2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress

Team Recognition in Big Scholarly Data: Exploring Collaboration Intensity

Shuo Yu*, Feng Xia*, Kaiyuan Zhang*, Zhaolong Ning*, Jiaofei Zhong† and Chengfei Liu‡
*School of Software, Dalian University of Technology, China
†Department of Computer Science, California State University, East Bay, USA
‡Department of Computer Science and Software Engineering, Swinburne University of Technology, Australia

***Abstract*—The scale of scholarly data has been expanded due to the fact that scientific productions are increasing rapidly and new scholars affiliate academia incessantly. Scholars are shifting their research patterns from individual research to academic teamwork due to the complexity of scientific issues. In order to achieve higher reputations and better performance, academic teams with leaders are constructed to speed up knowledge sharing and problem solving. It is significant to explore team-based issues with the increasing interests of information exploration in big scholarly data. However, existing academic team definitions are commonly not quantitative, which makes it difficult to identify real academic teams. In this work, we propose a collaboration relationship evaluation index called Collaboration Intensity Index (CII), which is a two-way and quantitative index to evaluate collaboration intensity between two scholars in the network. Then, we construct a new type of co-author network with edges weighted by CII, which differs from the original co-author networks. This network reflects the newly scientific research patterns inside or outside academic teams. Furthermore, we propose TRAC (Team Recognition Algorithm based on CII) to identify academic teams from large co-author networks. Finally, we use DBLP data set, which contains 1,250,440 scholars and 1,575,949 published papers, to identify teams by TRAC. Comparing with fast unfolding algorithm and real team data, the effectiveness of our method can be demonstrated.**

***Index Terms*—Big scholarly data, collaboration intensity, academic team recognition, scientific collaboration.**

I. Introduction

With the advancement of modern science, scholars around the world have produced an increasing large volume of research articles, which provides rapid growing in Big Scholarly Data (BSD) [1]. BSD contains vast quantity of data associated with scholarly undertakings, such as journal articles, conference proceedings, dissertations, books, patents, presentation slides, and experimental data. Mining knowledge from BSD provides great benefits for various stakeholders, i.e., it helps scholars understand the laws of science itself, sociologists investigate researcher interactions, policy makers address knowledge and resources sharing, etc. Whereas the expanding scale of BSD brings new challenges with respect to data analysis and exploration due to its variety and complexity. Algorithms on traditional networks may not perform well with both high accuracy and low complexity in large scale networks [2], [3]. Meanwhile, The rapid development of knowledge economy and technological level has made changes on scientific research environment. Scientific research entities, such as countries, universities, research institutions, and independent researchers have paid more attentions to the efficiency and quality of scientific researches in order to achieve higher scientific reputations. There are some researches that improve subgraphs finding efficiency [4], [5]. There are also some other researches study high performance computing in large scale network from different perspectives [6], [7], while few researches consider the particularity of social and collaborative networks.

Half past century has witnessed great changes in research patterns, where teamwork has been recognized as a more efficient way to pursue scientific knowledge than individual combat [8]. Moreover, teams are proved to produce higher impact science than individual researchers [9]. The reasons leading to this phenomenon are manifold: (1) Though collaboration appears with scientific specialization, systematic interdisciplinary teams are still needed to solve complex scientific problems. With the increasing complexity of scientific problems, basic scientific research work requires more scientists with professional knowledge [10]; (2) Funding agencies are more inclined to academic teams due to the fact that proper collaboration has a significant impact on research quality [11]; (3) Scientists involved in team research have been proved to improve productivity, reduce errors, and achieve higher reputations. Under these circumstances, administrative meacertains (e.g. establish the advanced infrastructure) are taken to improve reputations, efficiency, and quality of scientific research. Scientific teams are organized with leaders in modern science as their cores, which indicates that scientific collaboration and teamwork have become a latent scientific rule instead of expectation [12]. In view of the above reasons, it is significant to explore inner patterns of teamwork, especially in the event that it has already played a key role in scientific research.

A variety of disciplines research on academic teams from different perspectives. Team-based researches contain many research issues like team composition, team formation, leadership and team performance evaluation, credit allocation, team dynamics, etc. A branch of science called Team Science comes up. Börner *et al.* proposed a multi-level, mixed-method approach for Team Science [13]. Unfortunately, most current studies on academic teamwork are stuck by lacking academic team data sets. Staša proposed a "core+extended" team model based on article data sets, and then explored that large team (10-1,000) size changing with time corresponds to a power-law tail [14]. Some other researches took similar methods to avoid

978-1-5386-1956-8/17 $31.00 © 2017 IEEE
DOI 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.155

the lack of real academic team data sets [8], [9]. Article data sets that are generated to explore team properties definitely can achieve direct and meaningful conclusions. However, it is more efficient and accurate to use data sets of teams or labs while existing team data sets are in small scale. In addition, academic team is dynamic just like other kind of teams. In order to ensure the accuracy of academic team data sets, high frequency update is necessary. That means, even if there exist such data sets (by collection recognition), it still requires an amount of efforts to maintain it. Some works put efforts on designing an effective academic team recognition algorithm by teamwork output. However, there is neither a clear, unified and quantitative concept, nor an effective recognition method of academic team despite the organization of academic team has been widespread.

