Commun. Comput. Phys.
doi: 10.4208/cicp.190909.161209a

Vol. 8, No. 3, pp. 690-700
September 2010

Comparisons of the English and Chinese Language Networks: Many Similarities and Few Differences

Dujuan Wang[1,*], Ru Wang[2] and Xu Cai[1]

[1] *Complexity Science Center, Institute of Particle Physics, Central China Normal University, Wuhan 430079, China.*
[2] *College of Information Science and Engineering, Huaqiao University, Quanzhou 362021, China.*

Received 19 September 2009; Accepted (in revised version) 16 December 2009

Communicated by Dietrich Stauffer

Available online 20 April 2010

Abstract. With words as nodes, and a link exits between two neighboring words, the weighted directed English and Chinese written human language networks are constructed from one English novel and two Chinese ones. We hereby analyze in detail the topological structure of them, in order to clarify their similar and different statistical properties. The empirical results show that the English and Chinese language networks all possess the shifted power law (SPL) degree distribution, the small-world property and the hierarchical structure, the connections among the words have positive assortativity coefficient and reciprocal characteristics. We also investigate the features of the strength and the centrality, which describe the importance of a specific word. Furthermore, considering the growth properties of the language networks and part of topological property, we find that the English written human language network grows slower than the Chinese one, which implies different mechanisms of the English and Chinese languages.

AMS subject classifications: 05C82, 91D30, 68M10

PACS: 89.75.-k, 89.75.Fb, 43.71.Sy

Key words: Topological structure, shifted power law, correlation, growth property.

1 Introduction

The past few years have witnessed people's great interest with regard to the complex networks. By investigating many real world networks, the small-world behavior [1, 2] and the scale-free property [3] were successfully confirmed, typical examples comprise

[*]Corresponding author. *Email addresses:* wangdj@iopp.ccnu.edu.cn (D. Wang), wr0124@gmail.com (R. Wang), xcai@mail.ccnu.edu.cn (X. Cai)

the World Wide Web [4], the collaboration network [5], the public transportation networks [6] and the graph of human language [7, 8], etc. As well known, the characterization of the topological structure [9–11] of a network is the basic factor to analyze its intrinsic functions and dynamics [12–16]. These empirical analyses have inspired people to probe the universality of the real world systems, and thus to provide an appropriate framework [17, 18] for developing techniques and models of the complex networks.

Composed of a number of words, novels and poems are simply normal examples of written human language networks in nature, and thus can be studied from the aspect of the complex network theory. Caldeira and Lobao [19] study the structure of meaningful concepts in written texts, they find the small-world effect as well as the scale-free structures. Li and Zhou [20] emphasize the Chinese character structure, supposing that the radical is the vertex and two vertices are linked if they can form a character or a part of it. Their work shows that the character networks also display the small-word property and the non-Poisson distribution. Masucci and Rodgers [21] investigate the English novel named 1984, they find the existence of different functional classes of vertices, the significance of the second order vertex correlations in the network architecture.

The previous works are of great importance to understand the nature of the written human language networks. However, to our best knowledge, they just concentrate on one language, and there are no comparisons about the characteristics of different languages. Therefore, in this paper, our main purpose is to investigate the similarities and differences between English and Chinese written human language networks. We select three novels [22] as our empirical objects, which include a Chinese one named *"A Q Zheng Zhuan"* (*AQC*) written by Lu Xun in 1921, the English version *"The true story of Ah'Q"* (*AQE*) translated by Yang Hsien-yi in 1960, and another Chinese one entitled *"Kun Lun Shang"* (*KLS*) written by Bi Shumin in 1986.

In our studies, word represents the vertex, an edge exists between two vertices if they are neighbors, and the edge directs from the former word to the latter one. Neglecting the punctuation marks and the paragraph gaps, we construct the weighted directed English and Chinese written human language networks, the system sizes are 21118, 17204, 23270, the number of different words are 1553, 2661, 2048 for these three networks respectively.

