

Motif-based Classification in Journal Citation Networks

Wenchen Wu¹, Yanni Han¹, Deyi Li²

¹(State Key Lab of Software Development Environment, Beihang University, Beijing, 100083, China), ²(Institute of Electronic System Engineering, Beijing, 100039, China)

Email: wuws@nlsde.buaa.edu.cn, ziqinli@public2.bta.net.cn, libra_hyn@sina.com

Received November 16th, 2008; revised November 20th, 2008; accepted November 27th, 2008.

ABSTRACT

Journals and their citation relations are abstracted into journal citation networks, basing on CSTPC journal database from year 2003 to 2006. The network shows some typical characteristics from complex networks. This paper presents the idea of using motifs, subgraphs with higher occurrence in real network than in random ones, to discover two different citation patterns in journal communities. And a further investigation is addressed on both motif granularity and node centrality to figure out some reasons on the differences between two kinds of communities in journal citation network.

Keywords: Motif, Classification, Journal Citation Networks

1. Introduction

As an effective method, complex networks have been widely used to describe many complicated real world systems. It can be regarded as the topology abstract of many real complex systems, whose structure do not rely on node position or edge form, but with two essential attributes-small-world [1] and scale-free [2].

Though many networks present common global characteristics, they could have entirely different local structures. Recent researches indicate that network motifs, interconnected patterns occurring in numbers that are significantly higher than those in identical randomized networks, may be the “simple building blocks” in complex networks [3]. The concept and applications of motifs are first appeared in biological field. They present in biological systems as characteristic modules to carry out some certain kind of functions. For example, the same motifs, defined as feed-forward loops, have been found in organisms from bacteria and yeast [4], to plants and animals [5,6]. This kind of motifs plays an important role of persistence detectors, or pulse generator and response accelerators. These kinds of research results always make some direct biology meanings [7]. Besides, with certain iterations, many small, highly connected topologic motifs could combine in a hierarchical manner into larger but less cohesive modules [8].

Ron Milo proposed two concepts in 2002, which are shown below to find motifs in networks. And then they gave the concept of “superfamilies” in 2004 and also the significance profile (SP) method to compare the local structure between different kinds of complex networks [9]. The result shows that networks from different fields can share similar characteristic of local structures.

- Z-Score, valuing the statistic importance of each

network motifs

$$Z_i = \frac{N_{real_i} - \langle N_{rand_i} \rangle}{std(N_{rand_i})}$$

- P-Value means the probability of network motifs appearing in a randomized network an equal or greater number of times than in the real network.

Milo and his fellows also published the motif detection software, named MFinder in the homepage of Uri Alon lab. In MFinder, the subgraphs need to satisfy the default settings to make themselves network motifs, in which their Z-Score should bigger than 2, and P-Value should less than 0.05. Figure 1 shows a motif detection in real network and random network respectively by Ron Milo.

The reminder of this paper is organized as follows. Section 2 outlines the construction and essential attributes of journal citation networks. Section 3 presents the degree analysis of the networks. Section 4 analyses the motif structures and citation patterns in journal communities, and Section 5 concludes the whole work and discusses future research directions.

2. Construction Principles and Attribute Analysis

This article obtains the original data from a project led by the Institute of Scientific and Technical Information of China in 2004 [10]. The project formed both the citing and cited matrixes of each journal, which is embodied in China Scientific and Technical Papers and Citations (CSTPC) database from year 2003 to 2006 respectively. The journal citation networks are constructed according to those matrixes.

