Scientometrics (2018) 114:1327–1343
https://doi.org/10.1007/s11192-017-2619-0

Identifying important scholars via directed scientific collaboration networks

Jianlin Zhou[1] · An Zeng[1] · Ying Fan[1] · Zengru Di[1]

Received: 7 August 2017 / Published online: 16 December 2017

Abstract Scientific collaboration plays an important role in the knowledge production and scientific development. Researchers have investigated numerous aspects of scientific collaboration by constructing scientific collaboration networks. And we can perform node centrality analysis on the scientific collaboration networks to identify important scholars. In these collaboration networks, two scientists are linked if they have coauthored at least one paper and the way of constructing these networks is based on the assumption that each author's contribution to an article is the same. However, the authors' contributions to an article are unequal in reality and we should pay attention to the impact of this unequal credit allocation on the understanding of scientific collaboration. In this paper, we regard the first author as the most important contributor to an article and build a directed scientific collaboration network. Then we identify important scholars by analyzing this directed network. For one thing, we investigate the difference between the undirected and directed scientific collaboration network in network properties and centrality analysis. For another, we apply different centrality indices: betweenness, PageRank, SIR and HITS to the directed scientific collaboration network. As a result, we find that each indicator has a different performance and the PageRank algorithm and SIR show highly positive correlation with in-degree. The HITS algorithm also shows better property which can hep us distinguish potential young scholars and identify important collaborators.

Keywords Scientific collaboration network · Credit allocation · Centrality analysis

✉ An Zeng
anzeng@bnu.edu.cn

✉ Ying Fan
yfan@bnu.edu.cn

[1] School of Systems Science, Beijing Normal University, Beijing 100875, People's Republic of China

Introduction

Research collaboration as one of important scientific activities is of great significance for knowledge creation and dissemination (Tijssen 2004; Wuchty et al. 2007; Sonnenwald 2007). As one of the most important elements of the research collaboration, scientists promote the emergence of novel ideas and increase the scientific research output by combining diversified knowledge and techniques of different research fields. Important scholars have more opportunities to get access to the knowledge and financial resources, thereby they could form more excellent research teams and create more influential and innovative knowledge. In order to achieve more effective collaboration and promote knowledge creation and spreading, identifying important scholars in research collaboration becomes more necessary and significative.

The scientific collaboration network is a social network where the nodes represent authors and the edges represent the co-authorship (Newman 2001). The researchers who occupy different network positions have their own characteristics and play different roles (Ebadi and Schiffauerova 2015). So the impact of scientists can be measured by analyzing their positions in collaboration networks. Some traditional network centrality measures such as degree, closeness, betweenness and PageRank and so on have been applied in the collaboration network to estimate the importance of scientists (Yan and Ding 2009). In addition, the sufficient analysis of the scientific collaboration network can help to reveal the cooperation mechanisms and characteristics of scientists and deepen the understanding of scientific collaboration. In the last 20 years, the studies of scientific collaborations networks have attracted wide attention in different fields (Newman 2001; Fan et al. 2004; Li et al. 2005; Hou et al. 2007; Ding 2011). Based on the empirical data, several aspects of scientific collaboration networks have been investigated and a series of research findings have been achieved, including basic structural properties (Newman 2001), assortative mixing (Newman 2002) and rich-club ordering (Colizza et al. 2006; Opsahl et al. 2008), community detection (Girvan and Newman 2002; Radicchi et al. 2004; Palla et al. 2005), network evolution patterns and mechanisms (Barabási et al. 2002; Evans et al. 2011; Zhai et al. 2014; Liu and Xia 2015), influence evaluation (Lu and Feng 2009; Yan et al. 2011; Yan and Ding 2011). The above-mentioned findings are all based on the network topology of scientific collaboration network which is undirected and unweighted. And the construction of this kind of network is based on the assumption that all authors contribute equally to an article. Moreover, many metrics such as h-index and its variants, PageRank and its variants, Q value (Sinatra et al. 2016) and so on have been developed to evaluate the scientists (Zeng et al. 2017). And these indicators can be used as an important reference when selecting important scientists.

A large amount of effort has been made by network scientists to further understand and model the evolution of the collaboration networks, with also numerous methods designed for ranking and prediction in these networks. However, all these works characterized the scientific collaboration data with undirected networks, leading to a huge amount of information loss. In the undirected networks, one scientist can be linked to other scientists due to many different cases, e.g. by either collaborating with them in a paper or collaborating individually with each of them in five separate papers. More importantly, whether a scientist is collaborating with others as a first author or not cannot be distinguished in the undirected networks. These factors are the key components determining the influence and future potential of scientists. As such, by analyzing only the undirected collaboration network will very likely result in a biased ranking and prediction of scientists behavior. We also learn that the collaboration networks can be modelled as directed network (Liu et al.

2005; Kim and Diesner 2015). For example, Yoshikane et al. (2006) regarded the first author of a paper as the leader and the other coauthors as followers and they built a directed collaboration network. They compared the co-authorship networks between the theoretical and application areas in computer science and identified important researchers. Inspired by Yoshikane et al. we also divide authors of a publication into two types: important contributors and ordinary contributors. Because the first author of an article is easily identified and makes a great contribution for the publication, for simplicity, we also only regard the first author as important contributor and other coauthors as ordinary contributors. Based on the above assumption, we construct the directed scientific collaboration network and identify important scholars via this network.

