Journal of Computers Vol. 27, No. 3, 2016, pp. 71-80
doi:10.3966/199115592016102703008

Empirical Analysis of Centrality Characteristics in Real Online Social Networks

Jun-Jun Cheng[1] Wei Cao[1] Hai-Qiang Chen[1]
Xin Zhou[1] Fei Xiong[2]

[1] China Information Technology Security Evaluation Center, Beijing, China
chengjj@itsec.gov.cn, caow@itsec.gov.cn, chenhq@itsec.gov.cn, zhouxin@itsec.gov.cn

[2] Beijing Jiaotong University, Beijing, China
xiongf@bjtu.edu.cn

Received 1 April 2015; Revised 15 April 2015; Accepted 24 April 2015

Abstract. In this paper, we analyzed topological characteristics of four famous centrality indexes (including degree, closeness, betweenness, and the k-core) and their correlations (including Pearson correlation and Kendall Rank correlation) in two real data sets. It's the fundamental work of identifying the influential nodes in complex networks. After simulations on two real data sets, we found that the distribution of degree, betweenness, and the k-core totally follow the power-law distribution. The Pearson correlation between degree and betweenness is the highest, however, the Kendall Rank Correlation between degree and k-core is relatively larger than values between degree and other two indexes.

Keywords: centrality indexes, correlation analysis, social networks

1 Introduction

Recent years, many scholars and researchers are concerned about the social network topology analysis, the relationship between the network topology and dynamics of behavior based on the various types of complex networks [1-5]. As we all know, a small part of the influential nodes for complex dynamical behavior of various types of networks (such as network cascade, information dissemination and network node synchronization, etc.) plays a very important role in the dynamic evolution [6-11]. Therefore, mining key nodes in the network has a certain theoretical research value.

In the computer science, the above problem is defined as the influence maximization, which is to find a small subset of nodes (k seed nodes) in a social network that could maximize the spread of influence based on a certain influence cascade model. D. Kempe et al [12] firstly defined this problem as the discrete formulation optimization problem, and studied three classic cascade model at the same time, which are independent cascade model, weighted cascade model and linear threshold model. Kempe et al. had proved the above optimization problem is NP-hard, and improved a greedy approximation algorithm. However, because of its low computational efficiency, it was not suited to large-scale social networks. To overcome these problems, many scholars had improved some new heuristic algorithm [13-16].

But in Interdisciplinary physics, many scholars pay more attentions on the problem of mining influential nodes base on certain topological statistics just like degree, closeness, betweenness and k-core (or k-shell). Some hybrid model for example, local weight index [17], influence factor [18], LeadRank [19], and so on, are also researched. However, the correlation between these four fundamental topological characteristics are still not very impressive. Based on this, our article will focus on some features of these four characteristics on two real social networks, and then give a clear understanding of their correlations.

2 Data Sets

The two data sets used in this chapter are both taken from real networks [20]. One comes from BlogCatalog, whose homepage will recommend some distinctive published blog and list the most popular and latest blog users. In this data set, each node corresponds to a blog user and the edge between two nodes represents a friend relationship between two blog users. Another data set comes from Delicious. This website is currently the world's largest bookmarking website. Four basic functions provided through the website: collect, tag, review and automate enable users to easily store, share and discover their favorite website links. The two real networks used in this paper are both undirected graphs. Degree distribution of two data sets are shown in Fig.1.