The whole text is organized as follows: we show the topological property of the networks such as degree distribution and clustering in Section 2. Section 3 presents the weighted network. Section 4 depicts the centrality and betweenness measures. In Section 5, we exhibits the growth properties of the networks. Conclusion and discussion are given in Section 6.

2 The topology of network

The foremost quantity that describes the characteristic of the network is the degree distribution. In order to reduce the statistical errors arising from the limited system size, we introduce the Pareto distribution, which is regarded as the same thing as Zipf, power-

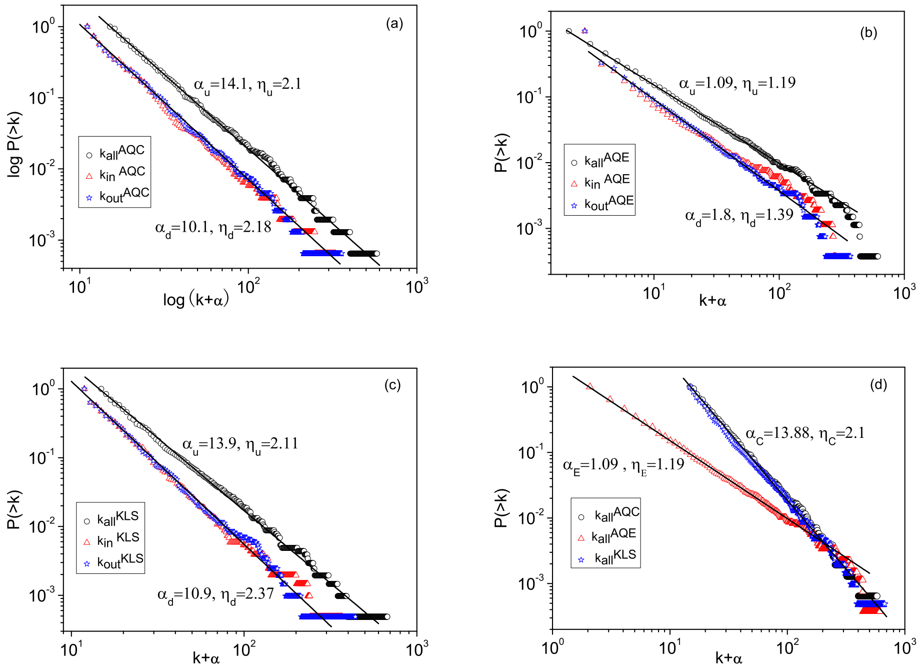

Figure 1: The cumulative distributions of in-degree, out-degree and all-degree for (a) AQC, (b) AQE, (c) KLS, respectively. (d) The cumulative distribution about the all-degree of the three networks. α_d, η_d, α_u, η_u stand for the directed and undirected networks, α_C, η_C, α_E, η_E stand for the Chinese and the English language networks respectively.

law distribution [23]. We represent k_{in} and k_{out} as the incoming and outgoing degree of a given node, and k_{all} as the degree when we do not distinguish incoming and outgoing. We show the cumulative in degree, out degree and all degree distributions of the three networks in Fig. 1.

The distributions show the so-called shifted power law (SPL) function form [24], which is analytically deduced partially by the random selection and partially by linear preferential principle, as this format:

$$P(>k) \propto (k+\alpha)^{-\eta}, \tag{2.1}$$

where $0<\alpha<\infty$, for $\alpha=0$, the SPL shows a power law distribution, while $\alpha \rightarrow \infty$, the SPL tends to an exponential distribution, the typical SPL functions can be shown only with values of α between 1 and 100 [5].

From Fig. 1(a)(b)(c), we find that the in-degree and out-degree distributions are almost consistent with each other, while great deviations exist in the undirected degree distributions. This is a prominent distinction, since most of the directed and undirected degree distributions of many other real world networks lap over with each other.

As well known, due to the strict grammar rules and matching relations, the language networks are directed ones, and they are not bidirectional, so this extraordinary

Table 1: The probabilities of the in degree, out degree and all degree of the three networks.