In this paper, we model the scientific collaboration data in the field of complex systems with directed networks which well capture the detailed structure of the collaboration behaviors completely neglected in the undirected network case. Considering the influence of the structure of collaboration network on identifying important scholars, we compare this directed scientific collaboration network with undirected scientific collaboration network in network property and node centrality. The results show that there are certain similarities such as power-law degree distribution and clear community structure as well as differences in cluster coefficients and degree correlation between these two networks. We apply three centrality indicators: betweenness, PageRank and SIR to the undirected and directed scientific collaboration network. We find that the ranking results on these two collaboration networks show obvious differences in terms of each index, especially PageRank and SIR. Therefore, the way of constructing collaboration network exerts an obvious influence on ranking scholars. Finally, we compare the ranking results of several centrality indicators applied into the directed scientific collaboration network. This research helps to understand the collaboration among scientists in the field of complex systems and contributes to the identification of important scholars in this field.

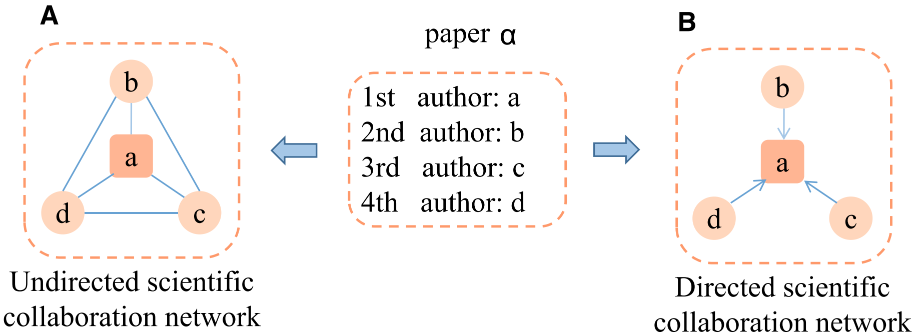
A
b
a
d
c
Undirected scientific collaboration network
paper α
1st author: a
2nd author: b
3rd author: c
4th author: d
B
b
a
d
c
Directed scientific collaboration network

Fig. 1 (Color online) The schematic diagram of two forms of collaboration networks for a scientific publication. **a** The undirected scientific collaboration network, where all authors in a paper are connected to each other. **b** The directed scientific collaboration network, where the directed edges are from non-first authors of a paper to the first author

Methods

Network construction and data

In this article, we study two kinds of scientific collaboration networks which are undirected and directed respectively shown in Fig. 1. The undirected scientific collaboration network is the most frequently mentioned and traditional collaboration network. This type of network is an undirected and unweighted. In this network, there exists an edge between the two scientists if they have written at least one article together. For the directed scientific collaboration network, currently it is much less studied. In this network, the directed edge tells us which node is ever as the first author or non-first author between two authors. For example, if there is an edge from author A to author B, it means that author A writes a scientific paper with author B and the author B is the first author in the paper. The directed scientific collaboration network which is also an unweighted network cuts off the relationship among the non-first authors in the publication.

To verify the differences between the two collaboration networks mentioned above, we focus on a field of science: complex systems, which is about studying how parts of a system give rise to the collective behaviors of the system, and how the system interacts with its environment [http://necsi.edu/guide/]. Our database used in this paper is collected according to Physics and Astronomy Classification Scheme(PACS) code which was developed by the American Institute of Physics (AIP) to identify fields and sub-fields of physics [https://journals.aps.org/PACS]. The relevant PACS codes of complex systems are 89.75.-k, 89.75.Da, 89.75.Fb, 89.75.Hc, 89.75.Kd. Because an article can involve several PACS codes and we will incorporate this article into our data set as long as it contains one of PACS codes mentioned above. We have collected 2323 papers from 2001 to 2009 in complex systems. For each paper, the information about its title, DOI, authors and their affiliations, publication time, PACS code and the DOI of its citing papers can be obtained. In our data set, each author has a unique name and we do not need to deal with author name disambiguation problem. Based on the above information, we construct the undirected and directed scientific collaboration network, respectively.

Centrality measures

With the development of complex networks, many indicators used to measure the node centrality have been proposed. We considered five representative centrality measures in this paper. Frankly speaking, it is still a challenging issue to determine which centrality measure performs best in the analysis of directed networks, as confirmed by a recent review article (Lü et al. 2016). In this context, we try our best to select diverse centrality metrics in order to cover different aspects for central nodes in directed networks. The first measure is the most straightforward index called degree centrality which uses only the one-step local information to estimate the centrality of a node. The second measure, however, is a global index called betweenness. It depends on the shortest path lengths and promotes the importance of bridging nodes. The third one is the well-known PageRank algorithm which computes the centrality score of nodes based on an iterative process. The fourth one is the centrality measure based on SIR model. With this dynamical process, the node with the largest spreading coverage is considered as the highest centrality. This centrality is adopted as it mimics the knowledge spreading scenario in scientific collaboration. The final measure is the HITS algorithm which assigns two roles to each node in the directed networks.

Degree centrality

A directed network can be described by an adjacency matrix A in which an element $a_{ij} = 1$ indicates there is a directed edge from node i to node j. Degree centrality as a local centrality to measure node's importance is very simple. For directed networks, two degree centralities: in-degree and out-degree should be considered. The in-degree of node i is the number of links directly pointing to the node i and the out-degree of node i is the number of links from node i to other nodes. The in-degree and out-degree of node i are defined as:

$$k_i^{\text{in}} = \sum_{j=1}^{n} a_{ji}, \quad k_i^{\text{out}} = \sum_{j=1}^{n} a_{ij}. \tag{1}$$

where n is the number of nodes in the network.