In this work, academic team is redefined quantitatively as a group of scientists, who collaborate with each other with high collaboration intensity for common research purpose, knowledge exchange, and resource sharing. Academic teams are the clusters of scientists, who are in deeper and closer collaboration relationships with each other. To quantify how tight the tie (i.e. collaboration relationship) is, we propose a quantitative index called Collaboration Intensity Index (CII) to reflect real collaboration relationship between scholars. CII is a relative metric, which will be described in detail in Section 2. It is verified that this index is effective by our work. This index is proposed to evaluate scientific collaboration between scholars. We then build a multi-weight co-author network with several collaboration indices, including collaboration intensity. This co-author network reflects real collaboration between scholars, and it helps identify academic teams. Then we propose a team recognition algorithm named Team Recognition Algorithm based on CII (TRAC) to identify academic teams, in which members work together as a team in reality. Finally, we analyze 1,575,949 papers from DBLP and build co-author networks by 1,250,440 scholars. We find that: (1) CII is an effective index in scientific collaboration evaluation; (2) Real academic teams, especially high performance teams are composed with scientists, who collaborate with high collaboration intensity; (3) TRAC provides a new perspective on exploring relationships in BSD. We make the following contributions in this work.

- **The definition of CII:** We propose CII to evaluate scientific collaborations. Then we redefine academic team by this index quantitatively. This index is proposed to evaluate collaboration relationship between scholars. The effectiveness of this index is demonstrated by our experiments.
- **The establishment of weighted co-author network:** We establish a weighted co-author network based on several collaboration evaluation indices, including CII, collaboration frequency, and partnership ability index. This dynamic network reflects real collaboration relationship in academia.
- **The design of team recognition algorithm TRAC:** Based on CII, we propose a high efficiency team recognition algorithm called TRAC. This algorithm identifies academic teams in which members are in real scientific collaboration relationships with each other.
- **The recognition of teams in large scale network:** We implement TRAC in a large scale network, which is generated by 1,250,440 scholars and 1,575,949 papers in DBLP data set. Recognition results are of high accuracy comparing with real academic teams.

The rest of this paper is organized as follows. Section II introduces research background and related work. Section III describes several collaboration evaluation indices, especially the proposed CII. Academic team recognition algorithm, i.e., TRAC is introduced in Section IV. Section V implements the experiments and analyses results. Section VI concludes the paper.

II. Related Work

Academic teams are recognized as project groups, research groups, lab teams, or article teams at most time [15], [16]. Zulueta $et\ al.$ defined an academic team as a scientist community, where scientists share research methods, materials and financial resources but is not necessarily organized in an institution with fixed structure [12]. However, scholars in the same institution may belong to different academic teams while members in different institutions may belong to one same academic team in real science. One member may also belong to different institutions or teams at the same time. There also exists situation that members never collaborate with each other in the same team as well. However, such situations are hard to be defined in a quantitative way, which makes it difficult to identify academic teams.

Social network analysis can change this situation to some degree by means of constructing co-author networks. Many studies use social network analysis to define academic teams. Newman was one of the first to build collaboration networks based on bibliographical databases [17]. Calero $et\ al.$ presented a new bibliometric approach to identify research groups in a specific research field with a combination of bibliometric mapping techniques and network analysis [18]. Du $et\ al.$ came up with a community detection algorithm based on social networks, which can be used to identify academic team leaders by allowing community overlap [19]. There are also some other studies identifying academic teams by the similarity of scientific publications' titles, structures, and contents [2], [20]. In these co-author networks, nodes represent scientists and edges represent collaboration relationships which are always evaluated by collaboration frequency. However, collaboration frequency neglects the reciprocity of collaboration and may miss some team members. Besides, previous definitions may neglect the fact that academic teams are in different scales. In a word, previous academic team definitions lack a quantitative description of real collaboration relationships, which makes it difficult to identify academic teams efficiently.

III. Scientific Collaboration Evaluation

In order to identify academic teams by co-author networks, collaboration relationship evaluation is in critical need firstly. With the popularization of the scientific teamwork pattern, scientific collaboration evaluation causes attention.

A. Collaboration Indices

There are some widely used collaboration evaluation indices in scientometrics which will be introduced. In the following equations, j is the number of co-authors in one single article. f_j is the number of papers co-authored by j co-authors within one same team or institution. q is the maximum co-author number of one single article in team or institution. n is the total number of members in one team or institution. N is the total number of papers published by this team or institution.

(1) Collaborative Frequency (CF)

CF is one of the most widely used indices. CF refers to the number of papers that two co-authors are in co-author networks.

(2) Collaborative Index (CI)

CI refers to the average number of co-authors within an academic team or institution, which can be defined as follows,

$$CI = \frac{\sum_{j=1}^{q} j f_j}{N} \tag{1}$$

CI is easily computable, but not reasonable as a collaboration indice due to the fact that it has no upper limit. Besides, CI gives a non-zero weight to single-authored papers that involve no collaboration.