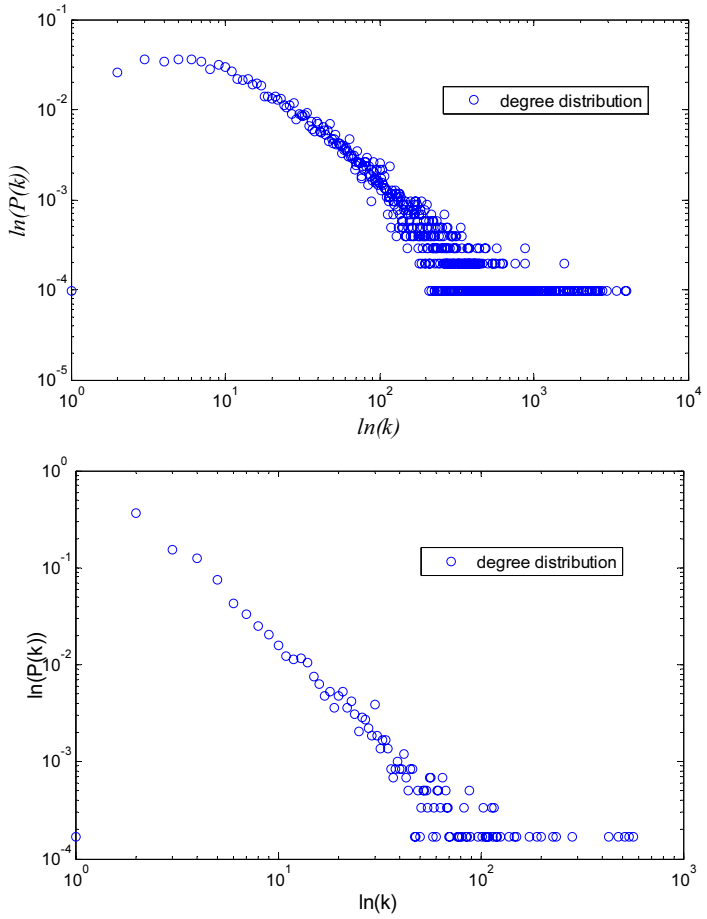

Fig. 1. Degree distribution of Blogcatalog (upper) and Delicious (lower)

It can be found that two real networks all have obvious characteristics of scale-free networks (the power exponents of the two networks can be approximately fitted as 2.2 and 1.9 via mathematical tools) that most users in the network have fewer friends while few users have more friends. The network diameter and clustering coefficient of Delicious are respectively 5 and 0.2128, indicating that it has obvious small-world characteristics. But compared with BlogCatalog, Delicious is much sparser.

3 Characteristics analysis

Degree centrality is usually used in the complex network, and degree distribution of two data sets are shown and analyzed in section 2, so we will not make any introduction of degree. In this section, we just focus on characteristics analysis of other indexes, such as closeness, betweenness, and k-core.

3.1 Closeness

In the complex network theory, closeness can also measure the nodes centrality. The closeness of node i $C_C(i)$ can be regarded as the reciprocal of geodesic distance sum of all the nodes. In order to simplify the calculation, the node closeness can be usually expressed by the following formula [21]：

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d_{ij}} \tag{1}$$

Among them, n is the total number of network nodes, and d_{ij} expresses the geodesic distance between nodes i and j. Under the condition of not considering the network edge weight, geodesic distance can be regarded as the minimum hop between nodes.

The probability distribution and cumulative distribution (CDF) of closeness in two data sets are shown in Fig.2. It can be seen that the closeness of the two data sets have the same distribution trend. Furthermore, we analyze the histograms of two data sets. Statistical results show that noses closeness of Delicious and BlogCatalog are respectively distributed in the interval [0.18, 0.48] and [0.26, 0.62] (see Fig.3). However, nearly 86.5% node closeness is concentrated between the interval (0.2, 0.3), and in BlogCatalog, nearly 69.8% node closeness is concentrated between the interval 0.4 and 0.5 and 28.5% node closeness is concentrated between the interval 0.3 and 0.4. All of this illustrate that, from the closeness aspect, more nodes are located in the relatively central position of the network.