	AQE			*AQC*			*KLS*		
k	P_{in}	P_{out}	P_{all}	P_{in}	P_{out}	P_{all}	P_{in}	P_{out}	P_{all}
1	0.588	0.569	0.000	0.372	0.367	0.000	0.364	0.361	0.000
2	0.164	0.159	0.540	0.158	0.178	0.321	0.168	0.171	0.331
3	0.074	0.077	0.051	0.106	0.099	0.054	0.094	0.092	0.045
4	0.042	0.042	0.127	0.071	0.064	0.129	0.067	0.072	0.133
5	0.025	0.026	0.032	0.048	0.047	0.046	0.046	0.037	0.030
6	0.016	0.021	0.052	0.032	0.037	0.062	0.035	0.045	0.056
7	0.017	0.015	0.019	0.029	0.027	0.035	0.035	0.031	0.026
8	0.009	0.014	0.026	0.019	0.027	0.034	0.024	0.026	0.047
9	0.005	0.006	0.011	0.017	0.018	0.030	0.017	0.017	0.025
10	0.007	0.008	0.014	0.016	0.014	0.021	0.015	0.016	0.032

Table 2: The fundamental parameters of the three language networks.

	AQE	*AQC*	*KLS*
N	2661	1553	2048
$\langle k \rangle$	7.878	12.546	12.474
C	0.315	0.256	0.186
C_{rand}	0.00296	0.008	0.006
L	3.372	3.079	3.548
L_{rand}	7.885	7.340	7.623
r	0.114	0.012	0.096
ρ	-0.015	-0.005	-0.010

and unique property results in the imbalance between the undirected degree distribution and the directed ones. For the whole network, the in degree and out degree of different words obey the principle of equality and mutual benefit, when the in degree of a word grows, the out degree of its neighboring word will grow correspondingly, therefore this inevitably leads to consistency of the cumulative in degree and out degree distributions. Table 1 intuitionally depicts the characteristic of the directed and undirected degree distribution, and confirms the above conclusions.

In Fig. 1(d), the degree distributions of the two Chinese language networks coincide with each other, while the degree distribution of the English one is different, though the story of the English novel *AQE* and the Chinese one *AQC* are the same, which may be attributed to the different vocabulary size and the grammar rules of different languages.

The assortative and reciprocity properties of the network can measure the degree correlations of the network:

The assortativity coefficient r shows whether nodes of high degree connect to the nodes of high degree or low degree. From Table 2, the assortativity coefficients r of the three networks are all greater than 0, which means the language networks are all assortative networks, vertices of high degree tend to connect with vertices of the same kind,

Table 3: The frequencies (Fre) of the top ten binary structure.

AQE		*AQC*		*KLS*	
structure	Fre	structure	Fre	structure	Fre
Ah Q	279	阿 Q	275	一个	61
of the	80	没有	92	什么	60
he had	76	一个	61	自己	54
in the	75	知道	53	昆仑	43
to the	66	自己	46	拉练	36
he was	53	因为	45	没有	33
it was	37	什么	45	起来	32
to be	33	他的	43	他们	27
had been	32	然而	41	他的	26
did not	29	有些	39	知道	24

hence the positive correlations exist in the degree. Generally the social networks are the assortativity ones [25] and our empirical analysis confirms this conclusion.

While the reciprocity property [26] of the degree reflects whether the in-degree equals the out-degree of a given node, the reciprocity coefficient ρ indicates this trend of the whole network. In Table 2, the reciprocity coefficients ρ are all close to 0, which indicates the written human language networks are not balanced in connecting and being connected, since there exists so many binary structures in the language networks, they appear as a whole in the daily use. Table 3 lists the most frequent binary structure in the three networks, take 'it was' for example, there's just one arrow point from 'it' to 'was', for 'it', its out degree will not grow as quickly as its in degree, while for 'was', its in degree will grow much slower than its out degree, and this results in the disproportion of the in degree and out degree of a specific word.

