Statistical Characteristics of an Evolving Co-citation Network: The Distribution of Betweenness Centrality

Deming Lin[1], Chaomei Chen[2] and Zeyuan Liu[3]

[1]*lin.deming@163.com*
WISE Lab, Dalian University of Technology, Dalian (China)

[2]*chaomei.chen@cis.drexel.edu*
WISE Lab, Dalian University of Technology & College of Information Science and Technology
Drexel University, Philadelphia (USA)

[3]*liuzy@vip.163.com*
WISE Lab, Dalian University of Technology, Dalian (China)

Abstract
We investigate statistical characteristics of an evolving co-citation network, primarily in term of the dynamics of betweenness centrality measures, we generate co-citation network of papers published in journal of Scientometrics. Our study shows that the overall co-citation network is a small-world and scale-free network. The co-citation network has a relatively small number of nodes with high betweenness centrality, most nodes have low betweeness centrality scores. Furthermore, the betweenness centrality distribution of the co-citation network follows segmented Zipf-Pareto distribution. We found a tendency that high betweenness centrality measures tend to reduce over time.

Introduction
Co-citation network is a network mapping of co-citation relation, and it is a powerful tool for co-citation analysis. Small (1973) did not only introduce the concept of co-citation originally, but also drew the first co-citation network from particle physics. Co-citation network could be used to characterize many research fields (Usdiken, 1995, Braam, Moed & Van Raan, 1999, Chen C., 1999, Gmur, 2003, Reid & Chen H. C., 2007, Vargas-Quesada & Moya-Anegon, et al., 2008). Along with the development of co-citation analysis, social networks and complex networks, co-citation networks are advancing rapidly (Otte & Rouseau, 2002, Egghe & Rousseau, 2002, 2003, White, 2003, Schildt & Mattsson, 2006, Quirin & Cordon, et al., 2008). But there aren't many articles that explore and discuss the statistical characteristics of co-citation networks.

In a social network, betweenness centrality is a useful indicator to measure the role of nodes in network (Freeman, 1977). In co-citation network, betweenness centrality could examine the citation environments of scientific journals as an indicator of the interdisciplinarity of scientific journal (Leydesdorff, 2007), also measure the position of component of scientific literature such as citations, authors (Chen C., 2006). A natural impulse is to survey the statistical characteristics of betweenness centrality of co-citation network.
Many indicators for scientometrics such as citation, citation rate (Braun, Glanzel & Schubert, 1985) follow Zipf-Pareto distribution (Liming L., 1993) that the plot between indicator y and its rank r are consistent with power law curve:

$$y = \frac{C}{r^{\alpha}} \quad (C \text{ is a constant, } \alpha > 0)$$

Whether the betweenness centrality of co-citaion network follows Zipf-Pareto distribution or Zipf-Mandelbrot distribution is questioned. Zipf-Mandelbrot distribution is given by:

$$y = \frac{C}{(r+m)^{\alpha}} \quad (C \text{ is a constant, } \alpha > 0)$$

Proceedings of ISSI 2009, pages 552–560.
Edited by B. Larsen and J. Leta.

Where m is a positive parameter; if $m = 0$, that is Zipf-Pareto distribution. So we select the articles published in *Scientometrics* to vestigate the distribution of beteweenness centrality of co-citation network.

Co-citation network

The betweenness centrality of a node i is a measure giving the probability that the node will occur on a shortest path between two arbitrary nodes in the given network, where the shortest path between tow arbitrary nodes is a path such that the sum of the amount of its constituent edges is minimized. Namely,

$$B_i = \sum_{i,j} \frac{\sum_{l \in S_{ij}} \delta_l^i}{\left| S_{ij} \right|}$$

Where S_{ij} ($\forall i, j \in N$) are all the shortest paths of network, $|S_{ij}|$ is the amount of S_{ij}, δ_l^i is the amount of the shortest paths crossing node i in $\{S_{ij}\}$.

