges is rare. From these features we know that each university connects to other universities through a small number of pages, as shown in Table 2.

3. The Evolution Model of CERNET

Using the mechanism of growth and preferential attachment, the scale-free model proposed by Barabasi et al. can to some degree disclose the nature of many complicated phenomena in the practical world. However, this model cannot be applied to CERNET. For example, every newly attached node has output edges in this scale-free model, but for the directed network of CERNET a larger amount of newly attached nodes have only one input edge; that is, these nodes have zero outdegree. Also in this model, the preferential attachment of newly added nodes will search the whole network for the best node to connect to, while in CERNET the newly added pages will generally choose some pages in the same university to connect to. Only occasionally, the newly added pages will choose pages in other universities, but these pages will not search the whole CERNET for the best pages to connect to. From these features of CERNET, we propose the evolution model of CERNET, as follows.

(i) The CERNET starts from m_0 nodes and e_0 edges. The m_0 nodes are randomly divided into l subsets. There are $m_{01}, m_{02}, \ldots$, and m_{0l} nodes and $e_{01}, e_{02}, \ldots$, and e_{0l} edges in each subset, respectively, where $\sum_{i=1}^{l} m_{0i} = m_0$, and $\sum_{i=1}^{l} e_{0i} = e_0$.

(ii) At each moment, a new node will randomly be added into one of the subsets of the network. There are 5 cases for the edges that are added together with the new node:

(1) the new node has only one input edge;
(2) the new node has only m output edges;
(3) the new node has one input edge and one output edge;
(4) the new node has one input edge and $m - 1$ output edges;
(5) the new node has one output edge and $m - 1$ input edges,

where $m \leq (m_{0\min})$ and $m_{0\min}$ is the minimum initial number of nodes among l subsets and $m_{0\min} = \min(m01, m02, \ldots, m0l)$.

(iii) When the new node with one input edge is added to the network with probability α, this node will randomly choose a subset and let itself be connected by a preferentially selected node in this subset. Let $\prod_{\text{out}}(i)$ denote the probability of node i to be selected as the source node; then $\prod_{\text{out}}(i)$ is determined by k_{out}^{i}, the outdegree of i.

(iv) When the new node with m output edges is added to the network, there are 2 cases we should consider. The probabilities of the two cases are β_1 and β_2, respectively.

(1) For the first case, the new node will randomly choose a subset and let itself connect to a preferentially selected node in this subset. Let $\prod_{\text{in}}(i)$ denote the probability of node i be selected as the target node; then $\prod_{\text{in}}(i)$ is determined by k_{in}^i, the indegree of i. For the rest of the $m-1$ output edges, at each moment only one edge randomly chooses a subset which has not been connected by the new node and connects itself to a preferentially selected node in this subset, till all $m-1$ output edges are processed.

(2) For the second case, the new node will still randomly choose a subset, but this time this node will preferentially choose $m-1$ nodes in this subset and let itself be connected. Let $\prod_{\text{in}}(i)$ denote the probability of node i be selected as the target node; then $\prod_{\text{in}}(i)$ is determined by k_{in}^i, the indegree of i. For the rest of the edges that this new node carries, it will randomly pick a subset which has not been connected by this new node and connect itself to a preferentially selected node in this subset.

(v) When the new node with one input edge and one output edge is added to the network, there are also 2 cases we should consider. The probabilities of the two cases are γ_1 and γ_2, respectively.

(1) For the first case, the new node will randomly choose a subset and let itself be connected by a preferentially selected node in this subset. The probability of node i to be selected as the source node is determined by k_{out}^i, the outdegree of i. The output edge of the new node will randomly select a subset which has not been connected by the new node and connect itself to a preferentially selected node. The probability of a node i to be selected as the target node is determined by k_{in}^i, the indegree of i.

(2) For the second case, the new node will randomly choose a subset and let itself be connected by a preferentially selected node in this subset. The probability of node i to be selected as the source node is determined by k_{out}^i, the outdegree of i. For the output edge that this new node carries, it will still pick a node in the same subset and connect itself to a preferentially selected node which has not been connected by the input edge of the new node. The probability of a node i to be selected as the target node is determined by k_{in}^i, the indegree of i.