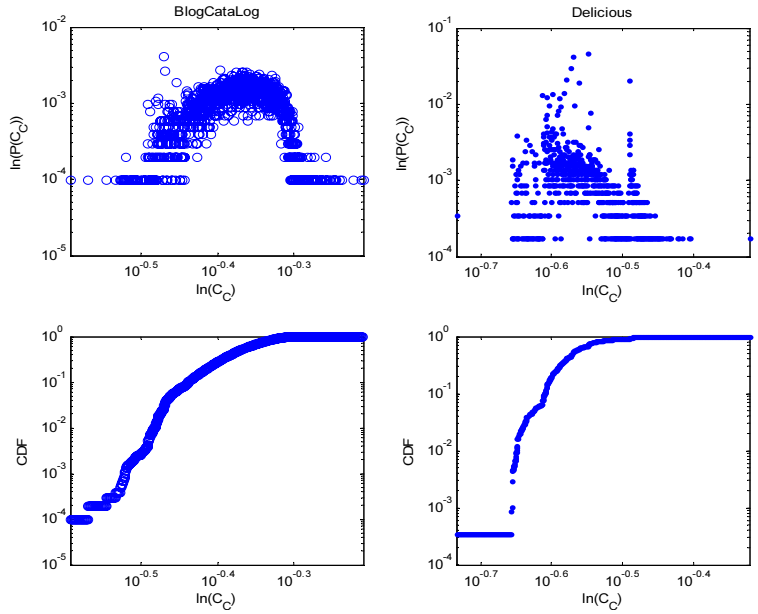

Fig. 2. Closeness distribution in two data sets

3.2 Betweenness

Betweenness is usually applied to evaluate the significance of nodes in the process of information spread and material transportation. The betweenness $C_B(i)$ of node i can be defined by the following formula [22]:

$$C_B(i) = \sum_{i \neq j \neq k \in V} \frac{g_{jk}(i)}{g_{jk}} \tag{2}$$

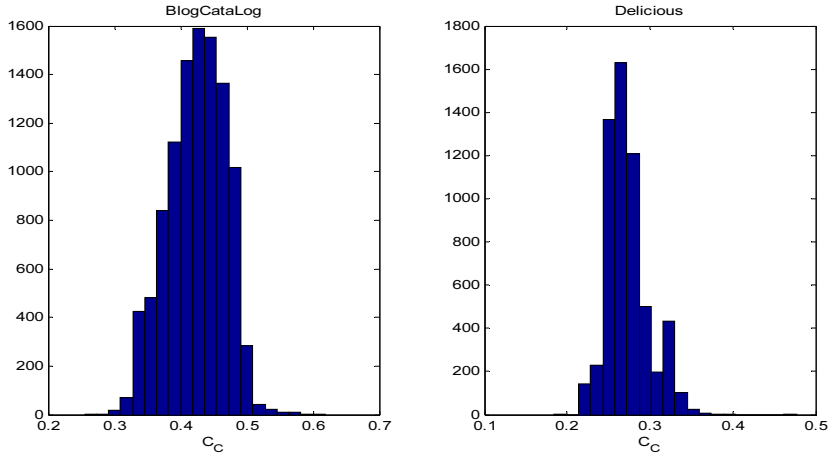

Fig. 3. Closeness histogram

Among them, $g_{jk}(i)$ signifies the number of shortest path between j and k which must go through node i. And $g_{jk}(i)$ is the number of all the shortest path between node j and node k. In order to improve calculation speed, betweenness can be solved by applying the following equation:

$$C_B(i) = \frac{2}{(n-1)(n-2)} \sum_{i \neq j \neq k \in V} g_{jk}(i) \tag{3}$$

The distribution and CDF of betweenness of two data sets are shown in Fig.4.

It can be seen from Fig.4 that betweenness centrality of the two data sets approximate to obey power-law distribution. The largest betweenness in BlogCatalog is 0.1186, and the node number whose betweenness are smaller than 0,001 accounts for 97.53%. The largest betweenness in Delicious is 0.7427, and the node number whose betweenness are smaller than 0.01 accounts for 98.58%. These statistical data strongly prove that the betweenness distribution in these two networks is extremely uneven, and it has strong heterogeneity.