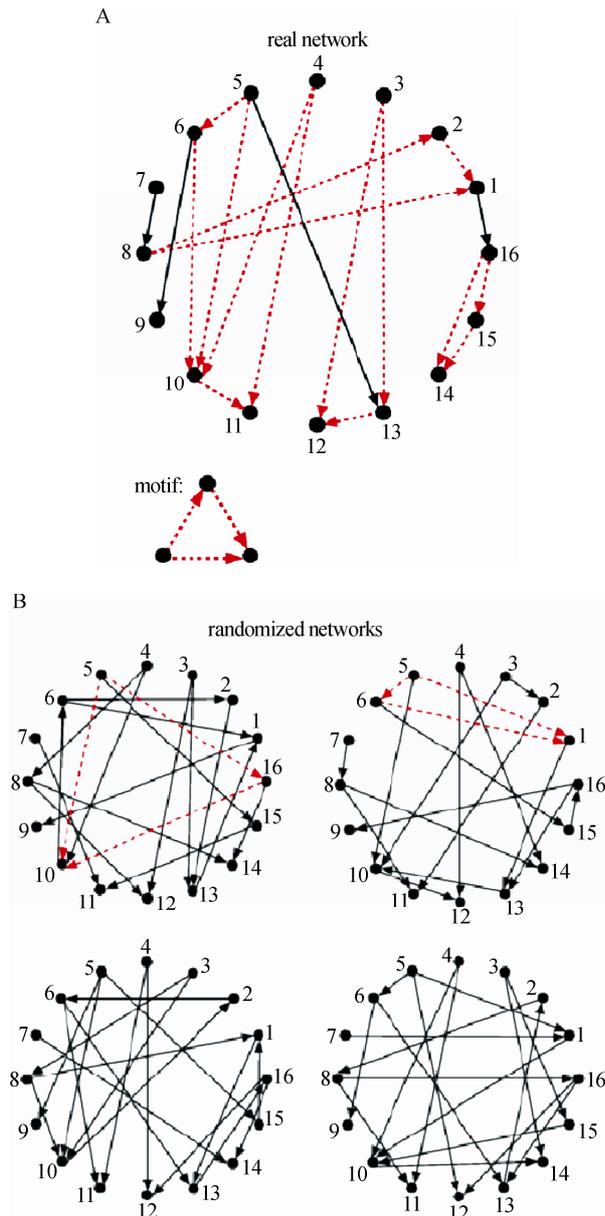


Figure 1. Motif detection in networks

In this paper, we define a journal citation network as follows: each journal expresses as a node of network, if one journal cites others, an edge is added beginning with

this journal and ending with the journal it cites. Otherwise, if one journal is cited by other ones, then add an edge beginning with the journal which cites it and ending with this journal. After sorting one-year journals this way, a directed journal citation network can finally be formed.

This article presents many essential attributes in journal citation networks of these four years. For instance, network connectivity, network diameter, average path length and also average clustering coefficient. Table 1 shows some fundamental statistic information. It can be found that network scale grows steadily from the year 2003 to 2006, except a sharply edge decrease in the year 2005. According to a further investigation, this phenomenon has something to do with a limited threshold in the original datasets, which is set up to filtrate the noise data. The average degree can explain average citation times between journals. It shows that the citation is more positive in 2006 than other years. Meanwhile, network diameters are no bigger than six in the year 2003, 2004 and 2006, and these networks also have big average clustering coefficient, which indicate typical small-world characteristic commonly in complex networks.

3. Degree Analysis

Degree is a simple but important definition to describe node attributes, which can reflect some network characteristics intuitively. When it comes to directed journal networks, a node's outdegree is the number of journals it cites, and its indegree is the number of journals citing it. Figure 2 shows the indegree and outdegree distributions of these four years. It is obvious that the nodes whose indegree or outdegree are bigger than 10, are in accordance with power-law distribution, which means a typical scale-free characteristic. Most nodes of small degrees have few cited or citing relations, but in contrary, large quantities of citation relations are held in only a few nodes. Particularly, though some nodes with really small degrees are not accordance with power-law distribution, their citations are totally rare when comparing with the entire network scale. To some extent, this kind of journals is the so-called fringe journals, and it does not play a vital part on the distribution characteristic of globe network.

Table 1. The statistical data of fundamental attributes

| | 2003 | 2004 | 2005 | 2006 |
|------------------------------------|----------|----------|----------|----------|
| node number | 1577 | 1659 | 1658 | 1787 |
| link number | 32823 | 42909 | 25923 | 47470 |
| ratio L/N | 20.81357 | 25.86438 | 15.6351 | 26.56407 |
| maximum indegree | 211 | 264 | 255 | 288 |
| maximum outdegree | 98 | 84 | 124 | 245 |
| average degree | 36.27774 | 44.51718 | 27.22799 | 47.6911 |
| network diameter | 5 | 6 | 8 | 6 |
| average path length(reachable) | 3.47 | 3.242 | 4.073 | 3.782 |
| average path length(bidirectional) | 2.709 | 2.616 | 2.969 | 2.642 |
| average clustering coefficient | 0.238 | 0.246 | 0.269 | 0.302 |