The results of the clustering coefficients C are shown in Table 2, we can find that all of them are less than 0.32, which suggests only a few triangles present in the language networks, and this should be attributed to the selectiveness of syntax structures. For comparison, random graphs of the same average degree $\langle k \rangle$ and the same nodes N are investigated, it is manifest that the average clustering coefficient of the language networks is much greater than that of the random graphs.

Furthermore, we analyze the clustering spectrum in Fig. 2(a), which shows the empirical degree-dependent clustering coefficient follows the pow-law decay $\langle C(k) \rangle \sim k^{-1.0}$. This indicates there exists the hierarchical and modular structure [27] in the language networks, which is universal for many other real world networks.

The path length distribution is presented in Fig. 2(b). Comparing with the corresponding random graph, $L_{rand} = \ln k / \ln N$, we find the average shortest-path lengths of the written human language networks are much smaller than that of the random graph. Some fundamental parameters of the three language networks are listed in Table 2. The high average clustering coefficient and the small average shortest-path length indicate the small-world property of the language networks.

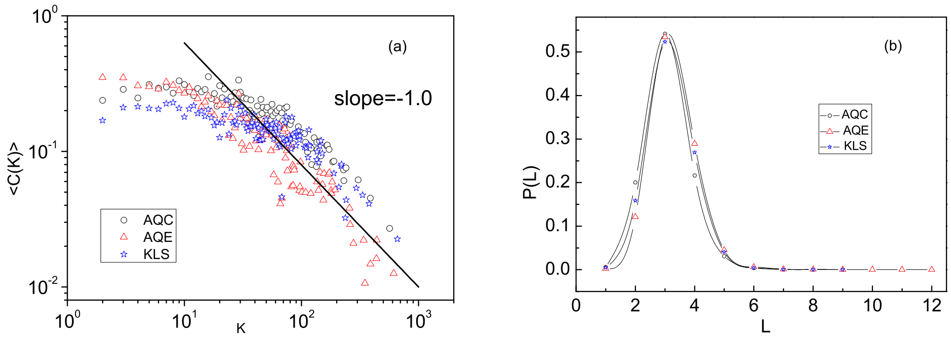

Figure 2: (a)The degree-dependent clustering coefficient $\langle C(k)\rangle$ as a function of the degree k. (b) The path length distribution.

3 The weighted network

In this part, we investigate the representation of weighted version, the weight of an edge is defined as the frequency of a connection appearing in the whole network, while the strength of a vertex (the sum of the weight values of all edges passing through it) denotes the frequency of a word appearing in the network. The weight and the strength can provide information about the importance of the edges and the words. Taking the Chinese language for example, the Chinese word vocabulary in common use has no more than three thousand, whereas they can constitute many novels with much more larger size, due to the property of strength, which is defined as $S_i=\sum_j W_{ij}$, where W_{ij} represents the weighted matrix of the language networks. We find the strength distributions exhibit the same properties as that of the degree distribution, that is, the strength distributions of two Chinese language networks coincide with each other and differ from the English one. From Fig. 3(a), it is found that all of them follow SPL function form:

$$P(>s)\propto(s+\beta)^{-\gamma}. \tag{3.1}$$

Resorting to the original data, we find that vertices with large strength tend to form the binary structures, and are likely to constitute the phrases or the short sentences (Table 3).

We depict the relationship of the average strength $\langle s(k)\rangle$ as a function of degree k in Fig. 3(b), and find that it follows a power-law, with the format $\langle s(k)\rangle\sim k^{1.15}$. It implies the larger the degree, the larger the strength and the more important the word in the network. To be specific, we list the top ten words with large degree and their corresponding strength in Table 4. We find that words of large degree and strength have the same meanings and functions in the two different languages, most of which are the adjectives and the conjunctions.