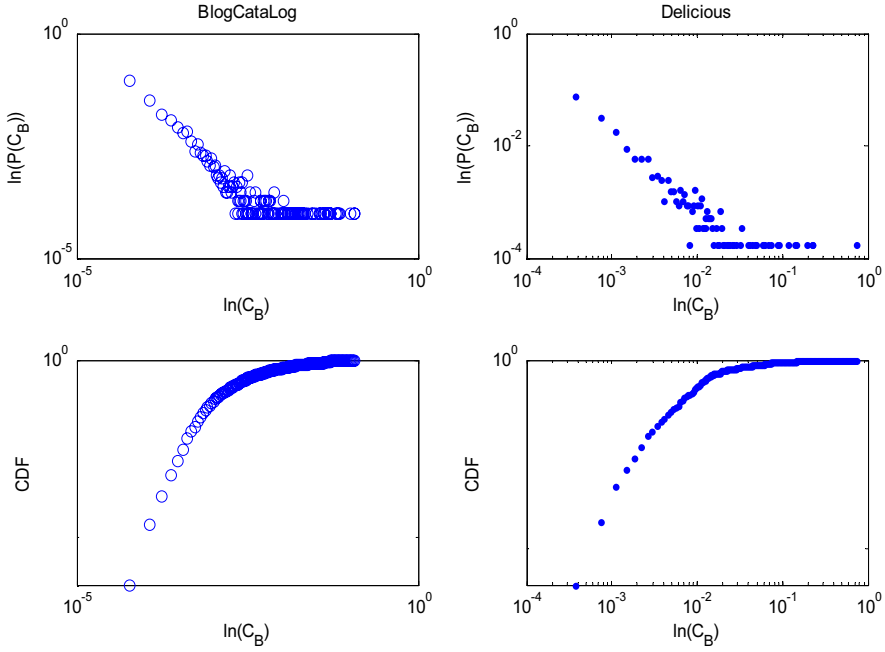

Fig. 4. Betweenness distribution in two data sets

3.3 k-core

Centrality index k-core can describe the node's position in the network. The larger the node k-core is, the closer the node is to the core place. The detailed calculation process of k-core can refer to the famous work [23]. The pseudo-code is shown in Fig.5.

```
1   init nodes list V;
2   init links list E;
3   init core=1;
4   while (|V| != 0)
5           while (node i ∈ V and degree (i) = core)
6                   delete node i and related links;
7                   update V and E;
8                   add removal node i into set V_k-core(core);
9           end while
10          core ++;
11  end while
12  reture V_k-core;
```

Fig. 5. The pseudo-code of k-core decomposition algorithm

After k-core decomposition, the network is divided into k layers (assuming that the maximum k-core is k). All nodes in each layer have the same k-core, but they are not necessarily to the same degree. Furthermore, two nodes with the same degree does not belong to the same layer. Fig.6 shows a sample of k-core decomposition, in which, node A and B belong to different k-core although they have same degree. Obviously, node B locates in the edge position, and node A is in the central area.