Betweenness centrality

The betweenness centrality is one of path-based centralities which can be applied into directed or undirected network and it measures the extent to which a particular node lies on shortest paths between the other pairs of nodes in the network (Kintali 2008). Therefore, the betweenness of node i is given by:

$$bc(i) = \sum_{i \neq s \neq t} \frac{g_{st}^{i}}{g_{st}}. \tag{2}$$

PageRank

PageRank was originally proposed to rank web pages in Google search engine (Brin and Page 1998). Its basic idea is that the importance of a web page is determined by both the quantity and the quality of the pages linked to it. Now it has been applied in different fields far beyond its origins (Gleich 2015). Especially, in scientometrics, PageRank and its variants have been used to evaluate the importance of different objects such as papers, scientists, journals and institutions. PageRank which is an iterative algorithm can also be applied in both directed and undirected network. Mathematically, the PageRank score of node i in the network at step t is given by:

$$PR_i(t) = d \sum_{j=1}^{n} A_{ji} \frac{PR_j(t-1)}{k_j^{\text{out}}} + (1-d)\frac{1}{n} \tag{3}$$

where n is the total number of nodes in the network, k_j^{out} is the number of outgoing links of node j, d represents the probability for a random walker to continue walking through the outgoing links of the present node and $(1-d)$ is the probability for a random walker to jump to a random node from the present node. In this article, we set $d = 0.85$ which is the typical value in computer science (Brin and Page 1998).

The SIR model

The SIR model is one of classic epidemic models and it could be extended to measure the spreading effect of specific nodes in the undirected or directed network (Lü et al. 2016). In this model, all nodes in the network belong to one of three states: susceptible(S), infected(I), or recovered(R). Initially, all nodes are in the susceptible state except for one infected node which is selected to examine its spreading performance. At each time step, each infected node will infect their susceptible neighbours with probability β and then enters the recovered state where the node cannot infect or be infected with a probability μ. For simplicity, in this paper we set $\mu = 1$. The spreading process will end until there is no longer new infected node appearing in the network. Then the spreading influence of the target node could be defined as the number of recovered nodes in the complete spreading process. In this paper, we set the infected probability $\beta = 1.5\beta_c$ where β_c is the approximate epidemic threshold:

$$\beta_c = \langle k \rangle / (\langle k^2 \rangle - \langle k \rangle) \tag{4}$$

HITS algorithm

HITS (hyperlink-induced topic search) algorithm is originally proposed to discover and rank web pages relevant for a particular topic (Kleinberg 1999). Now this method has been applied to rank publications, scientists and journals. The HITS algorithm measures a node's importance from two aspects: authority and hub. A good hub points to many good authorities and a good authority is linked to many good hubs. Thus each node in the directed network can be assigned two scores: authority score and hub score. For node i, We denote its authority score and hub score at time t by $a_i(t)$ and $h_i(t)$, respectively. Originally, the hub scores of all nodes are assigned one unit score and the authority and hub score of node v_i at time t can be computed as:

$$a_i'(t) = \sum_{j=1}^{n} a_{ji} h_j(t-1), \quad h_i'(t) = \sum_{j=1}^{n} a_{ij} a_j'(t). \tag{5}$$

where n is the number of nodes in the network. After each iteration, the value of $a_i'(t)$ and $h_i'(t)$ should be normalized as follow:

$$a_i(t) = \frac{a_i'(t)}{\|a'(t)\|}, \quad h_i(t) = \frac{h_i'(t)}{\|h'(t)\|}. \tag{6}$$

Results and analysis

Descriptive analysis for collaboration in complex systems (2001–2009)

First of all, we extract the basic information of the database we collected. In the period of 2001–2009, we collect 2323 papers in all and 3749 authors are involved. We investigate the distribution of number of authors per paper (2001–2009) shown in Fig. 2. We can see that more than half of the scientific papers have only two or three authors. And single-author papers account for about one tenth of the collected articles. Furthermore, there are few scientific articles ($<5\%$) whose the number of authors are more than 5. It indicates that the collaboration in the field of complex systems is relatively frequent and the scale of collaboration is also small. In order to better describe the development and collaboration of the field of complex systems, we make a more temporal analysis. For example, the

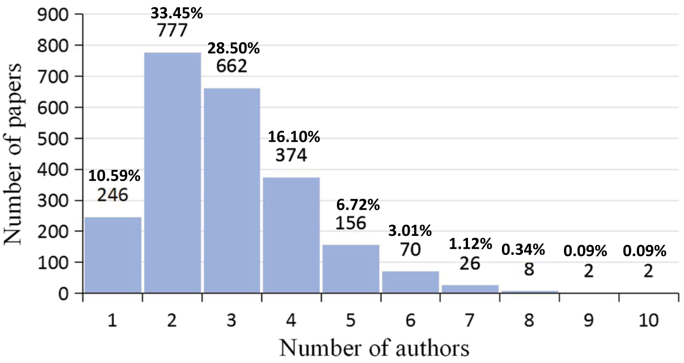

Fig. 2 (Color online) Distribution of the number of authors per paper in our data set

evolution of the number of articles and the number of authors per articles are shown in Fig. 3. One can see that the number of articles per year generally keeps increasing with a small decrease in year 2007. The positive growth trend of the the number of authors per articles also indicates that the collaboration in complex systems is becoming more closely. In order to further explore the collaboration, we analyze the evolution of the number of papers and the relative number of papers per year with different number of authors shown in Fig. 4. The relative number of papers per year is obtained by normalization of the number of papers per year with different number of authors. By observing the evolution of relative number of papers with different number of authors, we can find that overall the number of three-author papers shows a growing trend but the number of two-author papers shows a slight downward trend.

