16th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

16 - 20 October, 2017

WUHAN UNIVERSITY · WUHAN · CHINA

Conference Proceedings

www.issi2017.org
Contents

Why Do Some Research Articles Receive More Online Attention? Reasons for Online Success as Measured with Altmetrics ......................................................... 1

Can Twitter Increase the Visibility of Chinese Publications? .............................................. 10

Exploratory Analysis on the Construction of Journal Evaluation Model Using Altmetrics Indices ................................................................................... 20

Alternative Metrics Correlations: Do Academic Reviews Correlate with Library Holdings? 31

An Analysis of Scientific Co-author Network of Virtual Technology in China ......................... 44

A Weighted Method for Citation Network Community Detection ........................................ 58

Research on Domain Knowledge Network Based on Bibliometrics ........................................ 68

Detecting Social Communities in Spanish Theses Defense Committee in the Area of Computer Science .................................................................................. 83

Publication Patterns in the Social Sciences and Humanities in Flanders and Poland ............... 95

Assessment Criteria for Early Career Researcher’s Proposals in the Humanities .................... 105

Cognitive and Organizational Classification of Publications in the Social Sciences and Humanities ........................................................................................................ 112

Peer Review as A Delineation Criterion in Data Sources for the Assessment and Measurement of Scholarly Book Publishing ................................................. 118

Bibliometric Analysis of Publications from Post-Soviet Countries in Psychological Journals in 1992–2016 ..................................................................................... 125

Global Overview of Unmanned Aerial Vehicles Research: Country-level and Organisation-level Bibliometric Analysis .............................................................. 136

A Parameter-free Index for Identifying Under-cited Sleeping Beauties in Science ........................ 148
A Weighted Method for Citation Network Community Detection

Yunwei Chen*, Xue Xiao, Yong Deng, Zhiqiang Zhang

* Corresponding Author: chenyw@clas.ac.cn
Scientometrics & Evaluation Research Center (SERC), Chengdu Library and Information Center of Chinese Academy of Sciences, Chengdu, 610041 (China)

Abstract
There are many reasons for studying the community structure in citation network, whose underlying community structure may help us understand both the obvious and the more subtle interrelations between subfields, which are so important for understand the research structure and paradigm structure of a subject. However, most methods used for citation network community detection are based on topological structure or content separately. Usually such link-based methods ignore the content of citations, and the content-based methods lack the consideration of citation networks' topological structure, which might be crucial for accurate and meaningful community detection. In this paper, we develop a weighted method to identify the community structure of citation network. The new weighted model incorporates both the topological links and content information in citation networks to find communities. A text similarity method is used for weighting the links in citation networks and then constructs a weighted citation network to carry out community detection analysis. Two case studies had been carried out in two fields of Scientometrics and Evolving Citation Networks. Results revealed that using weighted citation edges with VSM-based nodes similarity to detect communities can improve the modularity Q value, which might mean a better assignment result. Four nodes’ movement with the direction of better appropriate of community assignation in the 188 node citation network of the field Scientometrics had also indicates the advantage of our weighted method.

Conference Topic
Social network analysis.