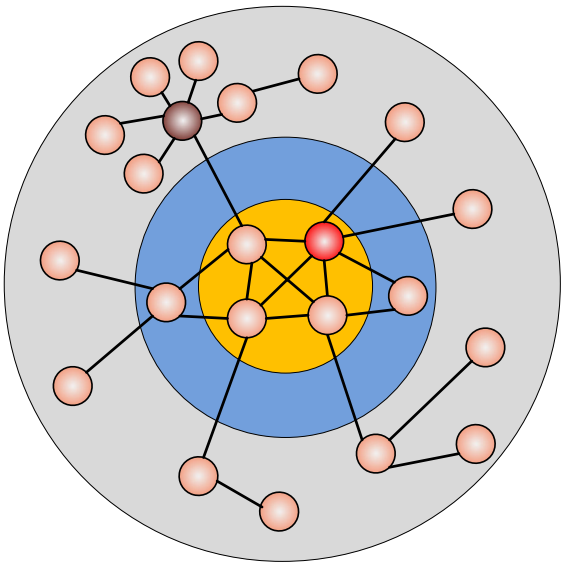

Fig. 6. Schematic diagram of k-core decomposition algorithm

In addition, by analyzing k-core distribution of the two data sets (as is shown in Fig.7, it can be known that each node k-core distribution also conform to the characteristics of power-law distribution, namely k-cores of most nodes in the network are relatively small (close to the edge of network), while a small amount of node k-cores are relatively large (located in the central position of the network). This kind of distribution way is extremely similar with the user distribution of real social network, namely many celebrities and opinion leaders are sought after by the majority of users as cores, while the vast majority of network users are located at the relatively marginal position in the network as followers.

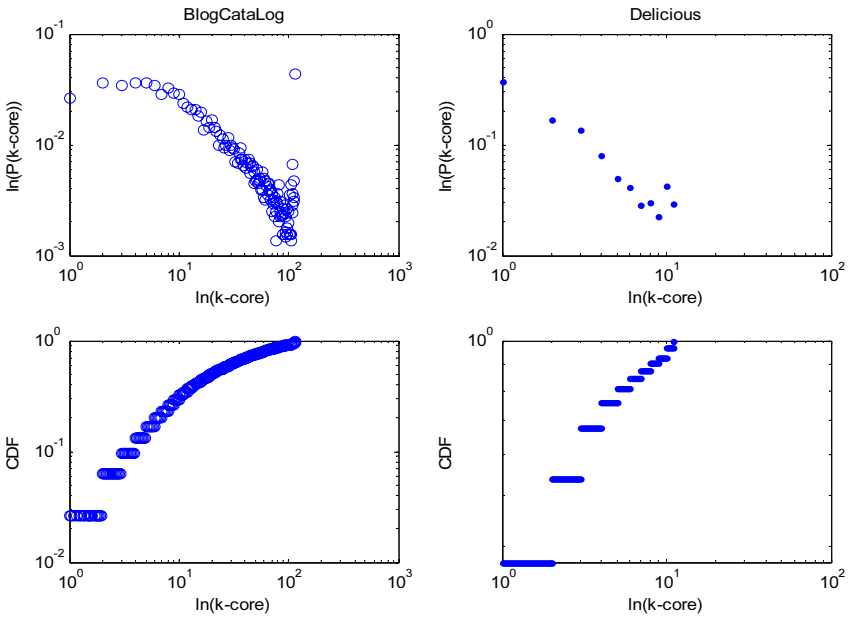

Fig. 7. K-core distribution of two real networks

4 Correlation Analysis

4.1 Pearson Correlation Analysis

Respectively calculate the Pearson Correlation Coefficient of four centrality indexes in BlogCatalog and Delicious, and specific results are shown in Table 1 and Table 2.

Table 1. The Pearson Correlation Coefficient of four centrality indexes in BlogCatalog

	Degree	Closeness	Betweenness	k-core
Degree	1	0.5065	0.8258	0.5760
Closeness	-	1	0.2292	0.8449
Betweenness	-	-	1	0.2010
k-core	-	-	-	1

Table 2. The Pearson Correlation Coefficient of four centrality indexes in Delicious

	Degree	Closeness	Betweenness	k-core
Degree	1	0.4214	0.7016	0.4783
Closeness	-	1	0.2558	0.5944
Betweenness	-	-	1	0.1675
k-core	-	-	-	1

We can find that the correlation between degree and betweenness is the highest, especially in the BlogCatalog where the Pearson correlation coefficient has arrived above 0.8. The weakest correlation is betweenness and k-core. In Delicious, the coefficient between betweenness and k-core is only 0.1675. Furthermore, degree has higher correlation with the other three centricity indicator, while betweenness has the weakest correlation with k-core. The correlation distribution between concentration of the two data sets and centrality indexes of the other three is shown in Fig.8.