Comparative analysis of network properties

Based on the above-mentioned data set which is related to the field of complex systems (2001–2009), we have constructed two kinds of networks: undirected scientific collaboration network (USCN) and directed scientific collaboration network (DSCN). There are 3749 nodes and 6550 undirected edges in the USCN. And in the DSCN, it contains 3749

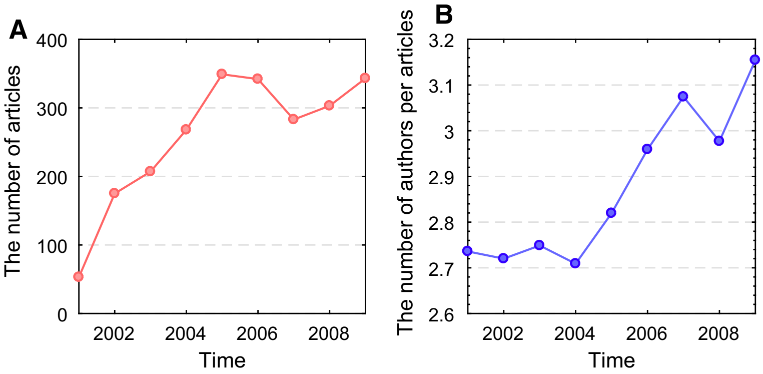

Fig. 3 (Color online) **a** is the number of articles over different years in our dataset. **b** is the number of authors per articles published in different years

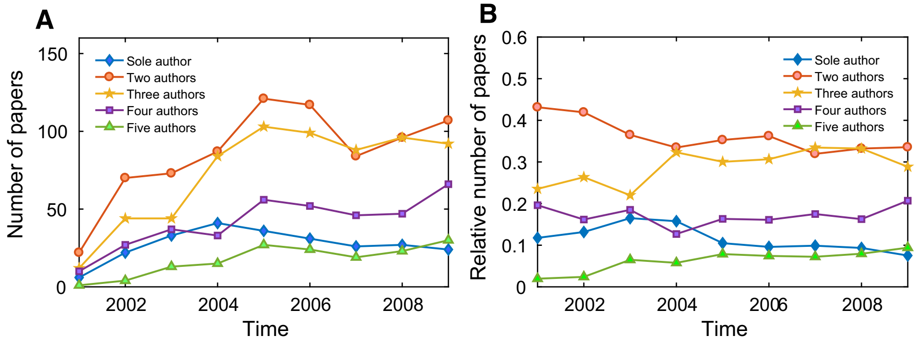

Fig. 4 (Color online) **a**, **b** are the evolution of the number of papers and the relative number of papers with different number of authors in the period 2001–2009, respectively

nodes and 3618 special directed edges. We can find that these two networks are not fully connected graphs. When we consider the weak components of DSCN, which ignores the edge direction of directed network, the USCN has the same distribution of components with DSCN shown in Fig. 5a. We find that there are 730 components in the networks, of which the giant component includes 1075 nodes and the second largest component contains 62 nodes. We also observe that the size of 475 components which contain 1012 nodes (26.99% of the network) is smaller than 4. Because the small components are not enough to reveal the network statistical properties, so we mainly focus on the statistical properties of the giant component (GC) whose scale is relatively large to analyze the differences between two networks.

There exist essential differences between USCN and DSCN due to the edge direction. Firstly, we investigate the network properties of the GC of USCN. For the undirected scientific collaboration network, the degree of author *i* is the number of collaborators who are ever with author *i* appeared in the author list of articles. We study the cumulative degree distributions of the GC shown in Fig. 5b. We can find that the cumulative degree distribution of the GC approximately follows the power-law distribution and the max degree is 56, which indicates that few scholars have a large number of collaborators and most people have a small number of collaborators.

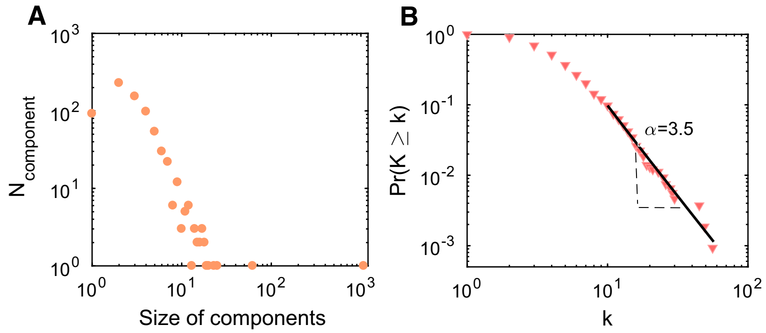

Fig. 5 (Color online) **a** is the distribution of the size of components of the USCN and **b** is the cumulative degree distribution of the GC of the USCN

Then we calculate other network properties of the GC of USCN such as cluster coefficient, average path length, degree–degree correlation presented in Table 1. The clustering coefficient of network measures the extent to which nodes in the network tend to cluster together. The average path length of the network is the average distance between all pairs of nodes in the network. We can observe the giant component of USCN shows the small-world property compared to the random networks of the same size and the GC of USCN has much bigger cluster coefficient and smaller average path length. Degree correlation describes the relationship between the degrees of nodes that link to each other and could reveal the the network topology which belongs to assortative network, disassortative network or neutral network. The degree correlation coefficient of the GC is close to 0 and not clearly distinct from random networks', which indicates the hierarchical structures of the GC of USCN is not obvious and it is a neutral network. Community structure as one of common meso-scale features means that there exist many subgroups of nodes that are densely connected internally but with few links outside the subgroups. To detect whether the GC has clear community structure, we adopt the index of modularity Q (Newman and Girvan 2004)which is one of widely used methods of community detection in networks. The optimal Q values are 0.8424 for the GC of USCN, which shows the network has obvious community structure.