Introduction
In this work, we develop a weighted method to identify the community structure of citation network. There are many a kind of reasons for studying the community structure in citation network, whose underlying community structure may help us understand both the obvious and the more subtle interrelations between subfields, as well as the growth and the ebb of subfields (Chen & Redner, 2010). Communities are the clusters of closely connected nodes within a network, which is a property found in many networks, such as biological networks, the World Wide Web, social networks, collaboration works, citation networks, et al (Girvan & Newman, 2002). To identify communities within networks, a variety of methods have been developed. A recent new and powerful method to detect communities in complex networks is the study of Girvan and Newman (2002), who boomed the community research for their new algorithm using information about edge betweenness (Newman, 2001) to detect community peripheries, we call GN algorithm. However, the GN algorithm and some other algorithms are computationally demanding, which limits their application only to small networks. Therefore, Newman (2004a) proposed a new much faster algorithm, typically thousands of times faster than previous algorithms, we call FN algorithm. FN operates on different principles to GN but gives qualitatively similar results. FN algorithm is based on the idea of modularity, with the optimization of Q values over all possible divisions to find the best communities. Modularity measure Q is calculated to know when the communities found by the algorithm are good ones (Newman & Girvan, 2004). It has become a popular approach of optimizing a modularity function to community detection (Waltman & Van Eck, 2013). Compared to the earlier community detection methods, Q requires no extra knowledge beyond the network structure itself and can be applied to any type of network (Chen & Redner, 2010).
Nevertheless, the above algorithms proposed in the literature are suitable only for small and medium-sized networks. Therefore, some optimization algorithms were introduced to produce high-quality results even for very large networks, such as citation networks with tens of millions of nodes and hundreds of millions of edges. The first existing algorithms for large-scale modularity optimization is proposed by Blondel et al. (2008), often referred to as the Louvain algorithm. The other one is Louvain algorithm with multilevel refinement proposed by Rotta and Noack (2011), which is an extension of Louvain algorithm. Built on the ideas from these two existing algorithms, Waltman & Van Eck (2013) proposed a smart local moving (SLM) algorithm. All the three algorithms rely heavily on a well-known local moving heuristic (LMH), the idea of which is to repeatedly move individual nodes from one community to another in such a way that each node movement results in a modularity increase. We do not want to discuss the details of these algorithms for the purpose of this paper is not to reconstruct or modify them, but to make more meaningful citation networks and provide for the algorithms to detect communities. The theories and steps of the algorithm could be found in these papers: Girvan & Newman (2002), Newman (2004a), Newman & Girvan (2004), Waltman & Van Eck (2013).

Benefitted from the progress of community detection algorithms recent years especially as discussed above, lots of works focusing on the community structure of many kinds of networks had been carried out, which includes citation networks.

In the next section, we outline the principle of modularity measure $Q$, the Louvain algorithm that we use to analyse community structure within citation network, and the related works about citation networks community detection.

Based on the below related works analysis, we found that although many methods had been used to disclose the community structure of citation networks, there were still very few works focused the difference of different citation links. Therefore, this study advanced the existing works by constructing a weighted citation network to carry out community detection analysis. To our knowledge, none of the existing studies has done such work as ever. The significance of this work is that using weighted citation edges with VSM-based nodes similarity to detect communities with an improvement value of modularity $Q$, which might mean a better assignment result. Our weighted method breaks the barriers of traditional citation network analysis that regarding all edges as the same importance. We find that which avoids some of the shortcomings of the traditional techniques. It would be very valuable for many kinds of scientists, policy makers and stakeholders to deal with it.

**Related works**

In this section we review some representative existing methods for detecting communities and make a brief introduction to their advantages and disadvantages.

**Modularity measure $Q$**

Modularity $Q$ is the fraction of edges that fall within communities, minus the expected value of the same quantity if edges fall at random without regard for the community structure. If a particular division gives no more within-community edges than would be expected by random chance we will get $Q = 0$. Values other than 0 indicate deviations from randomness, and in practice values greater than about 0.3 appear to indicate significant community structure (Newman, 2004a).

Newman & Girvan (2004) defined $Q$ as follows,

$$Q = \sum_i (f_{ii} - a_i^2)$$

where $f$ is a k×k symmetric matrix whose element $f_{ij}$ is the fraction of all edges in the network that link vertices in community $i$ to vertices in community $j$. So they further define the row (or
(column) sums $a_i = \sum_j f_{ij}$, which represent the fraction of edges that connect to vertices in community $i$. In a network in which edges fall between vertices without regard for the communities they belong to, we would have $f_{ij} = a_i a_j$.

Newman (2004b) defined $Q$ in weighted networks (edges have weights, for example, based on the number of co-authored papers between two authors) as:

$$Q = \frac{1}{2m} \sum_y [A_{ij} - \frac{K_i K_j}{2m}] \delta(c_i, c_j)$$

where $c_i$ denotes the community to which node $i$ has been assigned, $A_{ij}$ represents whether there is an edge between nodes $i$ and $j$ ($A_{ij} = 1$) or not ($A_{ij} = 0$) in un-weighted networks, or the weight of the edge between node $i$ and $j$ in weighted networks. $K_i$ is the degree of node $i$ in un-weighted networks or the sum of the weights of the edges linked to node $i$, the $\delta(c_i, c_j)$ indicates whether nodes $i$ and $j$ belong to the same community. It equals 1 if $c_i = c_j$ and 0 otherwise. And $m = \frac{1}{2} \sum_y A_{ij}$ is the total number of edges in the network.