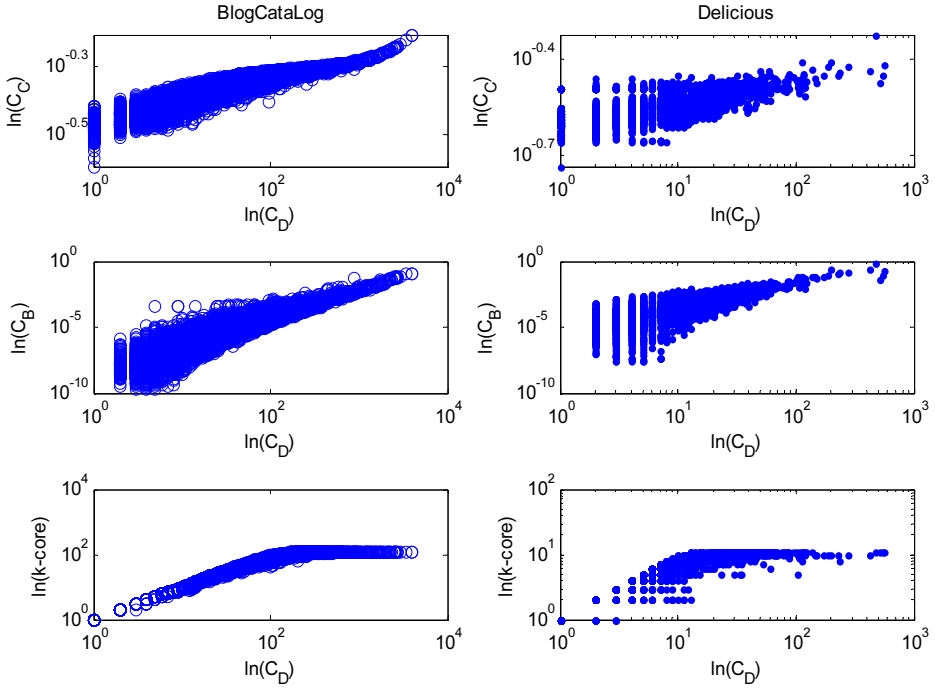

Fig. 8. The correlation of degree and the other three centricity indicators

In Fig.8, closeness, betweenness and k-core almost increase as the node degree increases. This trend is especially obvious in BlogCatalog. However, when the node degree is smaller, the differences among three centrality indexes of the same node are larger. In addition, there are 447 nodes in the kernel layer in BlogCatalog (k-core is 114), and all of these nodes' degree are distributed in the interval [174, 3992]. In the same way, the kernel lay in Delicious is 11, and it has 173 nodes whose degree are distributed in the interval [13 566]. This phenomenon shows that the k-core has relatively coarse distinguishing abilities.

4.2 Rank Correlation Analysis

Firstly, we enumerate top 10 nodes respectively in descending order based on degree, closeness, and betweenness centrality indexes in two data sets, and details are shown in Table 3. We should note that the number inside the brackets behind the node number is the corresponding k-core. It is known that the largest k-core in two data sets are respectively 114 and 11.

The consistency phenomenon can be easily found in the Table 3, especially between the degree and betweenness in BlogCatalog. It's a remarkable fact that top 10 nodes in each centrality indexes are not in the largest k-core layer. For example, a small amount of nodes come from the layer 10 and layer 8.