Next, we study the statistical properties of the GC of the DSCN. For the DSCN, the in-degree of the author i is the number of collaborators who are as non-first author but author i is as the first author in the publications. And the out-degree of the author i is the number of collaborators who are as the first author in their co-authored papers. Then the degree of author i is the sum of in-degree and out-degree. The cumulative degree distributions of its giant component are shown in Fig. 6. The cumulative distributions of degree, in-degree and out-degree all approximately follow the power-law distribution and their scale-free exponents are larger than 3. Compared with the cumulative in-degree distribution, the cumulative out-degree distribution has a longer tail, which indicates that many scientists are more involved in others' works rather than as the principal person of the projects. For directed networks, there are four different patterns of directed triangles that result in four different clustering coefficients (Fagiolo 2007) shown in Fig. 7a. We can see that overall these four clustering coefficients are much smaller compared with the clustering coefficient of the GC of USCN, which exhibits significant difference between the USCN and DSCN. To some extent these kinds of triangles describe the partnership of an author's neighbours. We find that the case that three authors are all ever as the first author is very few and its clustering coefficient C_{cyc} is equal to 0.0057. We also calculate the four kinds of degree correlation coefficients (Foster et al. 2010) which quantify the assortative or disassortative tendency of the network shown in Fig. 7b. The values of (in, out) correlation coefficient

Table 1 The statistical properties of the GC of USCN as well as its corresponding random network

Network	$<k>$	$<c>$	$<l>$	r
GC of USCN	4.8242	0.7225	7.1445	−0.0132
RN	4.8242	0.0116	4.1642	−0.0244

RN is the corresponding random networks of USCN, which are generated 100 times by randomly reshuffling links. $<k>$, $<c>$, $<l>$, r represent the average degree, average clustering coefficient, average path length, degree correlation coefficient respectively

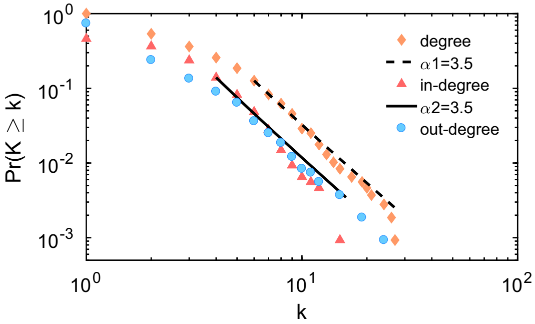

Fig. 6 (Color online) The cumulative distributions of degree, in-degree and out-degree in the giant component of the directed scientific collaboration network

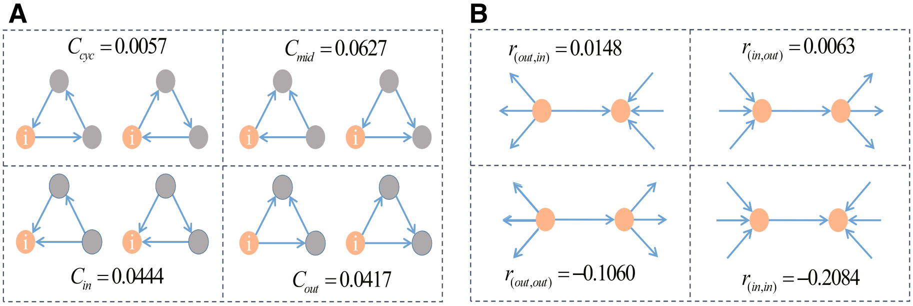

Fig. 7 (Color online) **a** is the four clustering coefficients with eight different directed triangles and **b** is the four degree–degree correlations in directed scientific collaboration network

and (out, in) correlation coefficient are close to zero, which reveals no significant correlation between the in-degree and out-degree. However, the (out, out) correlation coefficient is −0.106 and (in, in) correlation coefficient is −0.2084. They both show significant negative correlation, which indicates authors with many first-author papers tend to collaborate with the authors with few first-author papers. We also analyze the community structure of directed giant component using the method of modularity maximization (Leicht and Newman 2008) and the result shows that the directed giant component is divided into 72 communities with an optimal modularity of $Q = 0.8396$. So we can see that the undirected and directed scientific collaboration network has many similarities such as power-law degree distribution and community structure as well as differences such as cluster coefficient and degree correlation.

As the above analysis focuses on the static properties of collaboration network which is actually a complex evolving network, we also investigate the dynamical properties of these two kinds of collaboration network. Firstly, each year we will build two kinds of collaboration networks based on the cumulative data from 2001 up to a given year. And then we extract the giant components of each networks as well as analyze the network properties of these giant components. The results of the dynamical network properties of the giant components are presented in Tables 2 and 3. We can find that the network properties of two kinds of collaboration networks keep relatively stable in the periods of 2001–2009.