There is also a third formula of $Q$ can be written as:

$$Q = \sum_{i=1}^{n} \left[ \frac{L_v}{L} - \left( \frac{d_v}{2L} \right)^2 \right]$$

Here $n$ is the number of modules, $L$ is the number of links in the network, $L_v$ is the number of links between nodes in module $v$, and $d_v$ is the sum of the degrees of the nodes in module $v$.

**Louvain algorithm**

Louvain algorithm is proposed by Blondel et al. (2008), which finds high modularity partitions of large networks in short time and that unfolds a complete hierarchical community structure for the network, thereby giving access to different resolutions of community detection. This algorithm is based on the idea of the local moving heuristic is to repeatedly move individual nodes from one community to another in such a way that each node movement results in a modularity increase. It is divided in two phases that are repeated iteratively. Assume a network with $n$ nodes. First, we begin with each one node is a community. The algorithm then uses the local moving heuristic to obtain an improved community structure. Hence, in this initial partition individual nodes are moved from one community to another until no further increase in modularity can be achieved. At this point, a reduced network is constructed. Second, we build a new network whose nodes are the communities found during the first phase. The weights of the links between the new nodes are given by the sum of the weight of the links between nodes in the corresponding two communities. Once this second phase is completed, it is then possible to reapply the first phase of the algorithm to the resulting weighted network and to iterate until there are no more changes and a maximum of modularity is attained (Blondel et al., 2008).

**Communities in citation networks**

Citation network is originated from the reference relationships. There are many researches show that there is obvious trend of concentration and discrete in citation network, and publications in the same field seems like to cite semblable publications. Community in the citation network represent related papers on a single or some closely relevant topics. Community detection method is used for finding related papers in a given field. We could find the communities in citation networks and help researchers or policy makers to watch the whole pattern of a field at a glance. Subelj et al. (2016) presented a systematic comparison of the performance of a large number of citation networks clustering methods and found that
there is a trade-off between different properties that may be considered desirable for a good clustering of publications. At present, the methods used for citation network community detection could be divided into two classes. One is based on topological structure and another is based on content, such as text similarity.

Topological structure based methods usually use the community detection algorithms in the field of complex network to carry out empirical study in citation networks. For example, Wanjantuk et al. (2004) used the random walk graph clustering algorithm basing only on the link structure of a graph that can efficiently identify highly topically related communities to identify the communities within citation graph. Kajikawa et al. (2008) tracked the emerging research domains in energy research by using citation network analysis. They used FN algorithm to divide the largest connected component into some communities. Chen & Redner (2010) investigated the community structure of physics subfields in the citation network of all Physical Review publications between 1893 and August 2007. They focused only on well-cited publications that received more than 100 citations, and applied modularity maximization to uncover major communities that correspond to clearly identifiable subfields of physics. Ren et al. (2012) proposed a new clustering model for the high clustering in citation networks and got higher clustering coefficient and the size distribution of co-citation clusters in real networks. Wu et al. (2015) developed a local community detecting algorithm to find the corresponding research group. Kusumastuti et al. (2016) identified the clusters and sub-clusters of citation networks using the CitNetExplorer software to visualize timeline-based citation patterns.

Content based methods usually use traditional clustering algorithms to classify citation networks when the virtual links basing similarity are used to replace the actual links. For instance, Aljaber et al. (2010) presented a new approach for clustering scientific documents basing on the utilization of citation contexts. Chen et al. (2013) proposed a community discovery algorithm of citation semantic link network to find the semantic community in citation semantic link network. Liu et al. (2015) proposed a citation similarity based community detection method by transforming citation network to paper similarity network to detect the communities in citation networks.

Comparing the above two kinds of methods we could find that topological structure based methods usually lack the consideration of content-based text similarity, while the latter methods have also ignored the networks’ topological structure. Nevertheless, both the above two kinds of methods could not fulfill the need of citation network community analysis when used separately. Therefore, there are also some studies had paid attention to this question. For example, Cohn & Hofmann (2001) described a joint probabilistic model for modelling the contents and inter-connectivity of document collections such as sets of research paper archives. Erosheva et al. (2004) explored an internal soft-classification structure of articles based only on semantic decompositions of abstracts and bibliographies and compared it with the formal discipline classifications. Pathak et al. (2008) presented a Bayesian generative model for community extraction which incorporates both the link and content information in networks and its empirical study had been carried out on an email network. Although these papers did not analyse the citation network’s community, they indeed provided a powerful idea for community detection in citation networks. Indeed, most such studies incorporating both the topological link and content information had been studied in traditional complex networks, such as.