Table 3. Top 10 node number of four centricity indexes in descending orders in the two data sets

BlogCatalog			Delicious		
Degree	Closeness	Betweenness	Degree	Closeness	Betweenness
4839(114)	4839(114)	176(114)	320(11)	3(11)	3(11)
176(114)	176(114)	4839(114)	977(11)	1(10)	1(10)
4374(114)	4374(114)	4374(114)	3803(11)	636(11)	231(10)
8157(114)	8157(114)	8859(114)	3(11)	320(11)	320(11)
1226(114)	1226(114)	8157(114)	231(10)	1893(10)	354(10)
4997(114)	4984(114)	645(114)	354(10)	354(10)	198(8)
4984(114)	4997(114)	1226(114)	198(8)	269(11)	333(10)
8859(114)	8859(114)	7806(114)	333(10)	4046(11)	977(11)
645(114)	7098(114)	233(114)	1(10)	231(10)	1893(10)
446(114)	645(114)	446(114)	1893(10)	3441(10)	475(10)

Secondly, we analyze the Kendall Rank Correlation among closeness, betweenness, k-core and degree in the two data, and details are shown in Fig.9.

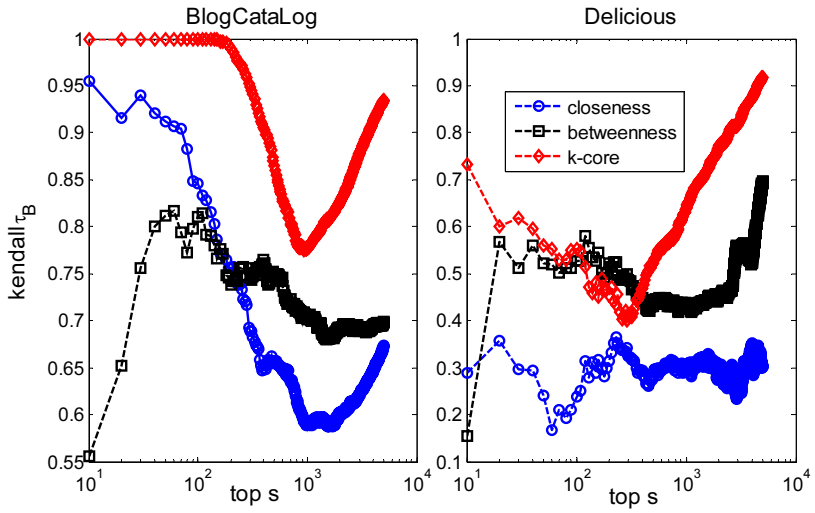

Fig. 9. Kendall Rank Correlation Coefficient among closeness, betweenness, k-core and degree

There exists a large rank correlation coefficient (positive correlation) between degree and k-core. Especially in the BlogCatalog, top 100 nodes in the degree ranking are approximately identical with top 100 nodes in the k-core ranking (Coefficient is close to 1), but there is not such high consistency in Delicious network. In addition, in two networks, the rank correlation between betweenness and degree violently increases as s increases, and when it reaches the critical value s_c, rank correlation will show the declining trend. Besides, the trends of rank correlation between closeness and node degree in these two networks are not the same. Totally speaking, the rank correlation coefficient between k-core and degree is relatively stable, betweenness ranks the second stable, and closeness is the most unstable.

5 Conclusion

In this paper, we just analyze four famous centrality indexes and their correlations in two real data sets. It's the fundamental work of identifying the influential nodes in complex networks. Based on the simulation of two real social networks, we found that the distribution of degree, betweenness, and the k-core totally follow the power-law distribution. The Pearson correlation between degree and betweenness is the highest, however, the Kendall Rank Correlation between degree and k-core is relatively larger than values between degree and other two indexes. From now on, we have an integral impression of the four centralities, next time we will focus on how to design a new index to identify influential nodes with lower complexity and higher precision.

Acknowledgement

This research was supported by the Fundamental Research Funds for the Central Universities under Grant No. 2014JBM018, the National Natural Science Foundation of China under Grant Nos. 61172072 and 61271308.

References

[1] Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47-97.

[2] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*(2), 167-256.