In co-citation network, ranking the betweenness centrality can express the importance degree of citations. The citations with high betweenness centrality always appear more frequently in gather of the shortest paths, and more citations formed the links of co-citation through it. Consequently, the high betweenness centrality citations have more influence in the co-citation network.

Selecting *all* 1262 *articles* published in *Scientomatrics* from 1998 to 2008 *(November)*, we establish a co-citation network using the Network Workbench Tool (http://nwb.slis.indiana.edu/). The network has 18237 nodes (citations), 515758 edges (links of co-citation) and 709 isolated nodes (isolated citation without link of co-citation between other citations).

Tabel 1. Details of the co-citation network

Nodes: N	**Edges:** M	**Average Degree:** $< k >$	**Average Distance:** L	**Average Cluster Coefficient:** C
18237	515758	56.5617	3.1143	0.8617

The details of the network (table 1.) show that the network has high average cluster coefficient and low average distance. Two arbitrary nodes in the network could be contacted via only three links. So the network is a small-world network. Moreover the average degree of the network is 56.5617, the degree distribution (figure 1) suggests that co-citation network is a truncated scale-free network similar to the network of movie actors in the paper of Amaral & Scala, et al., 2000. Anyway, the co-citation network is a small-world and scale-free network.

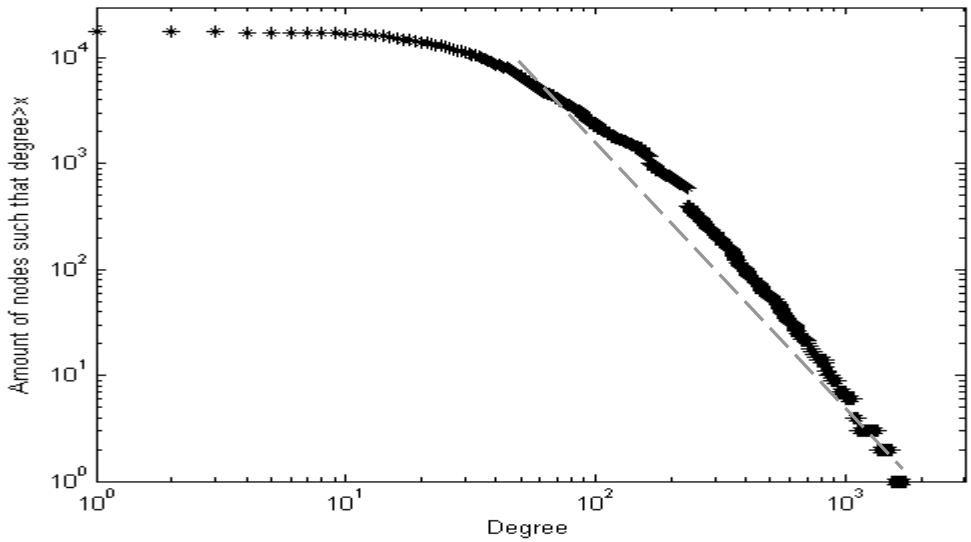

Figure 1.The distribution of degree of the co-citation network: the log-log plot of the degree distribution for the co-citation network is consistent with power law decay for the values of degree more than 40. Moreover it is truncated power law decay as many other realistic network (Amaral & Scala, et al., 2000).

Tabel 2. Top 20 of high betweenness centrality

Rank of Betweenness centrality	Citation	Cited frequency	Rank of cited frequency	Source
1	Price DJD, 1963, Little science, big science	65	1	BOOK
2	Garfield E, 1979, Citation indexing	60	3	BOOK
3	Egghe L, 1990, Introduction to Informetrics	65	2	BOOK
4	Lotka AJ, 1926, The frequency distribution of scientific productivity	46	4	J WASHINGTON ACADEMY
5	Merton RK, 1968, The Matthew Effect in Science	38	7	SCIENCE
6	Callon M, 1986, Mapping the dynamics of science and technology	31	16	BOOK
7	Gibbons M, 1994, The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies	37	8	BOOK
8	Rogers EM, 1995, Diffusion of Innovations	8	263	BOOK
9	Schubert A, 1989, Scientometric datafiles	40	5	SCIENTOMETRICS