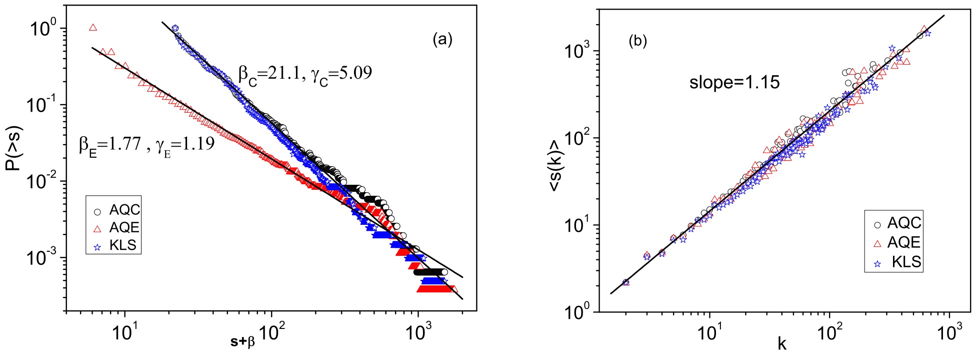

Figure 3: (a)The cumulative strength distribution of the three networks. (b)Average strength $\langle s(k)\rangle$ as a function of the degree k of nodes. β_C, γ_C, β_E, γ_E stand for the Chinese and the English language networks.

Table 4: Words of the top ten degrees.

AQE			AQC			KLS		
i	k_i	s_i	i	k_i	s_i	i	k_i	s_i
the	651	880	的	668	745	的	790	788
and	476	369	了	438	476	了	431	411
to	465	514	他	345	385	一	384	535
a	395	370	不	268	317	不	343	326
of	361	331	一	267	348	着	254	188
he	355	451	是	261	307	地	251	199
his	282	280	在	202	188	是	247	242
was	274	317	也	201	195	他	244	219
in	267	251	有	186	255	在	226	173
it	198	173	来	183	197	上	203	171

4 Centrality and betweenness

In order to shed more light on understanding the structural properties of the networks, we employ centrality and betweenness to quantify the importance of a word. The betweenness [28] b_i is defined as the probability of the shortest paths connecting any two words that involve a connection with the word i. We plot the cumulative distribution of the betweenness in Fig. 4, which follows the power-law decay, $P(>b)\sim b^{-1.16}$.

The average normalized betweenness $\langle b(k)\rangle$ as a function of degree k is analyzed in Fig. 5(a), we can find that $\langle b(k)\rangle \sim k^{1.33}$, which indicates words with large degree correspondingly have large betweenness, highly connected words are also the most central words.

The convenience of a given node to reach many other nodes in the network also reflects its centrality. We defined L_i as the average shortest path length from a certain vertex

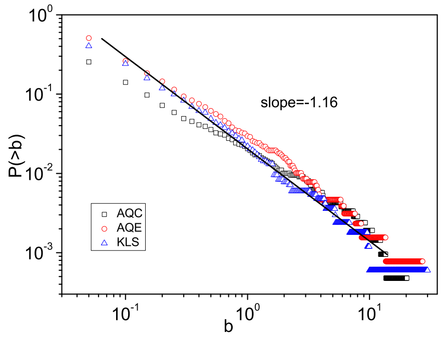

Figure 4: Cumulative distribution of the normalized betweenness b.

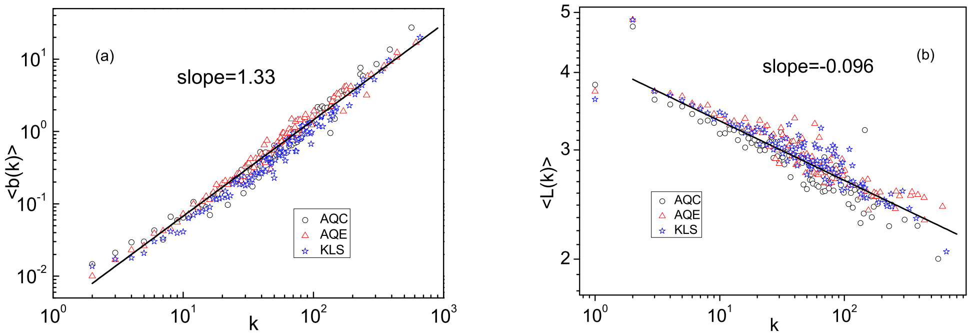

Figure 5: (a) Relationship between the average normalized betweenness $\langle b(k)\rangle$ vs degree k. (b)Average shortest path length from a certain node to all other nodes $\langle L(k)\rangle$ vs degree k.