Therefore, this study advanced the existing works by incorporating both the topological links and content information in citation networks to find communities. A text similarity method is used for weighting the links in citation networks and then constructs a weighted citation network to carry out community detection analysis.
Data and methods

Data

This paper prepared two datasets to carry out case study. One is all types of 4417 publications published in the journal of *Scientometrics* in 1978–2016 (publication year) were downloaded and chosen as base data of case study to construct citation network. Data was acquired from the Web of Science in March 13 2016. The citation network consists of 1483 publications published in *Scientometrics* between 1978 through January 2016 with at least one citation (the nodes) to the other articles among the 4417 publications. At the same time, publications have no any field of title, abstract and keywords had been removed from the network. Therefore, the 1483 nodes make up a network with 28 weakly connected components through 3072 arcs. Among which the largest connected component consists of 1420 nodes, which had been used to find communities according to different weighted methods or unweighted method.

The other is the 2046 publications (article, letter, proceedings paper and review) of Evolving Citation Networks published before 2014 in WOS. Search terms to retrieval papers are: TS=((citation* and (evolv* OR evolut* OR "knowledge flow" OR path OR trajectory* OR backbone)) NOT "citation classic"). The citation network consists of 340 publications with at least one citation (the nodes) to the other articles and also has all fields of title, abstract and keywords among the 2046 publications. Therefore, the 340 nodes make up a network with 32 weakly connected components through 531 arcs. Among which the largest connected component consists of 246 nodes, which had been used to find communities according to different weighted methods or unweighted method.

Although both the citation networks have more than one connected component, most nodes locate in the largest connected component. In this paper, only the largest connected components are included. On the other hand, the directed arcs in citation networks are treated as undirected edges used for community detection in this paper. Science of Science (Sci²) Tool (http://cns.iu.indiana.edu) and Gephi were used to citation network construction and community detection.

Weighted method- VSM-based nodes similarity calculation

The Vector space model (VSM) is a mathematical model first presented by Salton, Wong and Yang (1975) and generally defined by Salton (1989) as an algebraic model for representing text documents as vectors of identifiers such as index terms. It is now popular used in information filtering, information retrieval, indexing and relevancy rankings.

In VSM documents are represented as mutually independent ordered term vectors, \((T_1,T_2,T_3,\ldots, T_n)\). Each term \(T_i\) has its own weight \(W_i\) according to the contribution to the documents. Thus document \(d_j\) is represented as vector \(d_j= (W_{1,j},W_{2,j},W_{3,j},\ldots, W_{n,j})\). Each dimension corresponds to a separate term. The value of a term \(i\) in the vector is non-zero if it occurs in the document. Then we use the TF-IDF weighting method to computing the values of \(W_i\). However, the TF-IDF algorithm does not consider the position of a term in a document. There are some studies show that the importance of a term is different according to the position in the document, such as title, keywords and abstract. Therefore, three groups of position weight \(\gamma\) are added to \(W_i\) based on the following formula.

<table>
<thead>
<tr>
<th>(\gamma)</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>2.5</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>When term (i) lies in keywords</td>
</tr>
<tr>
<td>1.0</td>
<td>1.5</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>When term (i) lies in title</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>When term (i) lies in abstract</td>
</tr>
</tbody>
</table>

Thus, the weight of term \(i\) is calculated as follows.

\[W_i=TF(i) \times IDF(i) \times \gamma (i)\]
In the case of a term appears in more than one position in a document, the highest value of $\gamma$ is accepted.

Finally, we use cosine similarity to calculate the similarity between a pair of vectors. It has to be noted that although citation links have directions, the original citation networks that we use are regarded as un-weighted and undirected and do not have loops. In our citation networks the cosine similarity between two linked documents is regarded as the weight of the link.