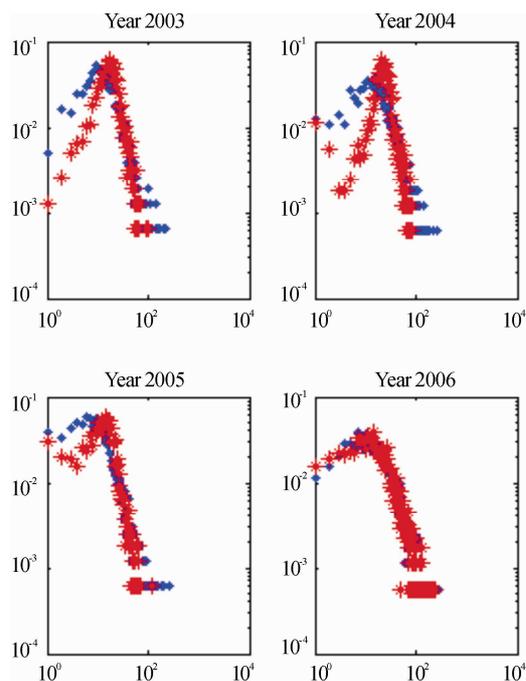


Figure 2. The indegree and outdegree distribution

In order to give a further look on the degree characteristic of these journals, Figure 3 shows the four-year correlations between indegrees and outdegrees in journal networks, in which each point corresponds to a node, and the x-position is determined by the node's indegree, the y-position corresponds to its outdegree. We can find that most nodes in journal network have significant distances with indegree and outdegree value. They are either with larger indegree but smaller outdegree, or vice versa. Only few nodes have both large indegree and outdegree. This characteristic is especially obvious in the year 2006.

It presents that nodes with large indegree have a good opportunity to be retrieved by SCI or EI, such typical examples including Chinese Science Bulletin, Chinese Journal of Computers, and etc. This kind of journals usually has a great influence in domestic journals of the same kind, which lead to a positive citations to them. But when it refers to their comparatively smaller outdegree, we believe it has a strong probability these influential journals prefer to make citations with those international journals. What have analyzed above indicates one citation characteristic of Chinese journal networks, that is journals retrieved by SCI/EI generally have a highly inclination to be cited, but with low positivity to cite other non-core journals in contrast.

4. Motif Structure in Journal Citation Networks

4.1 Motif Structure in Communities

It has been mentioned in the previous article that it is important to research on network local topology structure and generate mechanisms. In recent years, people find a clustering characteristic in complex networks [11-16].

Newman proposed the concept of community structure to indicate that an entire network is comprised of some communities or clusters. Nodes are joined together in tightly-knit groups, between which there are only looser connections. The community structure reflects high clustering and modularized characteristics. Many real networks, such as biology network, WWW network and social network have all been proved had obvious community structures. This article makes an analysis on 2004 journal citation network, and also finds the typical community structure in this network.

For the category differences, the citation times between different kinds of journals are extremely different. For example, there are only no more than ten times citations between class of physic and traffic, but thousands of times citation between all kinds of medical journals, such as pharmacy, clinical medicine and traditional Chinese medicine, etc. In principle, tight citation correlations make journals assemble in the same community, while loose citation correlations make journals separate into two communities. Through the designed experiment, journals of the same category or several similar categories generally appear in the same community with the partition of the whole journal network into twenty different communities in all.

In the following work, this article analyzes the different citation relations between those different communities. Motif kind presents in an exploding way with the increase of node number. For example, there are 13 kinds of motif with three nodes, while the number of kind rises to 199 with four-node motifs. Since journal network belongs to sociology field, and it is found that social networks are more likely to contain triangle relations. Therefore, in this paper, the research is mainly outspread on the granularity of three-node motifs.

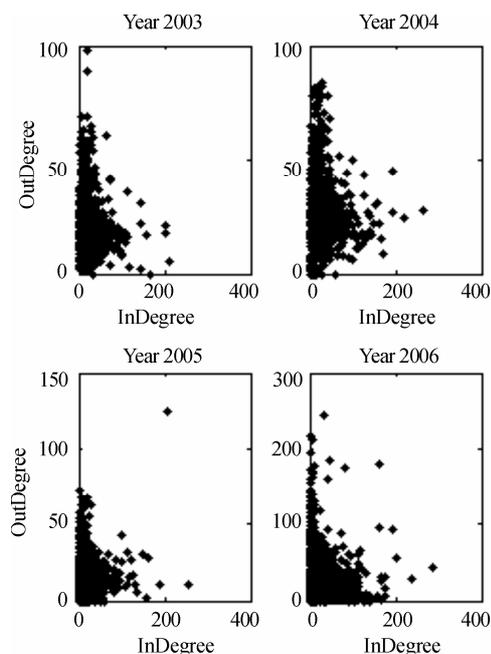


Figure 3. The indegree and outdegree correlations

According to the concrete meanings in journal citation networks, these thirteen motifs are classified into two kinds: one named “unidirectional citation clusters”, comprising with motif ID36, ID12, ID6, ID38 and ID140. This kind of motifs have a common characteristic, that is none of them contains any bidirectional edges, which means any pairs of nodes in these three-node motifs have no mutual citation relationships. To be contrast, the other kind named “mutual citation clusters”, with the motif members of ID164, ID14, ID78, ID166, ID174, ID46, ID102 and ID238, in which there could be one or even more bidirectional edges. In other words, it has at least one mutual citation relationship between the three nodes.