(vi) When the new node with 1 input edge and $m-1$ output edges is added to the network with probability δ, this node will randomly choose a subset and let itself be connected by a preferentially selected node in this subset. The probability of node i to be selected as the source node is determined by k_{out}^i, the outdegree of i. For the rest of the $m-1$ output edges, at each moment only one edge randomly chooses a subset which has not been connected by the new node and connects itself to a preferentially selected node in this subset, till all the $m-1$ output edges are processed. The probability of node i to be selected as the target node is determined by k_{in}^i, the indegree of i.

(vii) When the new node with 1 output edge and $m-1$ input edges is added to the network with probability ζ, this node will randomly choose a subset and connect itself to a preferentially selected node in this subset. The probability of node i to be selected as the target node is determined by k_{in}^i, the indegree of i. For the rest of the $m-1$ input edges, at each time only one edge randomly chooses a subset which has not been connected by the new node and lets itself be connected to a preferentially selected node in this subset, till all $m-1$ input edges are processed. The probability of node i to be selected as the source node is determined by k_{out}^i, the outdegree of i.

The definitions of $\prod_{\text{in}}(i)$ and $\prod_{\text{out}}(i)$ are listed in (3) and (4), respectively. The relation between different probabilities is listed in (5). We have the following equations:

$$\prod_{\text{in}}(i) = \frac{k_{\text{in}}^i}{\sum_{j=1}^{n_l}\left(k_{\text{in}}^j\right)}, \tag{3}$$

$$\prod_{\text{out}}(i) = \frac{k_{\text{out}}^i}{\sum_{j=1}^{n_l}\left(k_{\text{out}}^j\right)}, \tag{4}$$

$$\alpha + \beta_1 + \beta_2 + \gamma_1 + \gamma_2 + \delta + \zeta = 1. \tag{5}$$

In (3) and (4), n_l is the number of nodes of the subset that has new edges connected to it. The denominator of (3) is the sum of indegree of the same subset and the denominator of (4) is the sum of outdegree in this subset.

After t moments, we get a directed random network with N nodes and V edges, where $N = m_0 + t$, and

$$\begin{aligned} V = {} & e_0 + \alpha * t + (\beta_1 + \beta_2) * m * t \\ & + (\gamma_1 + \gamma_2) * 2t + (\delta + \zeta) * m * t. \end{aligned} \tag{6}$$

From the analysis of CERNET we set $\alpha = 0.60$, $\beta_1 = 0.2$, $\beta_2 = 0.12$, $\gamma_1 = 0.04$, $\gamma_2 = 0.02$, $\delta = 0.01$, and $\zeta = 0.01$. When $m_0 = 12$, $m = 3$, and $l = 3$, we get the distribution of outdegree and indegree of this simulated model. The outdegree and indegree distributions are illustrated in Figures 3 and 4, respectively. Figures 5 and 6 illustrate the comparison between the simulated data and the real data. From the comparison of outdegree distribution we can see that the slope of the simulated data is 2.48, the same as that of the real data, but the beginning part of the simulated data cannot

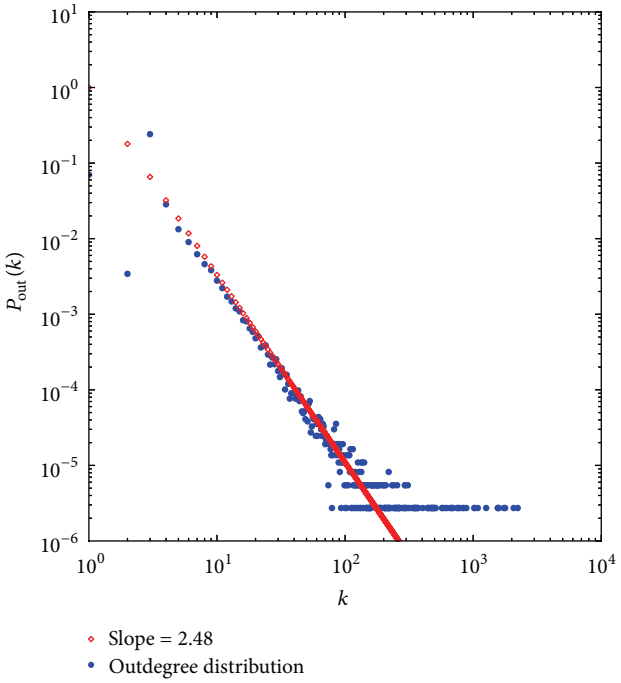

Figure 3: Distribution of outdegree of simulated data.