[3] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D. U. (2006). Complex networks: Structure and dynamics. *Physics Reports, 424*(4), 175-308.

[4] Zhang, J., Zhou, C. S., Xu, X. K., & Small, M. (2010). Mapping from structure to dynamics: A unified view of dynamical processes on networks. *Physical Reports E, 82*(2), 026116.

[5] Zeng, A., & Lü, L. Y. (2011). Coarse graining for synchronization in directed networks. *Physical Reports E, 83*(5), 056123.

[6] Motter, A. E., & Lai, Y. C. (2002). Cascade-based attacks on complex networks. *Physical Reports E, 66*(6), 065102.

[7] Zhou, T., & Wang, B. H. (2005). Catastrophes in scale-free networks. *Chinese Physics Letters, 22*(5), 1072-1075.

[8] Pastor-Satorras, R., & Vespignani, A. (2002). Immunization of complex networks. *Physical Reports E, 65*(3), 036104.

[9] Zhao, M., Zhou, T., Wang, B.-H., & Wang, W.-X. (2005). Enhanced synchronizability by structural perturbations. *Physical Reports E, 72*(5), 057102.

[10] Zemanová, L., Zhou, C., & Kurths, J. (2006). Cortical hubs form a module for multisensory integration on top of the hierarchy of cortical networks. *Front. Neuroinform, 4*(1).

[11] López, G. Z., Zhou, C. S., & Kurths, J. (2010). Cortical hubs form a module for multisensory integration on top of the hierarchy of cortical networks. *Front Neuroinform, 4*(1). Retrieved from http://journal.frontiersin.org/article/10.3389/ neuro.11.001.2010/full

[12] Kempe, D., Kleinberg, J. M., & Tardos, É. (2003, August). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 137-146). New York, NY: ACM.

[13] Kimura, M., & Saito, K. (2006). Tractable models for information diffusion in social networks. In J. Fürnkranz, T. Scheffer, & M. Spiliopoulou (Eds.), *Knowledge discovery in databases: PKDD 2006* (Vol. 4213, pp. 259-271). Berlin, Germany: Springer Berlin Heidelberg.

[14] Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., Van Briesen, J., & Glance, N. S. (2007, August). Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 420-429). New York, NY: ACM.

[15] Chen, W., Wang, Y., & Yang, S. (2009, June). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 199-208). New York, NY: ACM.

[16] Chen, W., Wang, C., & Wang, Y. ((2010, July). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1029-1038). New York, NY: ACM.

[17] Chen, D., Shang, L. L. M.-S., Zhang, Y.-C., & Zhou, T. (2012). Identifying influential nodes in complex networks. *Physica A: Statistical Mechanics and Its Applications, 391*(4), 1777-1787.

[18] Huang, X., Vodenska, I., Wang, F., Havlin, S., & Stanley, H. E. (2011). Identifying influential directors in the United States corporate governance network. *Physical Review E, 84*(4), 046101.

[19] Lü, L. Y., Zhang, Y. C., Yeung, C. H., & Zhou, T. (2011). Leaders in social networks, the delicious case. *PLoS ONE, 6*(6), e21202.

[20] Zafarani, R., & Liu, H. (2009). *Social computing data repository at ASU* [http://socialcomputing.asu.edu]. Tempe, AZ: Arizona State University, School of Computing, Informatics and Decision Systems Engineering.

[21] He, D. R., Liu, Z. H., & Wang, B. H. (2009). *Complex systems and complex networks*. Beijing, China: Higher Education

Press.

[22] Freeman, L.C. (1997). A set of measures of centrality based on betweenness. *Sociometry, 40*(1), 35-41.

[23] Carmi, S., Havlin, S, Kirkpatrick, S., Shavitt, Y. & Shir, E. (2007). A model of Internet topology using k-shell decomposition. *Proc. Natl Acad. Sci. USA, 104*(27), 11150-11154.