It indicates in Figure 3 the unidirectional citation clusters play an absolutely dominant part in journal networks, proving Chinese journal networks are more inclined to display unidirectional citation correlations. On the other hand, the occurrence of mutual citation clusters is much lower, but their Z-Score [13] values are generally much higher than motifs belonging to the unidirectional citation clusters. Z-Score is a certain variable to weigh the statistical significance in real networks with a comparison

to the corresponding randomized networks. Generally speaking, the higher Z-Score a motif has, the more significant for it to present typical characteristics in a network. Considering in the journal citation network, the mutual citation clusters’ high Z-Score value can partly illuminate the mutual citation pattern is a special pattern occurring in journal networks.

Figure 4 shows motif frequency distribution of partial communities, according to which we can classify these communities into two kinds. In the first kind, the motifs lying in the frontal part of coordinate show a higher frequency compared to the second kind, while the motifs lying in the latter part of coordinate have a lower frequency. The second kind displays in a completely opposite way. To a further analysis, the first kind communities are commonly large in node scale, for example, the Medical Sciences community has 437 nodes; the Biological and Agricultural community has 184 nodes. The second kind communities have relatively small node scale, with only 25 nodes in Light Industry & Textile community and 37 nodes in Chemical Sciences community.

Table 3. The Motif Frequencies in Several Network Communities (2004)

| |  |  |  |  |  |  |  |
|--|---|---|---|---|---|---|--|
| Civil & Water | 10.62% | 12.56% | 18.66% | 16.45% | 14.24% | 3.84% | 6.10% |
| Mathematical Sciences | 10.20% | 10.41% | 10.95% | 11.82% | 12.04% | 4.77% | 9.76% |
| Biology & Agriculture | 9.68% | 26.85% | 16.12% | 19.26% | 8.48% | 3.08% | 5.97% |
| Light Industry & Textile | 3.55% | 9.22% | 10.28% | 28.37% | 9.93% | 4.26% | 1.06% |
| Traffic Related | 9.24% | 11.14% | 17.08% | 19.97% | 14.27% | 7.01% | 3.80% |
| Mechanical Engineering | 6.38% | 16.53% | 10.55% | 25.48% | 9.38% | 6.04% | 3.72% |
| Chemical Sciences | 6.20% | 11.32% | 7.91% | 22.97% | 13.02% | 10.02% | 2.66% |
| Electronic Info. & Computer | 9.79% | 27.90% | 15.01% | 17.39% | 8.35% | 2.32% | 7.16% |
| Geological & Geophysical | 4.87% | 11.99% | 8.97% | 23.53% | 9.54% | 7.35% | 3.88% |
| Material Sciences | 5.51% | 22.65% | 9.32% | 26.60% | 5.63% | 5.01% | 5.63% |
| Comprehensive | 24.65% | 23.27% | 30.71% | 8.51% | 7.10% | 0.34% | 3.28% |
| Medical Sciences | 14.82% | 38.36% | 19.29% | 11.16% | 6.62% | 0.92% | 4.18% |
| Physical Sciences | 6.17% | 12.88% | 11.99% | 24.34% | 10.94% | 7.94% | 4.06% |
| |  |  |  |  |  |  | UCC |
| Civil & Water | 1.00% | 3.57% | 2.36% | 3.89% | 5.10% | 1.63% | 48.92% |
| Mathematical Sciences | 0.98% | 6.94% | 5.75% | 4.45% | 9.76% | 2.17% | 42.30% |
| Biology & Agriculture | 0.38% | 2.96% | 1.85% | 1.79% | 2.73% | 0.82% | 59.01% |
| Light Industry & Textile | 1.16% | 7.80% | 0.71% | 1.06% | 11.70% | 12.06% | 25.27% |
| Traffic Related | 0.35% | 2.56% | 2.48% | 2.97% | 6.19% | 2.15% | 41.60% |
| Mechanical Engineering | 0.27% | 5.57% | 2.16% | 2.88% | 6.98% | 3.98% | 37.44% |
| Chemical Sciences | 0.33% | 5.04% | 2.59% | 2.52% | 9.41% | 6.07% | 28.42% |
| Electronic Info. & Computer | 0.59% | 4.98% | 2.18% | 1.18% | 2.19% | 1.23% | 60.44% |
| Geological & Geophysical | 0.06% | 6.41% | 2.75% | 3.86% | 9.74% | 6.53% | 29.77% |
| Material Science | 6.38% | 2.32% | 2.13% | 6.82% | 1.94% | 0.08% | 49.50% |
| Comprehensive | 0.46% | 0.17% | 0.53% | 0.19% | 0.57% | 0.31% | 82.37% |
| Medical Sciences | 0.88% | 1.35% | 1.14% | 0.58% | 0.91% | 0.51% | 77.53% |
| Physical Sciences | 3.88% | 3.17% | 3.88% | 7.05% | 2.82% | none | 38.98% |