(3) Degree of Collaboration (DC)

DC refers to the percentage of co-authored papers within an academic team or institution, which is defined as follows,

$$DC = 1 - \frac{f_1}{N} \tag{2}$$

Obviously, a higher DC of academic team refers to a closer and intenser scientific collaboration relationship within this team.

(4) Revised Collaborative Coefficient (RCC)

Based on CI and DC, Egghe proposed the factor RCC [21], which can be calculated according to (3). RCC yields 1 when the collaboration is maximal and also be able to distinguish different situations of collaborations.

$$RCC = \frac{n}{n-1}\{1 - \frac{\sum_{j=1}^{q} \frac{1}{j} f_j}{N}\} \tag{3}$$

(5) Partnership Ability Index (PHI)

Schubert used h-index for reference and then proposed PHI, where numbers of co-authors and CF were taken into consideration [22]. A scholar owning a partnership ability index φ denotes that there are at least φ scholars collaborated with him/her at least φ times, and the rest $(n - \varphi)$ scholars collaborated with him/her less than φ times.

One's PHI reflects his/her collaboration relationship with his/her core collaborators. A higher φ means that the scholar keeps a much more wide and stable collaboration relationship with others. A lower φ represents loose and unstable collaboration relationship. PHI is vital to scientific collaboration evaluation. A scholar with $\varphi = 0$ means all of papers are published by the scholar him/herself. And a scholar with $\varphi = 1$ means this scholar only collaborates with one same scholar or this scholar collaborates with others each for one time. Thus, in this paper, we regard scholars with $\varphi > 1$ as ones who are of high active collaboration behavior. This is because that the number of collaborators and collaboration frequency are taken into account in PHI. Moreover, PHI reflects the stability of scientific collaboration to some degree.

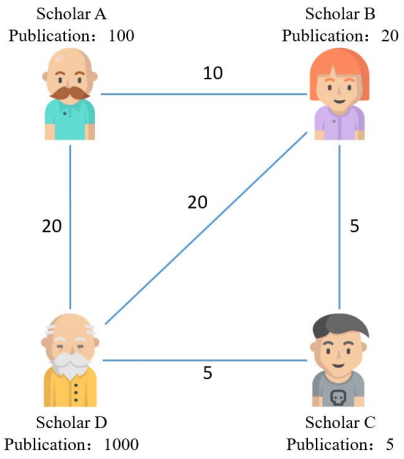


Fig. 1. An example of co-author network with edges weighted by CF.

Each of the indices evaluates scientific collaboration from different perspectives. However, collaboration is a two-way behavior. It is crucial to evaluate collaboration of both scholars in a collaborative relationship. Take the relationship in Figure 1 as an example. Scholars A and D collaborate for the same times as Scholars B and D. That means CF_{AD}, i.e., the CF between A and D, equals to CF_{BD}, i.e., CF between B and D. However, the ratio of A collaborated with D is less than that of B collaborated with D. That is, the collaboration relationship between A and D differs from that of B and D though $CF_{AD} = CF_{BD}$. Apparently, it is unfair to use CF to evaluate scholars' collaboration relationship due to the fact that scholars who published less papers may be in a strong collaboration relationship with each other. In another way, two scholars collaborate with a high CF may not refer to a high strength collaboration relationship. To evaluate the two-way collaboration behavior, we propose an index called CII.

B. Collaboration Intensity Index

We take into account the CF and the number of papers two scholars published when evaluating collaboration relationship between two scholars. The CII_{ij} of two scholars i and j who collaborated from year t_1 to year t_2 is defined in Equation 4,

$$CII = \frac{\Delta_{t_2-t_1} k_{ij}^2}{\Delta_{t_2-t_1} k_i \Delta_{t_2-t_1} k_j} \tag{4}$$

Herein, $\Delta_{t_2-t_1}k_i$ is the number of papers published by scholar i from year t_1 to year t_2 and $\Delta_{t_2-t_1}k_j$ is the counterpart of j. $\Delta_{t_2-t_1}k_{ij}$ is the number of papers co-authored by scholars i and j from year t_1 to year t_2.

CII is a relative, efficient and quantitative index in evaluating collaboration relationship. Two scholars collaborated with a higher CII refers that they are in a more intense collaboration relationship. Similarly, a lower CII refers a looser one. CII is relative since it is calculated based on the ratio of two scholars' numbers of publications. Take the collaboration relationship in Figure 1 as an example. The collaboration intensity of A and D (i.e., CII_{AD}) is 0.004, and the collaboration intensity of A and B (i.e., CII_{AB}) is 0.05. This indicates that A and B are in a closer collaboration relationship than A and D though $CF_{AB} > CF_{AD}$ in the view of A. This is because D published 1,000 papers, which is much more than B published. 20 times of collaboration may be significant to A, since it is a big percentage in A's collaboration relationships. To D on the contrary, 20 times may occupy a small percentage instead since D's other collaboration relationship is not shown in this figure. In the view of B, C and D, $CII_{BA} = CII_{AB} = 0.05$, $CII_{BC} = CII_{CB} = 0.25$, $CII_{BD} = CII_{DB} = 0.02$, $CII_{CD} = CII_{DC} = 0.005$. We can see that though C has collaborated with D for 5 times, CII_{CD} is even higher than CII_{AD}, while A has collaborated with D for 20 times. That is, CII can recognize whether two scholars are in a closer collaboration relationship or not by calculating CII of all collaboration relationships.