i to all other vertices as the follows: $L_i=(\sum_{j\neq i}L_{ij})/(N-1)$, where L_{ij} is the shortest path length between words i and j, a small value of L_i implies that it is convenient for the word i to connect with the other words in the network. Fig. 5(b) shows $\langle L(k)\rangle$ as a function of degree k, with the form $\langle L(k)\rangle\sim k^{-0.096}$, which can be easily understood for the reason that words with larger degree have shorter path length to go to many other words.

5 Growth property

The language network is an accelerated growing network, for the number of edges grows faster than the number of vertices. We define t as the time step when a new word is added to the text and $N(t)$ is the total number of words in the text, we find the empirical growing properties [21] of the three networks behave as:

$$N(t)\sim t^{\lambda}. \tag{5.1}$$

More specifically, t represents the number of different words used to comprise the

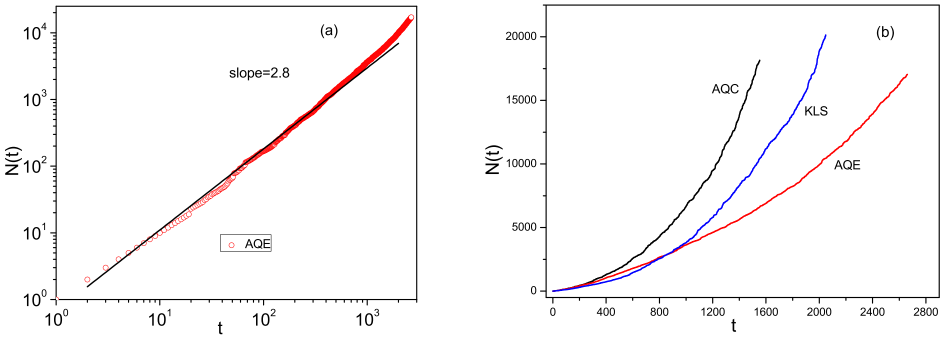

Figure 6: (a) The growing curve of AQE. (b) The comparison of the growing properties for the three networks.

text or the used vocabulary size, while $N(t)$ represents the total text size. Fig. 6(a) shows the growth of AQE in log-log scale, with the power law form $N(t) \sim t^{2.8}$. For comparison, we show the growing functions of the three networks in linear scale in Fig. 6(b), which shows the two Chinese language networks grow faster than the English one. This can be interpreted in the following way. Firstly, the vocabulary of the English language is much larger than that of the Chinese which lead to the large adding probability of the next new word for the English language. Secondly, the more flexible of the English word connections yield the slower growth of English language network.

6 Conclusion and discussion

In this paper, we implement an empirical analysis of the English and Chinese written human language networks. The empirical results show that the cumulative degree and strength distributions all follow the shifted power law function format, there exists hierarchical structure and small-world effect in the language networks, and the degrees are positively correlated. Moreover, large degree nodes are usually the center node according to the betweenness and centrality measure.

Furthermore, we find that the degree distributions with and without direction consideration are different for these two language networks. Secondly, the degree and strength distributions vary for different languages and have no relationship with the content. In addition, the English language network grows slower than the Chinese one. We explain this as the different mechanisms of the English and Chinese languages and future work will try to work on these mechanisms.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 10647125, 10635020, 10975057 and 10975062) and the Programme of Introducing Talents of Discipline to Universities under Grant No. B08033.

References

[1] D.J. Watts, S.H. Strogatz, Collective dynamics of ’small-world’ networks, Nature 393 (1998) 440.

[2] L. Guo, X. Cai, Opinion Dynamics of Sznajd Model on Small-World Network, Commun. Comput. Phys 6 (2009) 586.