10	Small H, 1973, Co-citation in scientific literature: A new measure of the relationship between publications	36	10	J AM SOC INFORM SCI
11	Wasserman S, 1994, Social Network Analysis: Methods and Applications	23	28	BOOK
12	Katz JS, 1997, What is research collaboration	37	9	RES POLICY
13	Garfield E, 1972, Citation analysis as a tool in journal evaluation	25	22	SCIENCE
14	Moed HF, 1995, New bibliometric tools for the assessment of national research performance	32	15	SCIENTOMETRICS
15	Etzkowitz H, 2000, The dynamics of innovation	22	32	RES POLICY
16	Griliches Z, 1990, Patent Statistics as Economic Indicators: A Survey	22	33	J ECON LIT
17	Narin F, 1997, The increasing linkage between US technology and public science	40	6	RES POLICY
18	May RM, 1997, The Scientific Wealth of Nations	36	11	SCIENCE
19	Rousseau R, 1997, Sitations: an exploratory study	25	23	CYBERMETRICS
20	Price DJD, 1965, Networks of Scientific Papers	26	20	SCIENCE

There are 7 books in the table. The citations ranked in 1-3 are all books. There are 4 citations published in *Science*, 3 citations published in *Research Policy* and 2 citations published in *Scientometrics*. Therefore, the more citations of higher betweenness centrality are classic books.

The average cited frequency of top 20 citations in table 2 is 35.7. (The average cited frequency of top 20 of high cited citations is 39). The average cited frequency of top 100 citations of high betweenness centrality is 10.12. (The average cited frequency of top 100 of high cited citations is 11.11). Moreover, there are 11 citations ranked in top 11 for cited frequency in the table 1. So the high betweenness centrality citations always have high cited frequency.

The citations whose betweenness centrality ranked in 1-4 are also ranked in 1-4 for cited frequency. We draw plots of the betweennss centrality of these four highly ranked citations distributed over time between 1998 and 2008.

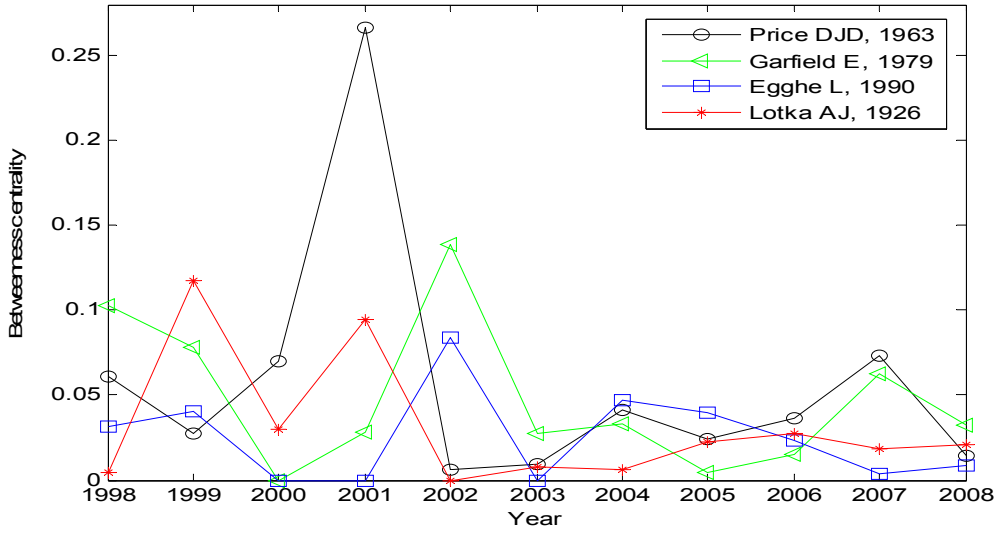

Figure 2. The distribution of betweenness centrality measures over time. Plots show that the peak values of the four citations are distributed in 1999-2002. The betweeness centralities of the four citations become smaller after 2002. This suggests that the literatures published in Scientometrics are becoming more multifold.