Table 2 The evolution of statistics for the undirected scientific collaboration network

Statistics	Time								
	2001	2002	2003	2004	2005	2006	2007	2008	2009
N	12	35	43	76	229	440	654	793	1075
$< k >$	7.1667	6.2286	6.2791	5.3684	4.5764	4.5273	4.7339	4.7617	4.8242
$< c >$	0.8876	0.8789	0.8702	0.7162	0.6956	0.6883	0.7058	0.7068	0.7225
$< l >$	1.3485	2.3546	2.3965	3.2751	9.2321	9.5832	7.7098	6.8796	7.1445
r	−0.3363	−0.0541	−0.0753	0.0194	0.0426	−0.0246	−0.0071	−0.0169	−0.0132
Q value	0.1255	0.4074	0.4475	0.5352	0.7543	0.7948	0.8278	0.8417	0.8451

N is the number of authors in the giant components

Node centrality analysis

The roles of researchers are tightly related to their positions in the collaboration network. Identifying these important researchers based on collaboration networks is one of important applications. Firstly, we compare the ranking results between the undirected and directed scientific collaboration network. We perform centrality analysis on the GCs of USCN and DSCN. Betweenness could revel a node's potential power in controlling information flow in a network and identify some important researchers who act as a broker or gatekeeper in the collaboration network. The PageRank algorithm could allocate nodes' resources to their neighbours iteratively to distinguish important scholars. And the SIR model which is one of the classical network diffusion models could identify the influential spreaders in terms of idea spreading and knowledge diffusion. We select the above three centrality indicators which they can be both applied into undirected networks and directed networks to identify important nodes. Based on the ranking results, we obtain the Spearman correlation coefficients in betweenness, PageRank and SIR between the two GCs are 0.6277, 0.1276, 0.3881, respectively. The correlation coefficients show us that the two network structures have marked differences in terms of ranking scholars, which indicates that the way of constructing collaboration network can directly affect the ranking results.

Further we compare the differences among ranking algorithms in the directed scientific collaboration network. Table 4 shows a correlation matrix which are calculated using Spearman correlation coefficient among the indices. We can find that the degree shows higher correlation with in-degree, PageRank and SIR, but lower correlation with out-degree. The in-degree shows negative correlation with out-degree but shows strong positive correlation with PagRank and SIR. Betweenness all shows positive correlation with other indicators. We can also find that the PageRank shows higher correlation with in-degree compared with degree. This is reasonable because the directed edges highlight the position of the first author which PageRank algorithm ranks the authors with many first-author papers higher just like in-degree. Different ranking algorithms have their own advantages and we need to combine different evaluation indicators to identify important scholars.

The HITS algorithm is also one of the well-known ranking methods which could capture nodes' importance from two aspects: authority and hub. We also make use of this method to evaluate the importance of scientists in the directed scientific collaboration network. The authority score and hub score of a scholar are related to in-degree and out-

Table 3 The evolution of statistics for the directed scientific collaboration network

Statistics	Time								
	2001	2002	2003	2004	2005	2006	2007	2008	2009
N	12	35	43	76	229	440	654	793	1075
$< k >$	2.5	2.4	2.5116	2.5789	2.6463	2.7136	2.8104	2.8222	2.8093
C_{cyc}	0	0	0	0	0.0008	0.0052	0.006	0.0049	0.0057
C_{in}	0.0278	0.0238	0.024	0.0278	0.0432	0.0473	0.0483	0.0458	0.0444
C_{mid}	0.0238	0.0221	0.0265	0.0564	0.0703	0.0668	0.0618	0.0569	0.0627
C_{out}	0.0278	0.0262	0.0205	0.0318	0.0359	0.0376	0.042	0.0387	0.0417
$r_{(\mathrm{out,in})}$	−0.7413	−0.4231	−0.324	−0.2405	−0.084	−0.0612	−0.0077	−0.0136	0.0148
$r_{(\mathrm{in,out})}$	−0.1321	−0.245	−0.2304	−0.1614	−0.0786	−0.039	0.0105	−0.0039	0.0063
$r_{(\mathrm{out,out})}$	−0.5854	−0.3269	−0.3132	−0.2824	−0.1376	−0.1402	−0.1284	−0.1272	−0.106
$r_{(\mathrm{in,in})}$	−0.8321	−0.6584	−0.6167	−0.4422	−0.4223	−0.308	−0.2557	−0.2594	−0.2084
Q value	0.2578	0.5709	0.607	0.6865	0.8034	0.8355	0.829	0.8572	0.8396

N is the number of authors in the giant components

Table 4 The Spearman correlation test of different ranking algorithms for the giant component of directed scientific collaboration network

	Degree	In-degree	Out-degree	Betweenness	PageRank	SIR
Degree	1	0.7198	0.2757	0.5946	0.6892	0.6895
In-degree	–	1	−0.3325	0.5651	0.9731	0.9782
Out-degree	–	–	1	0.3957	−0.3300	−0.3419
Betweenness	–	–	–	1	0.5452	0.5506
PageRank	–	–	–	–	1	0.9751
SIR	–	–	–	–	–	1

degree of the author respectively. Because the directed edge is from the the other coauthors to coauthor, which it highlights the important position of the first author, so the authors with many first author papers can have higher authority scores and the authors with many first-author collaborators can have higher hub scores. Based on the rank of authority scores, we can predict the development trend of young scholars' academic career and identify potential young scholars. High prestigious scholars are usually as the corresponding authors in the publications and guide young scholars to carry out the scientific research as well as they have higher hub scores. According to the rank of hub scores, we can identify the popular collaborators and choose scientific partners properly. The ranking result is shown in Fig. 8 and the highest ranking 5 scholars for authority scores and hub scores are listed in Tables 5 and 6, respectively. We could find that few scholars both have higher authority scores and hub scores like Eduardo López, Lidia A. Braunstein and Sergey V. Buldyrev. This kind of people can be predicted the potential young scholars and they are popular in the collaboration. We can also see that many scholars like H. Eugene Stanley and Shlomo Havlin who are domain experts in the field of complex systems have very high ranks in terms of hub scores. They have a better academic reputation and other scientists like to collaborate with them. There are also many scholars like Yiping Chen and Zhenhua