Academic team is a group of members who collaborated with each other with relatively higher CII. However, the specific value of CII, which is used to judge whether a scholar belongs to an academic team, needs to be calculated by real data in co-author networks.

To recognize academic teams from co-author networks, we need to build a new co-author network, where edge weight is CII. In the new co-author network, different nodes represent different scholars, and different edges represent different collaboration relationships.

IV. Team Recognition Algorithm Based on CII

Problem Definition: To a given $G = (V, E)$ and relationship constraint coefficient ω, there exists a subgraph $T \subseteq G$ such that for any two nodes $v, v' \in G$, $dis(v, v') \leq \omega$. $dis(v_1, v_2)$ is a function which can calculate the weight of edge between v_1 and v_2.

The basic idea of TRAC is shown in Figure 2. We first generate a subset T of scholars and then construct the new network, whose edge weight is CII. By the scientific relationship constraint coefficient ω, we cut off the edges with looser CII. The main advantages of TRAC are shown as below: (1) The algorithm achieves objective experiment results by co-author networks, which avoids subjectivity of other recognition methods like social investigations or peer reviews to some degree. TRAC uses publications to construct co-author networks, which combines objective data and objective algorithm to achieve objective results. (2) The iconicity of experiment results contains information, such as team structure, relationship between nodes, and team component. (3) TRAC is a real time team recognition algorithm that calculates ω in different time periods. This is critical to team based research since team is dynamic.

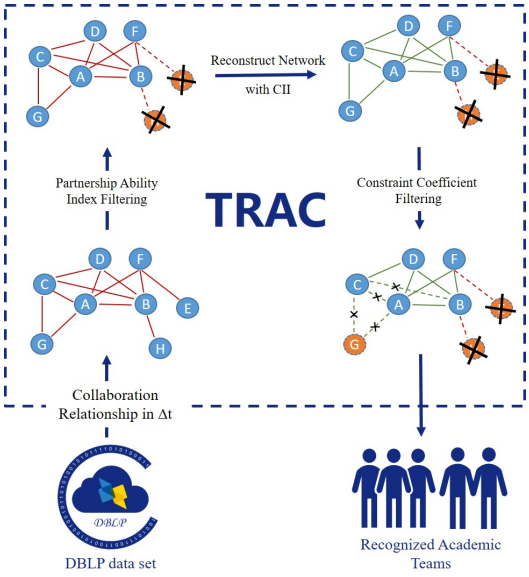


Fig. 2. The framework of TRAC.

TRAC identifies teams by mainly three processes. Step 1 calculates PHI of each nodes and filtrates nodes by PHI in a co-authorship graph (i.e., co-author network). Step 2 calculates CII between nodes, and then generates new network, where edge weight is CII. Step 3 identifies academic teams by filtrating edges with CII and ω. Now these steps will be introduced in detail as follows.

Algorithm 1 Network Filtering

Input:
$G = ((V, VWeight), (E, EWeight))$, which is the original co-authorship graph with node weight (i.e., number of published papers) and edge weight (number of co-authored papers);

Output:
$G_1 = ((V_1, VWeight_1), (E_1, EWeight_1))$, which is the filtered co-authorship network with node weight (i.e., number of published papers) and edge weight (number of co-authored papers);

1: **For** V_i in V
2: Calculate $PHI(V_i)$ according to co-authorship of V_i;
3: **If** $PHI(V_i)$ is less than 1
4: Remove (V_i, edge of V_i);
5: **Return** $G1$;

A. Step 1: PHI Filtering

Scholars of academic teams collaborate with their members in a higher frequency than those who are not in academic teams. For a given co-author network, if PHI of one node is less than 1, then we remove this node and the edges of this node. The constraint condition $PHI < 1$ is to filter those who are not active in scientific collaboration.

B. Step 2: Network Construction

The traditional co-authorship network, where edge weight is CF, fails on evaluating collaboration intensity. Thus we reconstruct the co-authorship network weighted by CII. By calculating CII between different nodes, we construct the new network G_{CII}.