Distribution of the betweenness centrality

If the betweenness centrality (bc) follows Zipf-Mandelbrot distribution with the rank (r), then

$$\log bc = \log C(r+m)^{-\alpha} = \log C - \alpha \log(r+m)$$

Where C and m are constants, $\alpha > 0$; if $m = 0$, that is Zipf-Pareto distribution. After defining $a = -\alpha$, $b = \log C$, $x = \log(r+m)$, $y = \log bc$, the log-log curve of betweenness centrality should be linear, namely $y = ax + b$. Figure 3 is the log-log curve of betweenness centrality ranked decreased power when $m = 0$, where abscissa is the log of r, ordinate is the log of bc. Because the plot in figure 3 is linear just for all ranks, it rather fits Zipf-Pareto distribution.

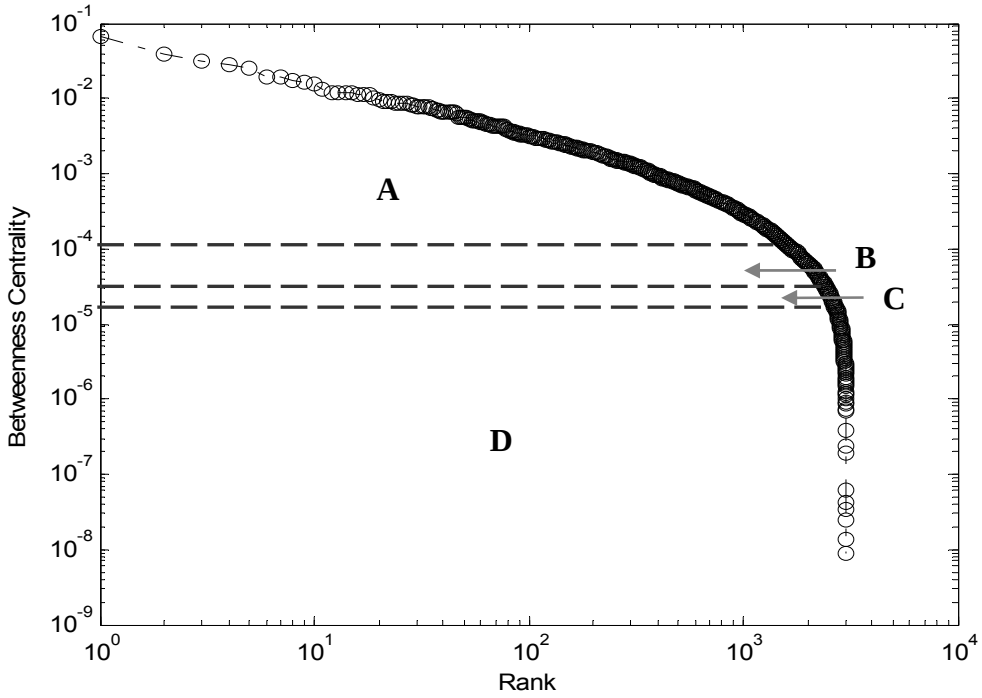

Figure 3. The log-log plot between rank and betweenness centrality of the co-citation network when *m*=0. The plot is consistent with power law line segmented four slices (in table 3). This shows that the curve of betweenness centrality of the co-citation network accord with a segmented Zipf-Pareto distribution intuitively.