[3] A.L. Barabasi, R. Albert, Emergence of Scaling in Random Networks, Science 286 (1999) 509.

[4] A.L. Barabasi, R. Albert, Mean-field theory for scale-free random networks, Physica A 272 (1999) 173.

[5] H. Chang, B.B. Su, Y.P. Zhou, D.R. He, Assortativity and act degree distribution of some collaboration networks, Physica A 383 (2007) 687.

[6] R. Wang, X. Cai, Hierarchical Structure, Disassortativity and Information Measures of the US Flight Network, Chin. Phys. Lett 22 (2005) 2715.

[7] J.Y. Ke, T. Gong and W.S.Y. Wang, Language Change and Social Networks, Commun. Comput. Phys 3 (2009) 935.

[8] R.Ferrer i Cancho, R.V. Sole, The small world of human language, Proc. R. Soc. Lond.B 268 (2001) 2261.

[9] X.L. Yu, Z.H. Li, D.M. Zhang, F. Liang, X.Y. Wang and X. Wu, The topology of an accelerated growth network, J. Phys. A: Math. Gen 39 (2006) 14343.

[10] Q. Liu, J.Q. Fang and Y. Li, Synchronization and Control of Halo-Chaos in Beam Transport Network with Small World Topology, Comman. Theor. Phys 47 (2007) 752.

[11] W.X. Wang, B.H. Wang, B. Hu, G. Yan, and Q. Ou, General Dynamics of Topology and Traffic on Weighted Technological Networks, Phys. Rev. Lett 94 (2005) 188702.

[12] L. Guo, and X. Cai, Continuous Opinion Dynamics in Complex Networks, Commun. Comput. Phys 5 (2009) 1045.

[13] Y.P. Yin, D.M. Zhang, G.J. Pan, M.H. He and J. Tan, Sandpile on scale-free networks with assortative mixing, Phys. Scr 76 (2007) 606.

[14] Y.Z. Zhou, J. Zhou and Z.H. Liu, Influence of network topology on the abnormal phase order, Europhys. Lett 84 (2008) 60001.

[15] X.B. Lu, X.F. Wang, J.Q. Fang, Topological transition features and synchronizability of a weighted hybrid preferential network, Physica A 371 (2006) 841.

[16] C. Castellano, S. Fortunato and V. Loreto, Statistical physics of social dynamics, Reviews of Modern Physics 81 (2009) 591.

[17] Z.H. Liu and B. Hu, Epidemic spreading in community networks, Europhys. Lett. 72 (2005) 315.

[18] W.X. Wang, B.H. Wang, C.Y. Yin, Y.B. Xie, T. Zhou, Traffic dynamics based on local routing protocol on a scale-free network, Phys. Rev. E 73 (2006) 026111.

[19] S.M.G. Caldeira, T.C. Petit Lobao, R.F.S. Andrade, A. Neme and J.G.V. Miranda , The network of concepts in written texts, Eur. Phys. J. B 49 (2006) 523.

[20] J.Y. Li, J. Zhou, Chinese character structure analysis based on complex networks, Physica A 380 (2007) 629.

[21] A.P. Masucci, G.J. Rodgers, Network properties of written human language, Phys. Rev. E 74 (2006) 026102.

[22] `www.cuiweiju.com, www.marxists.org.`

[23] M.E.J. Newman, Power laws, Pareto distributions and Zipf’s law, arXiv: cond-mat/0412004.

[24] Y.Z. Chen, D.R. He, A study on some urban bus transport networks, Physica A 376 (2007) 747.

[25] M.E.J. Newman, Assortative Mixing in Networks, Phys. Rev. Lett 89 (2002) 208701.
[26] D. Garlaschelli, M.I. Loffredo, Patterns of Link Reciprocity in Directed Networks, Phys. Rev. Lett 93 (2004) 268701.
[27] E. Ravasz, A.L. Barabasi, Hierarchical organization in complex networks, Phys. Rev. E 67 (2003) 026112.
[28] M.E.J. Newman, Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality, Phys. Rev. E 64 (2001) 016132.