Algorithm 2 Network Construction

Input:

$G_1 = ((V_1, VWeight_1), (E_1, EWeight_1))$, which is the filtered co-authorship network with node weight and edge weight;

Output:

$G_{CII} = ((V_{CII}, VWeight_{CII}), (E_{CII}, EWeight_{CII}))$, which is the co-authorship network with edge weighted by CII;

1: **For** scientist n_i in G_1
2: **For** its neighbor n_j in G_1;
3: $EWeight_{CII}(n_i, n_j) = EWeight_1^2(n_i, n_j)/(VWeight_1(n_i) * VWeight_1(n_j))$;
4: **Return** G_{CII}

C. Step 3: Team Recognition

In this process, we use scientific relationship constraint coefficient ω to restrict and filter academic teams. Academic teams are identified by removing edges and nodes. If edge weight is less than ω, we remove this edge. ω is the minimum edge weight of the top 20% edge weights, which ranks from the largest to the smallest in the network.

Algorithm 3 Team Recognition based on CII.

Input:

$G_{CII} = ((V_{CII}, VWeight_{CII}), (E_{CII}, EWeight_{CII}))$, which is the co-authorship network with edge weighted by CII; ω, which is scientific relationship constraint coefficient;

Output:

G_{new}, which contains all of the recognized academic teams;

1: $G_{new} = G_{CII}$;
2: **For** scientist n_i in G_{new}
3: **For** its neighbor n_j in G_{new}
4: **If** $G_{new}.EWeight(n_i, n_j) < \omega$
5: Remove $G_{new}.E$;
6: Remove isolated edges and nodes;
7: **Return** G_{new};

V. Experiments and Result Analysis

In this work, we implement our experiments on DBLP[1] data set, which is one of the most influential and largest scholarly data sets in Computer Science. DBLP provides open bibliographic information with more than 3,700,000 papers and 1,900,000 authors, and updates in a high frequency. In this work, we first preprocess the whole data set by filtering those papers published by sole author to reduce the complexity of the generated network and accelerate the algorithm efficiency. We use a volume of 1,250,440 scholars and 1,575,949 papers published from 2009 to 2017 in our experiments.

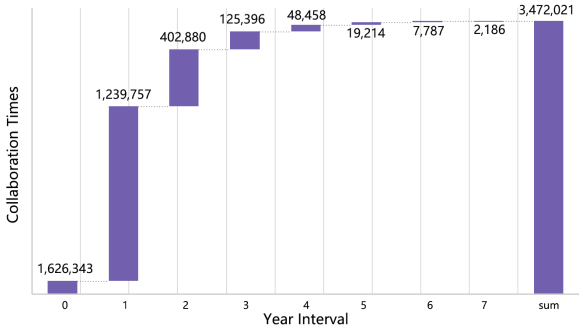


Fig. 3. The year interval of every two scholars collaborate successively in co-author network.

A long period collaboration helps to identify academic teams accurately. We analyze scholars' collaboration behaviors and find that two scientists rarely collaborate again after 5 years, which is shown in Figure 3. The horizontal axis is the year interval that every two scholars collaborate again in co-author network. The vertical axis is the times that every two scholars collaborate in co-author network. It can be seen from Figure 3 that collaboration times are less than ten thousand times when year interval is larger than 5, which occupies less than 0.29% of the total collaboration times. Therefore, we use 5 years as a whole collaboration period to identify academic teams, i.e., $m = 5$. That means, we only explore scientific collaboration relationship among scholars during 5 years period. Due to the statistics of DBLP, there is an obviously increasing collaboration trend since 2009. Then co-author networks are generated every 5 years, i.e., 2009 to 2013, 2010 to 2014, 2011 to 2015, 2012 to 2016, and 2013 to 2017. Scientific relationship constraint coefficient ω is significant for academic team recognition. Dunbar pointed out that human can only maintain a social network with 148 people, rounded to 150. Within the network, 20% of relationships are in strong ties with each other and 80% of them are in weak ties [23]. It is worth attention that the trend of scientific collaboration differs with time, which makes scientific relationship constraint coefficient ω differ with time as well. Moreover, academic teams are small organizations that members are generally in strong ties with each other, which

[1] http://dblp.uni-trier.de/

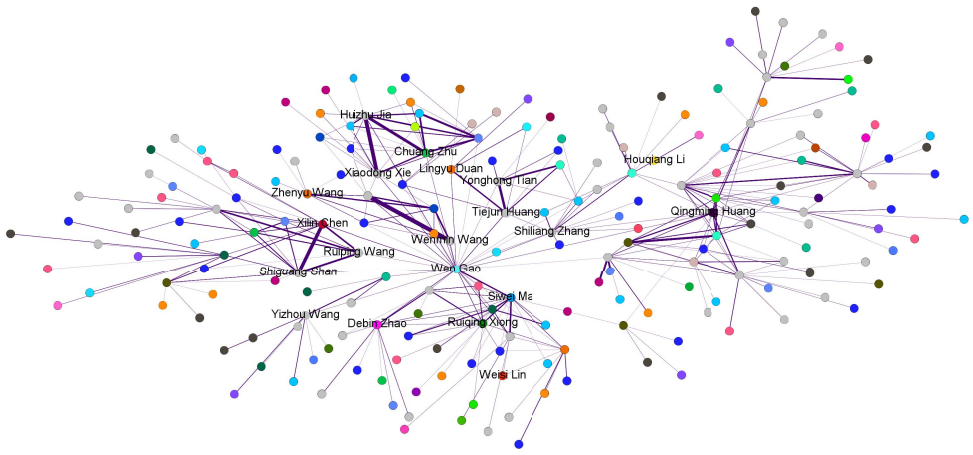


Fig. 4. The academic team leaded by Wen Gao identified by TRAC. There are 178 members in the identification team.

produce higher quality academic outputs. Thus we calculate ω according to the ratio mentioned above.