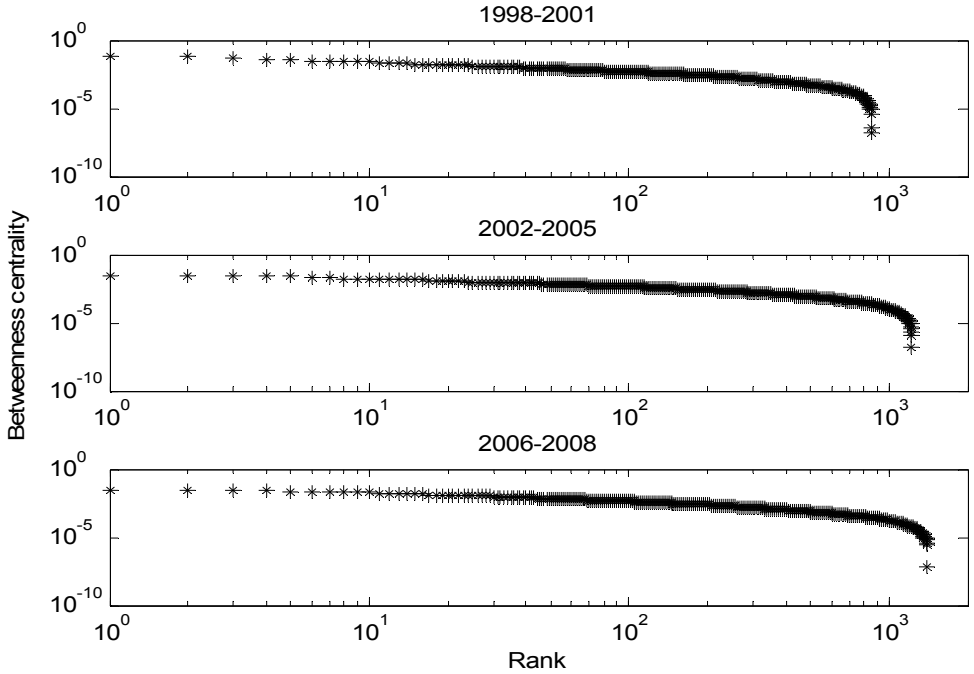

Figure 4. The log-log plot between rank and betweenness centrality of the co-citation network divided into three time slices of 1998-2001, 2002-2005 and 2006-2008. The plots of betweenness centrality all follow segmented Zipf-Pareto decay being similar to the plot of figure 3 intuitively.

Tabel 3. Slices of Zipf-Pareto distribution of the plot of figure 3.

Slice	A	B	C	D
Range	1-1685	1686-2961	2962-3022	3023-3038
$-\alpha$	0.30217	11.37	213.48	3604.4
$\log C$	-1.0848	-4.7179	-63.026	-1037.2
Confidence: σ	0.95	0.9216	0.9501	0.9171
Value	(9.96-6740)x10^{-5}	(9.95-0.42)x10^{-5}	(1.03-4.19)x10^{-6}	(0.88-101.7)x10^{-8}

In the Tabel 3, all the four confidence levels are more than 0.9. The spots of higher value of betweenness centrality are included in slice A, their descending trend is smoother, and slope of the plot is 0.30217. Then the decay is more and more rapid after the rank more than 1685. Entering slice D, the slope of plot has already reached 3604.4. Therefore, the value of high betweenness centrality node is concentrated, the value of low betweenness centrality node is dispersed. Further, defining the value range of betweenness centrality as from 10^{-i} to 10^{-i+1} （ $i=1,2,\cdots,5$） , such as $10^{-2} \le bc < 10^{-1}$ when $i=1$.

Tabel 4. Distribution of betweenness centrality of the co-citation network

Range of value	$10^{-2}-10^{-1}$	$10^{-3}-10^{-2}$	$10^{-4}-10^{-3}$	$10^{-5}-10^{-4}$	$0-10^{-5}$
Amount in range	19	367	1388	1424	15039
Proportion	0.001	0.0201	0.0761	0.0781	0.8246

The table shows that the majority of nodes in the co-citation network have relatively low betweenness centrality scores. Nineteen nodes in the co-citation network have betweenness centrality values in the range of 10^{-2}~10^{-1}. The proportion of the nodes that values of betweenness centrality are more than 0.01 in the co-citation network is only 0.1%. Therefore, there is Matthew effect in the betweenness centrality of co-citation network being similar to other scientometrics indicators.