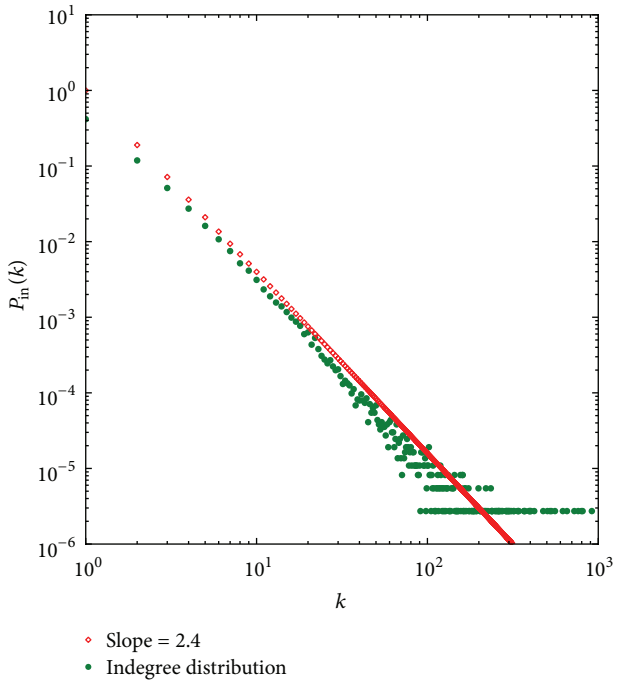

Figure 4: Distribution of indegree of simulated data.

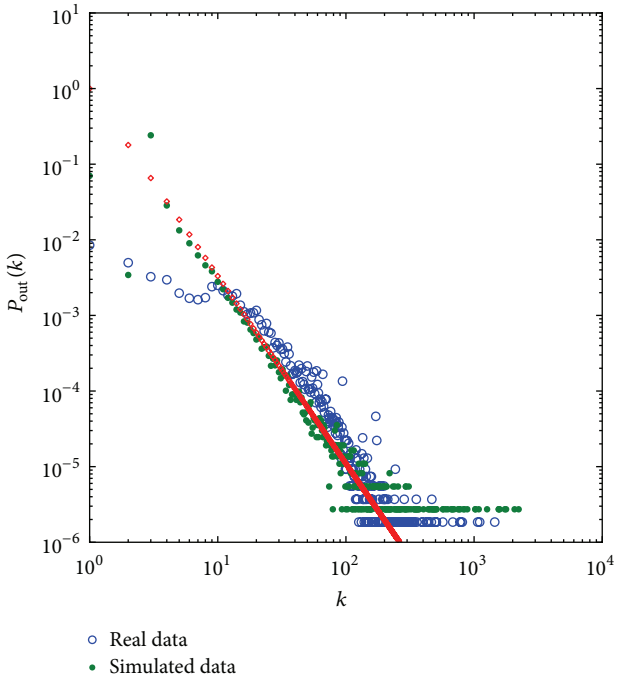

Figure 5: Comparison of outdegree distribution.

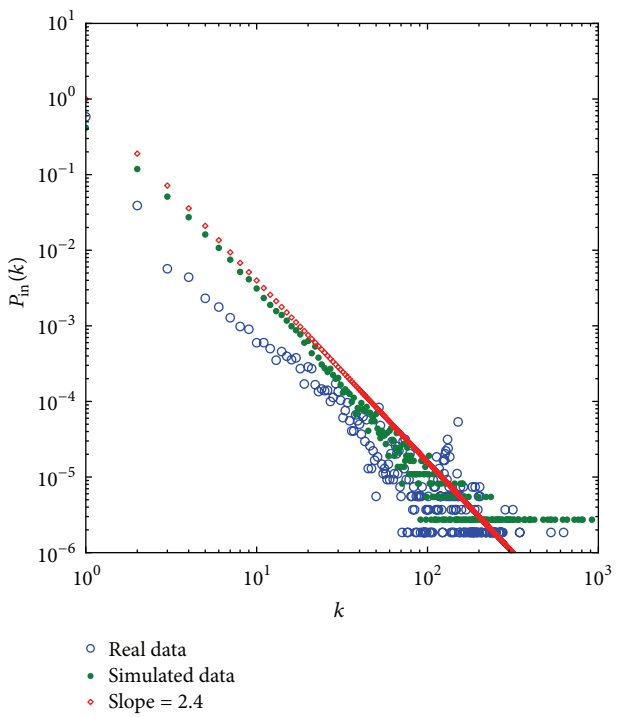

Figure 6: Comparison of indegree distribution.

fully reflect the statistical result of the real data. From the comparison of indegree distribution we see that the slope of simulated data is 2.40, the same as that of the real data, but the beginning part of the simulated data cannot fully reflect the statistical result of the real data. The tail is smoother than that of the real data. The slope is 2.40, the same as the real data.