Fig. 8 (Color online) The distribution of the ranking results based on the authority scores and hub scores of the HITS algorithm

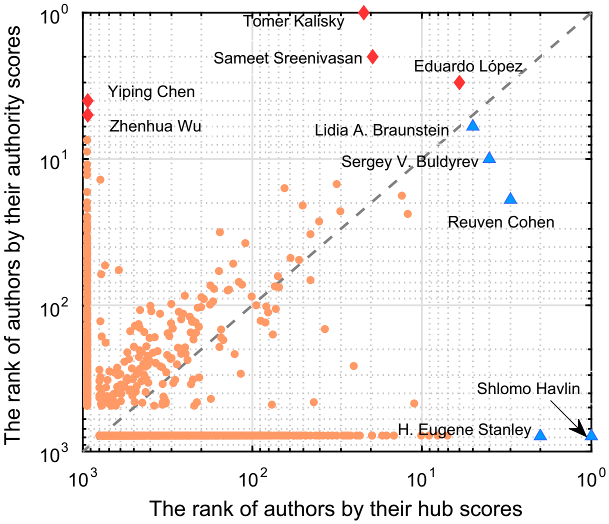

Table 5 The top 5 authors who are regarded as the potential scholars ranked by the authority scores

Rank	Name	Number of first-author papers	Number of non-first author papers	In-degree	Out-degree
1	Tomer Kalisky	5	1	10	1
2	Sameet Sreenivasan	2	1	8	1
3	Eduardo López	2	3	6	3
4	Yiping Chen	3	0	7	0
5	Zhenhua Wu	3	0	6	0

Table 6 The top 5 authors who are selected as the most popular collaborators ranked by the hub scores

Rank	Name	Number of first-author papers	Number of non-first author papers	In-degree	Out-degree
1	Shlomo Havlin	0	42	0	24
2	H. Eugene Stanley	0	32	0	19
3	Reuven Cohen	5	10	6	8
4	Sergey V. Buldyrev	2	8	3	7
5	Lidia A. Braunstein	1	8	4	5

Wu in our study who have very high ranks in terms of authority scores. These people as potential young scholars have better scientific research quality and creativity.

Discussion

In this work, we constructed a directed collaboration network which highlights the importance of the first author. We made an empirical analysis of the field of complex system based on this network with special focus on two main issues including (1) the difference between undirected and directed scientific collaboration network in terms of network properties; (2) the comparison of the ranking results between the undirected and directed scientific collaboration network. We perform our analysis on the giant components of these two networks. Compared with undirected collaboration network, the directed scientific collaboration network reduces the edge density but maintains the network connectivity. The results also show that there exist many similar network properties such as power-law degree distribution and community structures. However, due to edge direction, the directed scientific collaboration network shows its own unique characteristics in clustering coefficient and degree–degree correlation. For the directed collaboration network, one can see that the out-degree distribution is broader than the in-degree distribution, indicating that a scientist can have many non-first author papers but the number of their first author papers is usually much smaller as publishing a first author paper requires much more effort. The directed scientific collaboration network also has significant negative (in, in) degree correlation. It means that young scientists (e.g. small indegree) tend to collaborate with more senior scientists (e.g. large indegree). As pointed out by recent literatures, this tendency is more likely to result in a successful career of young scholars (Qi et al. 2017; Amjad et al. 2017).

We also apply several classical ranking methods such as betweenness centrality, PageRank and SIR model to undirected and directed collaboration networks to rank scholars. We can observe that the ranking results for these two networks show obvious differences especially in PageRank and SIR. This finding shows the influence of different network structure on the identifying important scholars. In addition, we apply the HITS algorithm with authority and hub scores to rank scientists in the directed collaboration network. Interestingly, we find that the scholars with high hub scores are mainly well-known experts in the field of complex systems such as H. Eugene Stanley and Shlomo Havlin. The scholars with high authority scores, on the other hand, are young scholars with remarkable potentials. Therefore, ranking scientists with directed collaboration networks provides information which cannot hardly observed in undirected networks.

The contribution of directed collaboration networks is twofold relate to the standing literature. Firstly, it is well-known that the undirected collaboration networks has high clustering coefficient (Newman 2001) and positive degree correlation (Newman 2002). However, one can observe a completely different properties in directed collaboration networks, i.e. a much lower clustering coefficient and negative degree correlation. These results indicates that directed collaboration networks provides a more detailed understanding of the collaboration data, complementary to the traditional undirected network modeling. Secondly, the top ranked scientists in the undirected collaboration networks mainly include the already outstanding scientists (Yan and Ding 2009). The directed collaboration networks instead can simultaneously rank scientists based on both influence and potential, offering a way to identify promising young scientists. Therefore, we believe that studying the directed collaboration network shows us a more complete picture of the scientific collaboration behavior.

It is undeniable that there exist many limitations related to the data and methodology in our study. On one hand, the size of our data is not big enough, which it may affect our analytical results. On the other hand, we should perform similar analysis on other more fields and investigate the common features of collaboration patterns of other fields. In our study, we only highlight the importance of first author and reduce the effect of other coauthors by means of the directed edges in the collaboration network. Of course, if we want to highlight the important position of corresponding authors, we can conduct a similar analysis. In future works, we will combine different algorithms of credit allocation and directed network to explore whether these methods have a significant effect on the rank of scientists and extend the directed scientific collaboration network to directed and weighted network.