A. Academic co-author network analysis

In the co-author networks we generated, nodes represent scholars, and edges represent collaboration relationships. The thickness of edges represents the value of CII. The higher the CII value is, the thicker the edge is, which can be seen from Figure 4.

From an overall perspective of the co-author network, there is a trend that the connectivity among scholars diverges from center margin in the network. Meanwhile, there is extra high connectivity in the center of co-author network and most of the nodes in center are connected with each other. Such connectivity elucidates that there exists "small world" phenomenon in co-author network. Besides, the connection is conducted by some "bridge scholars", where two nodes may be connected through these bridge scholars. In our work, bridge scholars refer to scholars (who occupy important positions in teams) or team leaders who make new connections with scholars (who occupy important positions in teams) or team leaders. The existence of bridge scholars tides the ties of academic network, which makes clusters of network gather closely. Ulteriorly, the existence of bridge scholars also reflects collaboration scope of scholars. Collaboration behavior has already expanded towards outer teams.

There are many academic teams pervading around the margin of the co-author network. These subgraphs are in a disperse distribution and smaller scale comparing to those in the center, which indicates that though collaborations spread, collaborations still need to be scaled up. Academic teams of larger scale generally perform better. By comparing with co-author networks of different periods, it can be found that scholars are tend to be a two-way collaboration in team contribution. The increasing number of teams reflects the widespread of academic teamwork in scientific model. More scholars intend to choose academic teams instead of solo working in order to improve their outputs, influences, and reputations.

By analyzing the structure of co-author network, some interesting phenomena are discovered in large scale (more than 4 nodes) academic teams (subgraphs with high CII). Nodes with higher CII relations form core academic teams. Like cohesive subgroups in social networks, core academic teams are explored in different dimensions in order to discover inner patterns. Core academic teams can also be used to find potential collaboration relationships to set up new academic teams.

In large scale academic teams, there always exists several core academic teams. This can be seen directly from the graphs since thickness of edges are different. Large scale teams generally contain several core teams, which will be discussed in Section V-C. Core teams can be regarded as basic components of team structure. These collaboration relationships are significant to team contribution. Such situations exist generally in universities since professors are core academic teams and the rest nodes in teams are students.

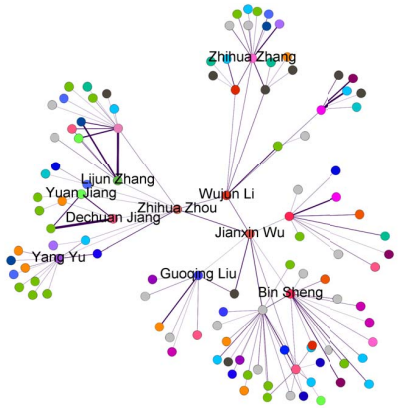
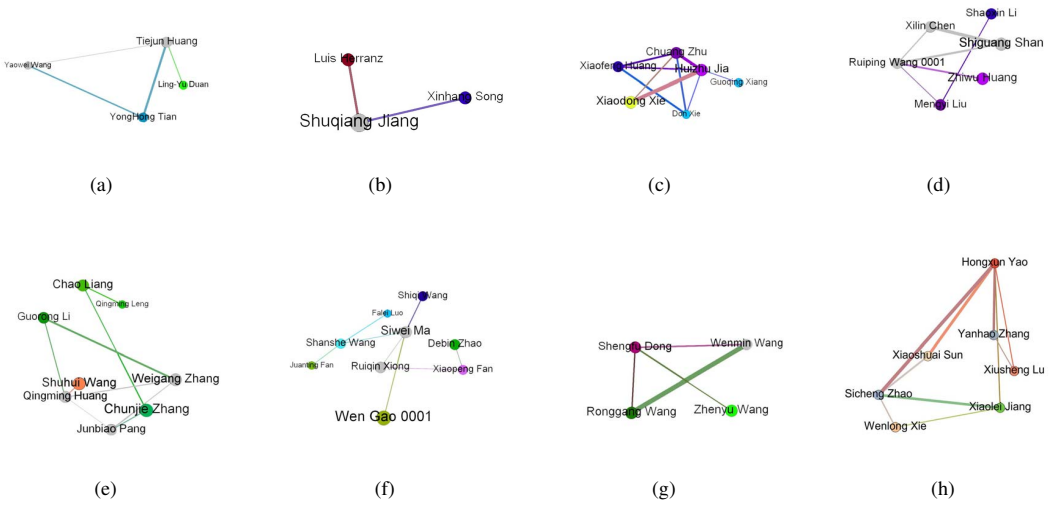


Fig. 5. The academic team leaded by Zhihua Zhou identified by TRAC. There are 87 members in identification teams and 115 members in real team.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 6. The core teams of Wen Gao's team. The average indices are list respectively as follows: (a) average CII=53.51, average publications=43.5, average h-index=10.5; (b) average CII=62.67, average publications=22.33, average h-index=7.67; (c) average CII=64.98, average publications=15.8, average h-index=2; (d) average CII=57.49, average publications=41.5, average h-index=7.5; (e) average CII=40.28, average publications=41.75, average h-index=7.5; (f) average CII=35.5, average publications=65.67, average h-index=11.22; (g) average CII=69.46, average publications=35, average h-index=4; (h) average CII=55.5, average publications=42.2, average h-index=6.14.