*UCC means unidirectional citation clusters

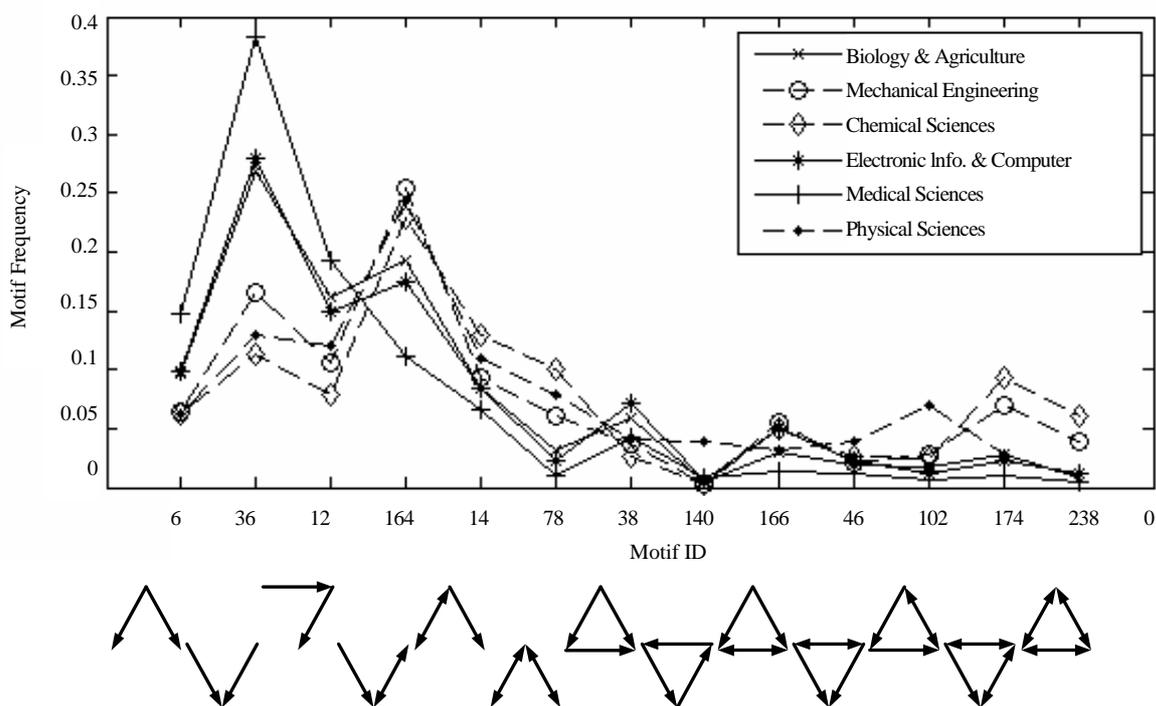


Figure 4. The comparison of motif frequencies in communities

Considering with nodes connection principles in the twenty communities respectively, it is found that in the first kind of communities, the sum frequencies of motifs in unidirectional citation clusters exceed 50%, meaning a dominance of unidirectional citation patterns. A few “hub nodes” have much larger indegree, and other nodes are inclined to connect to these nodes. However, these “hub nodes” always have very few citing connections with other nodes, even containing in the same community.

For the second kind of communities, the sum value of motif frequencies belonging to the mutual citation clusters is more than 50%, making a dominance of mutual citation pattern. We can see from above analysis that most nodes in this kind of community play a common role with no citation inclination in them.