Conclusions

In conclusion, the statistical characteristics of betweenness centrality distribution in a co-citation network have promising and encouraging results. The major finding is that

- the co-citation network will be a small-world, scale-free network when the amount of its nodes is large;
- the proportion of classic books in higher betweenness centrality nodes is always high;
- the cited frequency of high betweenness centrality nodes in co-citation network is large, high betweenness centrality nodes are cited more frequently than other nodes due to their intellectual impact;
- the value of betweenness centrality of co-citation network follow segmented Zipf-Pareto distribution intuitively;
- high betweenness centrality nodes form a small portion of the nodes in the network, but more nodes of low betweenness centrality;

Betweeness centrality as an indicator of scientometrics could mine the important literature of any research field and evaluate the influence of literature, author, organization or country. Finding the statistical characteristics of betweenness centrality is meaningful. Combining the

network analysis with the statistical analysis, locating the citation in co-citation network and plot of distribution together, the co-citation network would be improved further. Besides, challenges and opportunities also are identified in this article for future work. In summary, the major challenges are to

- exploring the statistical characteristics of other indicators of co-citation network;
- testing the influences of normalizations and "lower cut-offs" of references in the articles for the statistical characteristics of co-citation network;
- perfecting the betweenness centrality as a indicator for scientometrics;
- proposing a better algorithm to identify the high betweenness centrality when the data being massive.

Acknowledgements

This research is supported by China Postdoctoral Science Foundation funded project under Grant 20080441114, National Natural Science Foundation of China under Grant 70773015, 70431001, 70620140115, National Social Science Foundation under Grant 07CTQ008, 08BPQ025, Research Fund for the Doctoral Program of Higher Education of China under Grant 20070141059.

References

Amaral, L. A. N., Scala, A., Barthelemy, M., et al. (2000). Classes of small-world networks. *PNAS*, 97, 11149-11152.

Braam, R. R., Moed, H. F., Van Raan, A. F. J. (1999). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science*, 42, 233-251.

Braam, R. R., Moed, H. F., Van Raan, A. F. J. (1999). Mapping of science by combined co-citation and word analysis. II: Dynamical aspects. *Journal of the American Society for Information Science*, 42, 252-266.

Braun, T., Glanzel, W., Schubert, A. (1985). *Scietometrics indicators*. Singapore: World Scientific Publishing Co.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management*, 35, 401-420.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57, 359-377.

Egghe, L. & Rousseau, R. (2002). Co-citation, bibliographic coupling and a characterization of lattice citation networks. *Scietometrics*, 55, 349-361.

Egghe, L. & Rousseau, R. (2003). A measure for cohesion of weighed networks. *Journal of the American Society for Information Science and Technology*, 54, 193-202.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40, 35-41.

Gmur, M. (2003). Co-citation analysis and the search for invisible colleges: a methodological evaluation. *Scietometrics*, 58, 27-57.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58, 1303-1309.

Liming, L. & Lihua, L. (1993). Scientific publication activities of 32 countries – Zipf-Pareto distribution. *Scientometrics*, 26, 263-273.

Otte, E. & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information science*, 28, 441-453.

Quirin, A., Cordon, O., Santamaria, J., et al. (2008). A new variant of the Pathfinder algorithm to generate large visual science maps in cubic time. *Information processing and management*, 44, 1611-1623.

Reid, E. F. & Chen, H. C. (2007). Mapping the contemporary terrorism research domain. *International journal of human-cumputer studies*, 65, 42-56.

Schildt, H. A. & Mattsson, J. T. (2006). A dense network sub-grouping algorithm for co-citation analysis and its implementation in the software tool Sitkis. *Scientometrics*, 67, 143-163.

Small, H. (1973). Co-citation in scientific literature: A new measure of the relationship between publications. *Journal of the America Society of Information Science*, 24, 265-269.

Usdiken, B. (1995). Organizational analysis in North America and Europe: a comparison of co-citation networks. *Organization Studies*, 16, 503-526.

Vargas-Quesada, B. De Moya-Anegon, F., Chinchilla-Rodriguez, Z., et al. (2008). Evolución de la estructura científica española: ISI web of science 1990-2005. *EI Profesional de la informacion*, 17, 22-37.

White, H. D. (2003). Pathfinder networks and author cocitation analysis: a remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology*, 54, 423-434.