Each specific team can be detailedly analyzed to explore changes of collaboration relationships inner teams. Take the team leaded by Wen Gao as an example, and the network structure of his team is shown in Figure 4. We can see that the co-author network contains 218 nodes and 932 edges, with average degree of 9.55 and average weighted degree of 41.78. The average CII of Wen Gao's team is 7.43351. The statistics of average degree and average weighted degree indicates that collaboration behavior becomes complicated and scholars are intended to tide the ties of inner team.

B. Academic Team Recognition

In this paper, by defining CII, improving PHI, and analyzing big scholarly data set DBLP, our experiment results can objectively describe the composition and collaboration inside teams. Due to the lack of team recognition algorithms, we use one of the most popular community detection algorithm, i.e., fast unfolding algorithm [24] for comparison. As a classical community detection algorithm, fast unfolding algorithm uses the modularity as measurement to detect communities. And then we use real team to verify and evaluate the above two algorithms.

We use two real academic teams, which are leaded by Wen Gao[2] and Zhihua Zhou[3] respectively, to verify the accuracy of TRAC. For the team leaded by Wen Gao, TRAC identifies 178 members within the team. The total number of team members is 218. The 40 unrecognized members are edge nodes of the team network. The error is caused by team member dynamic during these years. Most of edge nodes are master degree candidates or doctoral candidates, while the core academic team is formed by competent scholars from Peking University or Chinese Academy of Sciences. The identification team is shown in Figure 4. It can be seen that Wen Gao is not only a leadership node within this team, but also a bridge node of several subgroups. There are many core teams whose members are of higher CII with each other.

[2]http://www.jdl.ac.cn/htm-gaowen/en_index.htm
[3]https://cs.nju.edu.cn/zhouzh

C. Collaboration Intensity Index Analysis

In the team leaded by Zhihua Zhou, 87 of 115 members are identified by TRAC. The 28 unrecognized members are all student members and have graduated from university, which is difficult for identification by co-author networks. Core members of this team are of low CII, which indicates that the whole collaboration relationship is relative loose. It can be seen from Figure 5 that there are mainly three subgroups of this team, where each of the subgroups is leaded by one core member. This is because there are several research interests within this team, and each research interest may be leaded by one core member.

TRAC totally identifies 44,376 teams from the co-author network during 2013 to 2017. Fast unfolding identifies 4,498 communities, which performs worse than TRAC. The final modularity is 0.771. The academic teams indentified by fast unfolding are in extra large scale. Identification result of fast unfolding contains not only Wen Gao's team but also Zhihua Zhou's team. Similar mixed results are identified as well. This is because that academic teams differ from communities. At the same time, fast unfolding may regard scholars who work independently as members of academic teams, which expands team scale to some degree.

Compared with fast unfolding based on CF, TRAC recognizes more academic teams than fast unfolding does. Results of TRAC are closer to real team data with higher accuracy. By TRAC, we recognize 44,125 teams during 2009 to 2013 and 58,278 teams during 2012-2016. Academic teams keep growing with a high rate since 2009. This indicates that scientific research is shifting its way to a new pattern and scholars need to adapt to this change. Teamwork has been regarded as one of the most popular ways to figure out scientific issues, especially interdisciplinary problems.

We use the identification result of Wen Gao to verify the efficiency of CII and TRAC. In order to explore collaboration relationship between core members, we limit the relationship constraint coefficient ω to the top 20% in the team that we have recognized in Figure 4. Then we get the core teams of Wen Gao's teams in Figure 6. In many cases, members of large scale academic teams do not usually work all together. They may work in several groups instead. By limiting ω to the top 20%, we get 8 core teams totally. Each core team contains no more than 8 members. Among these teams, core team in Figure 6(f) owns the highest average h-index=11.22, whereas its average CII is the lowest. Likewise, core team in Figure 6(g) owns the highest CII, whereas its average h-index is far less than other core teams. Scholars with high reputation are generally in complex collaboration relationships, which makes it difficult to recognize their academic teams. However, teams which are composed of high reputation scholars can be recognized accurately with TRAC. Meanwhile, teams that members actually collaborated with each other can be recognized as well. This indicates that CII is an efficient index that can be used in team recognition.