When considering node or edge as the basic granularity, one characteristic of community structure is the loose connections between two communities. Then what characteristics it will show when take three-node motifs as the basic granularity? For a further investigation, we also take a research on motif constitution and citation patterns between these twenty communities. We find that citation pattern between communities are entirely inclined to the unidirectional pattern, meanwhile the frequency of unidirectional citation clusters is more than 70% between most communities. This statistical data is even up to 100% between biology and agriculture community and electronic information and computer community.

Meanwhile, it is also shown the frequency of the unidirectional citation clusters between any two

communities is generally much higher than the corresponding frequency in both two communities. For example, the electronic information & computer community and medical community both have an inclination to unidirectional citation pattern with the frequency of unidirectional citation cluster 60.44% and 77.53%, respectively. But this frequency rises up to 91.28% between these two communities.

4.2 Node Centrality in Communities

The structure of complex networks is typically characterized in terms of heterogeneous and topology differentiate of nodes. Take node centrality in different communities into consideration. Based on the classical centrality measures, here this paper mainly discusses degree centrality and closeness centrality. The former reflect the numbers of links incident upon a node, while closeness centrality defines as the reciprocal of geodesic distance between nodes. Since the lower closeness value a node has, the higher distance it reaches other nodes, here we take two typical networks from the two kinds of communities. One is electronic information & computer community as the unidirectional citation pattern community and the light & textile industry community as the example of mutual citation pattern. Figure 5 shows the frequency distribution of node centrality in the electronic information and computer community

It is shown that nodes with large indegrees generally corresponding to small outdegrees, and the ones with large outdegrees turn out to have small indegrees. The

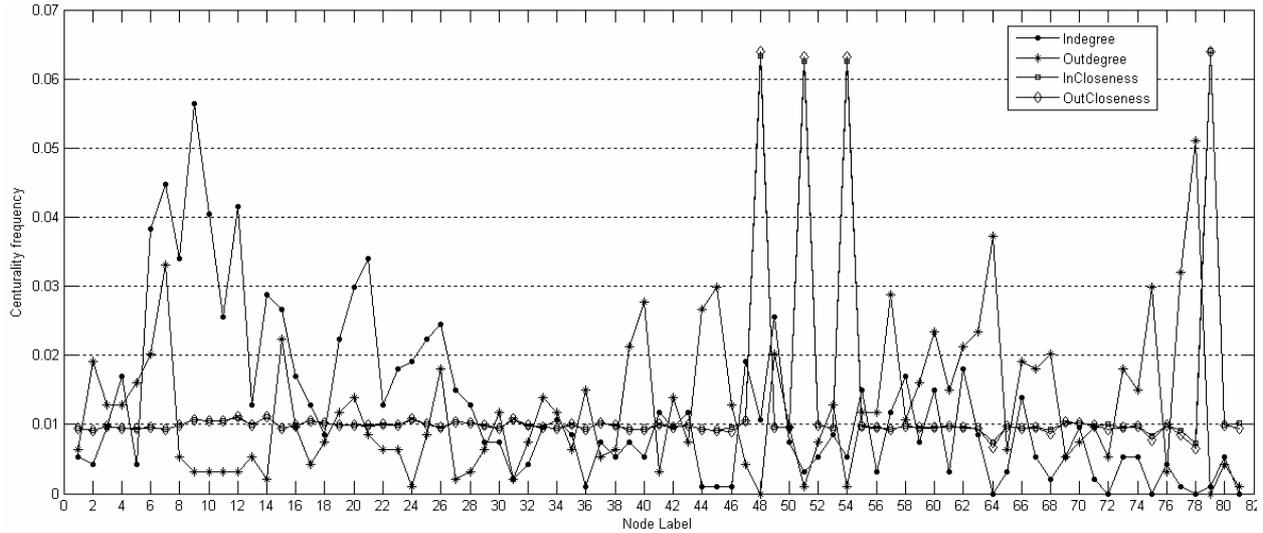


Figure 5. The centrality on electronic info. & computer communities

electronic information & computer community contains 88 nodes, and there are 19 nodes with indegrees bigger than 17, in which nearly 70% nodes with outdegrees smaller than 10. The journal with biggest indegree is “Computer Engineering and Applications”. Its indegree value is up to 59, but only has an outdegree value of 14. To be contrast, “Journal of Beijing University of Posts and Telecommunications” as the journal with biggest outdegree of 27, only having an indegree of 5. The degree distribution characteristic induces a strong unidirectional citation pattern in this kind of communities. On the other hand, though the incloseness and outcloseness of each node are approximately consistent, the whole picture shows sharp changes. Most nodes have small closeness, except four of them, which are “Computing Techniques for Geophysical and Geochemical Exploration”, “Robots”, “Piezoelectrics and Acoustooptics” and “Electronic Components & Materials”. These four nodes all locate near the edge of network, as shown in Figure 8. left. And their large closenesses

indicate the loose citation relations between them and other nodes.