4. Conclusions

From the figures of degree distribution, we can see that the simulated network can partly reflect the characteristic of CERNET. The degree distribution of the simulated network matches much better the real network than that in model [16].

We also compared other features of the simulated and the real networks. For example, the average shortest path length for the real network is 8.95, while for the simulated network, it is 7.81, which is much closer than that of the model listed in [16].

The main contribution of this paper is the evolution model of the CERNET. The result shows that the simulated model can partly disclose the property of this network. However, the model introduced in this paper is only the ideal model, which means that only the main features of the real network are considered. With the help of the fast growing computing power, we intend to adjust this model so that it can be used in the analysis of the ever increasing large scale complex networks.

Acknowledgments

The authors are grateful to colleagues in the German Research School for Simulation Science for their constructive suggestions. The authors are also grateful to the reviewers for their valuable comments and suggestions to improve the presentation of this paper. This work is supported by the National Natural Science Foundation of China under Grant no. 70971089, Shanghai Leading Academic Discipline Project under Grant no. XTKX2012, and Jiangsu Overseas Research and Training Programs for University Prominent Young and Middle-Aged Teachers and Presidents.

References

[1] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[2] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[3] A.-L. Barabási, R. Albert, and H. Jeong, "Mean-field theory for scale-free random networks," *Physica A*, vol. 272, no. 1-2, pp. 173–187, 1999.

[4] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[5] X. Li and G. Chen, "A local-world evolving network model," *Physica A*, vol. 328, no. 1-2, pp. 274–286, 2003.

[6] Q. Chen and D. Shi, "The modeling of scale-free networks," *Physica A*, vol. 335, no. 1-2, pp. 240–248, 2004.

[7] H. Che and J. Gu, "Scale-free networks and their signicance for systems science," *Systems Engineering-Theory & Practice*, vol. 24, pp. 11–16, 2004.

[8] J. Wu and Z. Di, "Complex networks in statistical physics," *Progress in Physics*, vol. 24, pp. 18–46, 2003.

[9] A. Lancichinetti and S. Fortunato, "Consensus clustering in complex networks," *Scientific Reports*, vol. 2, no. 336, 2012.

[10] X. Wang and D. Loguinov, "Wealth-based evolution model for the internet AS-level topology," in *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM '06)*, pp. 1–11, Barcelona, Spain, April 2006.

[11] S. Zhou, "Understanding the evolution dynamics of internet topology," *Physical Review E*, vol. 74, no. 1, Article ID 016124, 11 pages, 2006.

[12] A. Dhamdhere and C. Dovrolis, "Ten years in the evolution of the internet ecosystem," in *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement (IMC '08)*, pp. 183–196, ACM, Vouliagmeni, Greece, October 2008.

[13] L. Šubelj and M. bajec, "Model of complex networks based on citation dynamics," in *Proceedings of the 22nd International Conference on World Wide Web Companion (WWW '13)*, pp. 527–530, Rio de Janeiro, Brazil, May 2013.

[14] N. Zhang, "Analysis of China education network structure," *Computer Engineering*, vol. 33, no. 17, pp. 140–142, 2007.

[15] L. Su, N. Zhang, and L. Ma, "Comparative studies on the topological structure of China education network," *Journal of University of Shanghai for Science and Technology*, vol. 30, no. 3, pp. 297–299, 2008.

[16] N. Zhang, "Complex network demonstration—China education network," *Journal of Systems Engineering*, vol. 21, no. 4, pp. 337–340, 2006.

[17] X. Ni, N. Zhang, and M. Wang, "Parallel algorithms (MPI) on solving the shortest-path problem of china educational network," *Computer Engineering and Applications*, vol. 15, pp. 135–137, 2006.