VI. Conclusion

Scholars are shifting their research patterns from individual research to academic teamwork due to the importance of teamwork. Meanwhile, many researchers have realized the importance of using BSD to understand relationships of academic teamwork. However, existing team definitions are commonly neither unified nor quantitative. Under this circumstance, we first propose a collaboration relationship evaluation index named CII to evaluate collaboration intensity, which is also a new feature that can improve understanding of collaboration relationship in BSD. Then based on CII, we propose an academic team recognition algorithm named TRAC to identify academic teams. TRAC is also a new perspective to mine dynamic subgraphs in BSD networks. Compared with classical community detection algorithm fast unfolding, TRAC performs better in recognition accuracy, which indicates that the proposed CII is effective in both collaboration relationship evaluation and academic team recognition. We use real team data to verify the experiment results and find that errors are tolerable, which are mainly caused by team dynamic. This corresponds to the fact that academic teams tide the ties of scientific collaboration. Our approach provides a new method, which can be used to identify weak or strong collaboration relationships and real academic teams.

References

[1] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big scholarly data: A survey," *IEEE Transactions on Big Data*, vol. 3, no. 1, pp. 18–35, 2017.

[2] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, 2017.

[3] E. Stai, V. Karyotis, and S. Papavassiliou, "A hyperbolic space analytics framework for big network data and their applications," *IEEE Network*, vol. 30, no. 1, pp. 11–17, 2016.

[4] R. Zhou, C. Liu, J. X. Yu, W. Liang, B. Chen, and J. Li, "Finding maximal k-edge-connected subgraphs from a large graph," in *International Conference on Extending Database Technology*, 2012, pp. 480–491.

[5] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, 2016.

[6] R. Han, L. K. John, and J. Zhan, "Benchmarking big data systems: A review," *IEEE Transactions on Services Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[7] C. A. Ardagna, P. Ceravolo, and E. Damiani, "Big data analytics as-a-service: Issues and challenges," in *IEEE International Conference on Big Data*, 2016, pp. 3638–3644.

[8] S. Wuchty, B. F. Jones, and B. Uzzi, "The increasing dominance of teams in production of knowledge," *Science*, vol. 316, no. 5827, pp. 1036–1039, 2007.

[9] B. F. Jones, S. Wuchty, and B. Uzzi, "Multi-university research teams: Shifting impact, geography, and stratification in science," *Science*, vol. 322, no. 5905, pp. 1259–1262, 2008.

[10] N. Hara, P. Solomon, S.-L. Kim, and D. H. Sonnenwald, "An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration," *Journal of the American Society for Information science and Technology*, vol. 54, no. 10, pp. 952–965, 2003.

[11] W. Wang, S. Yu, T. M. Bekele, X. Kong, and F. Xia, "Scientific collaboration patterns vary with scholars academic ages," *Scientometrics*, pp. 1–15, 2017.

[12] L. Reyes-Gonzalez, C. N. Gonzalez-Brambila, and F. Veloso, "Using co-authorship and citation analysis to identify research groups: a new way to assess performance," *Scientometrics*, vol. 108, no. 3, pp. 1171–1191, 2016.

[13] K. Börner, N. Contractor, H. J. Falk-Krzesinski, S. M. Fiore, K. L. Hall, J. Keyton, B. Spring, D. Stokols, W. Trochim, and B. Uzzi, "A multi-level systems perspective for the science of team science," *Science Translational Medicine*, vol. 2, no. 49, pp. 49cm24–49cm24, 2010.

[14] S. Milojević, "Principles of scientific research team formation and evolution," *Proceedings of the National Academy of Sciences*, vol. 111, no. 11, pp. 3984–3989, 2014.

[15] F. Xia, X. Su, W. Wang, C. Zhang, Z. Ning, and I. Lee, "Bibliographic analysis of nature based on twitter and facebook altmetrics data," *PloS one*, vol. 11, no. 12, p. e0165997, 2016.

[16] X. Bai, F. Xia, I. Lee, J. Zhang, and Z. Ning, "Identifying anomalous citations for objective evaluation of scholarly article impact," *PloS one*, vol. 11, no. 9, p. e0162364, 2016.

[17] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[18] C. Calero, R. Buter, C. Cabello Valdés, and E. Noyons, "How to identify research groups using publication analysis: an example in the field of nanotechnology," *Scientometrics*, vol. 66, no. 2, pp. 365–376, 2006.

[19] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[20] S. E. G. Villarreal and S. E. Schaeffer, "Local bilateral clustering for identifying research topics and groups from bibliographical data," *Knowledge and Information Systems*, vol. 48, no. 1, pp. 179–199, 2016.

[21] C. H. Liao and H. R. Yen, "Quantifying the degree of research collaboration: A comparative study of collaborative measures," *Journal of Informetrics*, vol. 6, no. 1, pp. 27–33, 2012.

[22] A. Schubert, "A hirsch-type index of co-author partnership ability," *Scientometrics*, vol. 91, no. 1, pp. 303–308, 2012.

[23] R. Dunbar, *How many friends does one person need?: Dunbar's number and other evolutionary quirks*. Faber & Faber, 2010.

[24] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.