In the same way, taking the light & textile industry community as an example for the second kind of community to analyze its centrality, and the tendency is shown in Figure 6. Nodes with large indegrees usually also have large outdegrees, and vice versa. The node with maximum in-degree and maximum out-degree all belongs to “Food Science”, the two values of which don’t have too much difference (in-degree=11, out-degree=7). This kind of degree distribution presents tight connections among the nodes in communities. With further consideration on closeness centrality, it is easy to figure out nodes in the light & textile industry distribute quite even from the globe network, because the closeness curve displays in a gentle way. Figure 7 gives a directly look on node closeness, where node sizes are consistent with their closeness value. This phenomenon tells a further illustration on the bidirectional citation tendency of the light & textile industry community.

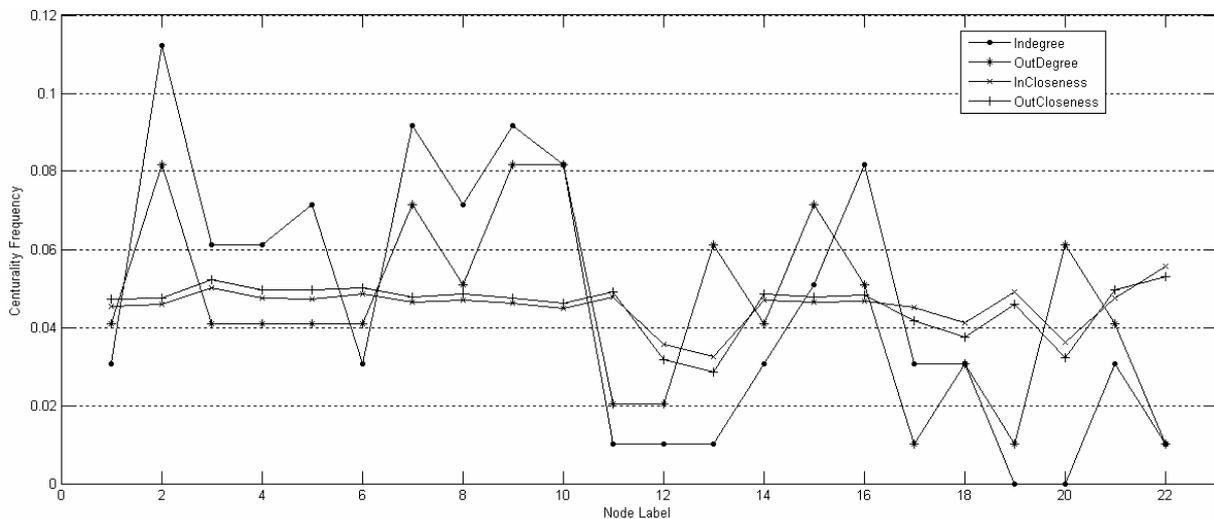


Figure 6. The centrality on light & textile industry communities

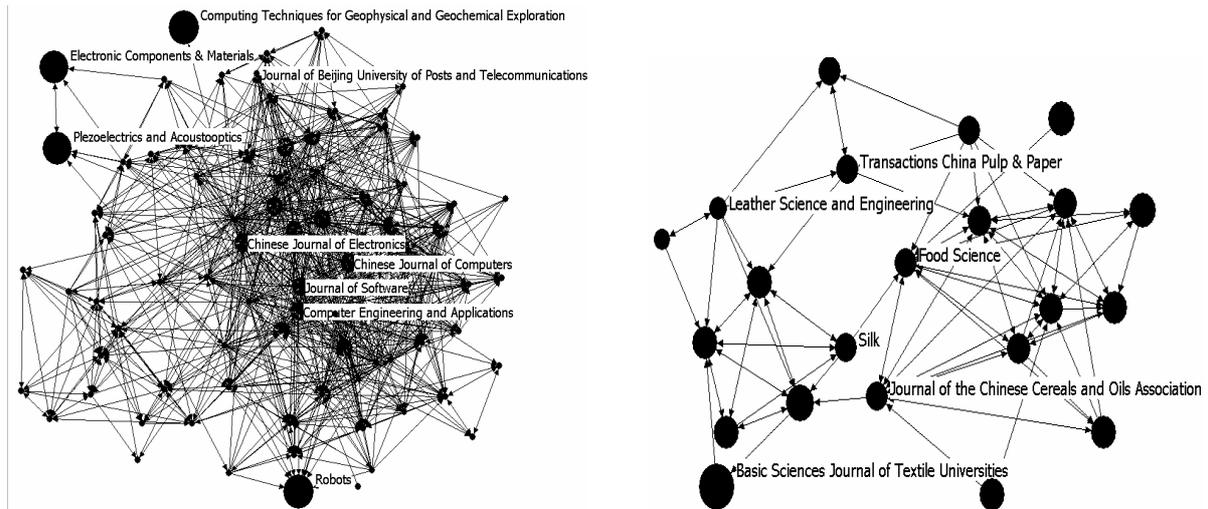


Figure 7. Centrality of journal network communities (left shows the electronic info. & computer community, right shows the light & textile industry community)

5. Conclusions

Chinese journal citation network is abstracted from more than one and a half thousands of Chinese journals of science and technique by CSTPC index. It is found these networks have obvious clustering characteristic and small-world pattern. This paper also borrows the motif concept into consideration to present some structure differences between two different kinds of network communities. One kind is more inclined to unidirectional citation pattern, while the other prefers the bidirectional citation ones. Then we give a further investigation on the reason of these two different kinds of citation patterns, according to node centrality in the communities. With a detailed statistics on node degree and its closeness, it illustrates communities of different kind also share different centrality characteristics.

Unlike general methods, this research takes three-node motifs as a basic granularity to find the discrepancy between different communities, rather than on a traditional node granularity. And it also gets some interesting ideas on journal citation networks. In the future, we could probably consider using motifs, a higher granularity to be a community partition criterion, instead of only using the system units, node or edge.