Acknowledgements This work is supported by the National Natural Science Foundation of China (Grant Nos. 61374175, 61573065 and 61603046) and the Natural Science Foundation of Beijing (Grant No. L160008).

References

Amjad, T., Ding, Y., Xu, J., Zhang, C., Daud, A., Tang, J., et al. (2017). Standing on the shoulders of giants. *Journal of Informetrics*, *11*(1), 307–323.

Barabási, A. L., Jeong, H., Nda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, *311*(3), 590–614.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117.

Colizza, V., Flammini, A., Serrano, M. A., & Vespignani, A. (2006). Detecting rich-club ordering in complex networks. *Nature Physics*, *2*(2), 110–115.

Ding, Y. (2011). Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of Informetrics*, *5*(1), 187–203.

Ebadi, A., & Schiffauerova, A. (2015). How to become an important player in scientific collaboration networks? *Journal of Informetrics*, *9*(4), 809–825.

Evans, T. S., Lambiotte, R., & Panzarasa, P. (2011). Community structure and patterns of scientific collaboration in business and management. *Scientometrics*, *89*(1), 381–396.

Fagiolo, G. (2007). Clustering in complex directed networks. *Physical Review E*, *76*(2), 026107.

Fan, Y., Li, M., Chen, J., Gao, L., Di, Z., & Wu, J. (2004). Network of econophysicists: A weighted network to investigate the development of econophysics. *International Journal of Modern Physics B*, *18*(17n19), 2505–2511.

Foster, J. G., Foster, D. V., Grassberger, P., & Paczuski, M. (2010). Edge direction and the structure of networks. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(24), 10815–10820.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(12), 7821–7826.

Gleich, D. F. (2015). PageRank beyond the Web. *SIAM Review*, *57*(3), 321–363.

Hou, H., Kretschmer, H., & Liu, Z. (2007). The structure of scientific collaboration networks in Scientometrics. *Scientometrics*, *75*(2), 189–202.

Kim, J., & Diesner, J. (2015). Coauthorship networks: A directed network approach considering the order and number of coauthors. *Journal of the Association for Information Science and Technology*, *66*(12), 2685–2696.

Kintali, S. (2008). Betweenness centrality: Algorithms and lower bounds. arXiv:0809.1906.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*(5), 604–632.

Leicht, E. A., & Newman, M. E. (2008). Community structure in directed networks. *Physical Review Letters*, *100*(11), 118703.

Li, M., Fan, Y., Chen, J., Gao, L., Di, Z., & Wu, J. (2005). Weighted networks of scientific communication: The measurement and topological role of weight. *Physica A: Statistical Mechanics and its Applications*, *350*(2), 643–656.

Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information processing and management*, *41*(6), 1462–1480.

Liu, P., & Xia, H. (2015). Structure and evolution of co-authorship network in an interdisciplinary research field. *Scientometrics*, *103*(1), 101–134.

Lü, L., Chen, D., Ren, X. L., Zhang, Q. M., Zhang, Y. C., & Zhou, T. (2016). Vital nodes identification in complex networks. *Physics Reports*, *650*, 1–63.

Lu, H., & Feng, Y. (2009). A measure of authors centrality in co-authorship networks based on the distribution of collaborative relationships. *Scientometrics*, *81*(2), 499.

Newman, M. E. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, *64*(1), 016131.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 404–409.

Newman, M. E. (2002). Assortative mixing in networks. *Physical Review Letters*, *89*(20), 208701.

Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113.

Opsahl, T., Colizza, V., Panzarasa, P., & Ramasco, J. J. (2008). Prominence and control: The weighted rich-club effect. *Physical Review Letters*, *101*(16), 168702.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814.

Qi, M., Zeng, A., Li, M., Fan, Y., & Di, Z. (2017). Standing on the shoulders of giants: The effect of outstanding scientists on young collaborators' careers. *Scientometrics*, *111*(3), 1839–1850.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(9), 2658–2663.

Sinatra, R., Wang, D., Deville, P., Song, C., & Barabsi, A. L. (2016). Quantifying the evolution of individual scientific impact. *Science*, *354*(6312), aaf5239.

Sonnenwald, D. H. (2007). Scientific collaboration. *Annual Review of Information Science and Technology*, *41*(1), 643–681.

Tijssen, R. J. (2004). Is the commercialisation of scientific research affecting the production of public knowledge? Global trends in the output of corporate research articles. *Research Policy*, *33*(5), 709–733.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the Association for Information Science and Technology*, *60*(10), 2107–2118.

Yan, E., & Ding, Y. (2011). Discovering author impact: A PageRank perspective. *Information Processing and Management*, *47*(1), 125–134.

Yan, E., Ding, Y., & Sugimoto, C. R. (2011). P Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the Association for Information Science and Technology*, *62*(3), 467–477.

Yoshikane, F., Nozawa, T., & Tsuji, K. (2006). Comparative analysis of co-authorship networks considering authors' roles in collaboration: Differences between the theoretical and application areas. *Scientometrics*, *68*(3), 643–655.

Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., et al. (2017). The science of science: From the perspective of complex systems. *Physics Reports*, *714–715*, 1–73.

Zhai, L., Li, X., Yan, X., & Fan, W. (2014). Evolutionary analysis of collaboration networks in the field of information systems. *Scientometrics*, *101*(3), 1657–1677.