**Results and Analysis**

We have used both the two largest connected components of the two datasets to perform basic tests of the significance of the communities that are discussed in this work based on the three group of $\gamma$. First, we performed the same community detection analysis described above for the two set of the two largest connected components without weights. Second, VSM based weighted method had been accepted to reconstruct citation network at the three ways of $\gamma$. Results of community detection in the two datasets are listed in table 1.

<table>
<thead>
<tr>
<th></th>
<th># communities</th>
<th>modularity Q</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scientometrics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(largest connected component consists of 1420 nodes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unweighted</td>
<td>20</td>
<td>0.672</td>
</tr>
<tr>
<td>weighted group 1</td>
<td>26</td>
<td>0.764</td>
</tr>
<tr>
<td>weighted group 2</td>
<td>30</td>
<td>0.780</td>
</tr>
<tr>
<td>weighted group 3</td>
<td>28</td>
<td>0.769</td>
</tr>
<tr>
<td><strong>Evolving Citation Networks</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(largest connected component consists of 340 nodes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unweighted</td>
<td>13</td>
<td>0.722</td>
</tr>
<tr>
<td>weighted group 1</td>
<td>14</td>
<td>0.743</td>
</tr>
<tr>
<td>weighted group 2</td>
<td>14</td>
<td>0.730</td>
</tr>
<tr>
<td>weighted group 3</td>
<td>12</td>
<td>0.739</td>
</tr>
</tbody>
</table>


Table 1 shows the increasing values of modularity Q through weighted method for community detection in both case of Scientometrics and Evolving Citation Networks no matter which $\gamma$ was used. However, the result could not show which group of $\gamma$ is a better choice to improve the performance of community detection. For example, $\gamma$ group 2 seems to be a better choice for the case of Scientometrics but it looks like the worse for the case of Evolving Citation Networks, for which $\gamma$ group 1 might be better. The reason might be related to the particular feature of a field or the size of a network.

Thus what we can conclude is that our weighted method truly improved the ability to find communities in citation networks when evaluated by the values of modularity Q. In order to find what this difference is, we decided to investigate further the difference between weighted and unweighted methods. Weighted method with $\gamma$ group 2 was used to be carry out comparison analysis to unweighted network. Meanwhile, in order to keep the scope manageable to get a more careful look at the nodes moving between weighted and unweighted methods, we restrict ourselves to well-cited 188 publications by k-core >4 in the 1420 nodes of the case of Scientometrics, defined as those with higher degrees, to be shown in visualization maps. Then we got five communities. We present the topics of the five communities by co-keywords in Table 2, which are patent analysis, collaboration, research evaluation, university evaluation and science mapping.
Table 2. Topics of communities in Scientometrics dataset

<table>
<thead>
<tr>
<th>Communities</th>
<th>Topics</th>
<th>keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Patent analysis</td>
<td>Patent citation, Patent mining</td>
</tr>
<tr>
<td>2</td>
<td>Collaboration</td>
<td>Scientific collaboration, Collaboration network</td>
</tr>
<tr>
<td>3</td>
<td>Research evaluation</td>
<td>Research evaluation, h-index</td>
</tr>
<tr>
<td>4</td>
<td>University evaluation</td>
<td>Arwu, Ranking of universities</td>
</tr>
<tr>
<td>5</td>
<td>Science mapping</td>
<td>Mapping of science</td>
</tr>
</tbody>
</table>

Fig.1 illustrates the communities of the 188 publications in both unweighted and weighted networks, which shows that four nodes move between the four networks. They are n46, n1116, n686, and n1386. Their titles are listed in table 3. In order to judge whether the move of the four nodes from unweighted network to weighted network is positive or not, we read the full texts of the four publications carefully and summarized the themes comparing to the topics of the five communities.