6. Acknowledgments

This work is partially supported by the National Grand Fundamental Research 973 Program of China under Grant No. 2007CB310803 and the National Natural Science Foundation of China under Grant No. 60675032.

REFERENCES

- [1] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, 393(6684): pp. 440–442, 1998.
- [2] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, 286(5439): pp. 509–512, 1999.
- [3] R. Milo, S. Shen-Orr, et al., "Network motifs: Simple building blocks of complex networks," *Science*, 298: pp. 824–827, 2002.
- [4] T. I. Lee, et al., "Transcriptional regulatory networks in *Saccharomyces cerevisiae*," *Science*, 298: pp. 799–804, 2002.
- [5] D. T. Odom, et al., "Control of pancreas and liver gene expression by HNF transcription factors," *Science*, 303: pp. 1378–1381, 2004.
- [6] N. Iranfar, D. Fuller, and W. F. Loomis, "Transcriptional regulation of post-aggregation genes in *Dictyostelium* by a feed-forward loop involving GBF and LagC," *Developmental Biology*, 290: pp. 460–469, 2006.
- [7] R. Prill, P. Iglesias, and A. Levchenko, "Dynamic properties of network motifs contribute to biological network organization," *PLoS Biology*, 3: pp. e343, 2005.
- [8] E. Ravasz, A. L. Somera, D. A. Mongru, et al., "Hierarchical organization of modularity in metabolic networks," *Science*, 297: pp. 1551–1555, 2002.
- [9] R. Milo, S. Itzkovitz, N. Kashtan, et al., "Superfamilies of evolved and designed networks," *Science*, 303: pp. 1538–1542, 2004.
- [10] P. Zhou, L. Leydesdorff, and Y. S. Wu, "The visualization of Chinese Journal of Scientific and Technic in citation environment," <http://users.fmg.uva.nl/lleydesdorff/istic03/index.htm>.
- [11] G. Kossinets and D. J. Watts, "Empirical analysis of an evolving social network," *Science*, 331: pp. 88–90, 2006.
- [12] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, 99: pp. 7821–7826, 2002.
- [13] J. Tyler, D. Wilkison, B. Huberman, "Email as spectroscopy: Automated discovery of community structure within organizations," *International Conference on Communities and Technologies*, pp. 81–96, 2003.
- [14] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences*, 101: pp. 2658–2663, 2004.
- [15] S. Fortunato, V. Latora, and M. Marchiori, "A method to find community structures based on information centrality," *Physical Review E*, 70: 056104, 2004.
- [16] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, 69: 066133, 2004.
- [17] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, 435 (7043): pp. 814–818, 2005.