Table 3. Four nodes distributed in different communities in Scientometrics dataset

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Title</th>
<th>Communities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>unweighted</td>
</tr>
<tr>
<td>n46</td>
<td>Knowledge network centrality, formal rank and research performance: evidence for curvilinear and interaction effects</td>
<td>3</td>
</tr>
<tr>
<td>n686</td>
<td>Bibliometrics evaluation of research performance in pharmacology/pharmacy: China relative to ten representative countries</td>
<td>5</td>
</tr>
<tr>
<td>n1386</td>
<td>The 100 most prolific economists using the p-index</td>
<td>5</td>
</tr>
</tbody>
</table>
Publication n46 (Badar, et al., 2015) explored the curvilinear association of co-authorship network centrality, degree, closeness and betweenness and the research performance, which considered formal rank of the authors as a moderator between network centrality and research performance. Its theme relates to both the topics of community 2 (collaboration) and community 3 (research evaluation). But by reading the paper we found that the central theme of n46 is scientific collaboration network analysis, and the usage for research evaluation is just an extended function. Therefore, assigning n46 to community 2 might be a better choice.

Node 1386 (Prathap, 2010) developed a new indicator called performance index (p-index), which was used to rank 100 most prolific economists. Node 686 (Ding, et al., 2013) evaluated the productivity of China in the field of pharmacology/pharmacy during 2001-2010 comparing to ten representative countries. Both of these two papers had made some mapping analysis but the key purposes were to carry out evaluation analysis. Therefore, these two publications were assigned preferably to community 3 by our weighted method.

Node 1116 (Zhao & Zhang, 2011) used the methods of co-word analysis, social network analysis and mapping knowledge domains as theory basis to construct the co-word network of the field of digital library research. It presented the study status quo and the research paradigm structure of digital library in China. Although it had also made some comparative analysis to that of international digital libraries research, the main contribution is illustration by mappings but not evaluation. Therefore, on the contrary to n1386 and n686, n1116 was moved to community 5 from community 3 by our weighted classification method.

The above analysis indicates that four nodes all move with the direction of better appropriate of community assignation through our weighted method. In fact, the four nodes are only key nodes with highly degrees among 188 nodes. There are much many nodes with lower degrees surrounded with the four moved nodes will also move with them to new communities in process of our new weighted method. In other words, our new weighted method will rebuild the research structure and paradigm structure of a subject, such as Scientometrics in this case. It might contribute to the better insight into the development of the field of Scientometrics.

**Conclusion**

Our study provides initial insights on the weighted method for community detection in citation network. An interesting advantage of using weighted community detection method is the fact that breaking the barriers of traditional citation network analysis that regarding all edges as the same importance. We find that using weighted citation edges with VSM-based nodes similarity to detect communities can improve the modularity Q value, which might mean a better assignment result. Four nodes’ movement with the direction of better appropriate of community assignation in the 188 node citation network of the field Scientometrics had also indicates the advantage of our weighted method.

However, by the result of this paper, we are not sure which part of a paper contributes more to the theme similarity between papers. Thus, an important question for further research is whether there is a common weights distribution feature among different parts of a paper or not, for example title, keywords, abstract and even full text.

Another problem with the citation network is that citation is strictly ordering in time and there is a problem of publication delay about months to years from a manuscript’s submission to publication or online when other authors can read it then cite it. As a result, a lot of highly relevant papers did not cite each other. That is to say, some correlative papers have no links in citation networks and suppressed the function of community detection in citation network to
depict the research structure and paradigm structure of a subject. Under this situation such kinds of highly relevant papers might be co-cited by later publications or have relation of bibliographic coupling by citing at least one common earlier publication. Thus, we suggest such publication pairs should be constructed as virtual edges and also be added selectively to our citation network. So we are facing quite a disturbing puzzle: how many or what kinds of such pairs need to be added to our network? We will take this question as one of our future research works.

We also hope that future research will further examine this question in line with deeper investigation into much more factors that could affect citation behaviour, such as co-author, co-institution, even the bias on language, institutions or countries.

Beyond all the problems identified above, we should also pay close attention to the dynamic evolving patterns of citation network, particularly on the transition of communities including merging, dividing, expanding, et al. For instance, Jung et al. (2014) proposed methods to analyse how communities change over time in the citation network graph without additional external information and based on node and link prediction and community detection.

Acknowledgements
This work is funded by the Documentation and Information Special Project of Chinese Academy of Sciences (2016). This work is supported in part by the West Light Foundation of the Chinese Academy of Sciences, China under grant no. [2013]165(3-6).

References
Liu, T.P. & Li, K. (2015). A Citation Similarity Based Community Detection Method in Citation Networks. 2015 Ieee Advanced Information Technology, Electronic and Automation Control Conference (Iaeac), 146-149.


