STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

A research team identification method based on relationship analysis[1]

Ma Tingcan*, Li Ruinan*, Ou Guiyan* and Yue Mingliang*

*matc@whlib.ac.cn; lirn@whlib.ac.cn; ougy@whlib.ac.cn; yueml@whlib.ac.cn
Wuhan Documentation and Information Center, Chinese Academy of Sciences, Wuhan (China)

Introduction

As one of the most fundamental units in scientific activities, research team has drawn lots of attentions in many scientific processes, e.g., scientific evaluation, discipline planning and construction, expert team formation and problem solving, etc. Consequently, finding and identifying research teams (from large sets of researchers) becomes an essential task (Yan & Ding 2010, Li et al. 2014).

Generally, research team refers to a set of closely related researchers that always cooperate with each other. Based on the notation, many team identification methods have been developed. Most of the methods take co-author network as input, identify team boundary based on network reachability, and use co-authorship time (number of co-authorships) as threshold to represents the concepts of closeness to identify core team member (Shen et al. 2013, Perianes-Rodríguez et al. 2010, Borner et al. 2010, Calero et al. 2006). Others may further consider the structure of the network by restricting the network features like network denseness, node closeness and betweenness, and so on (Ren & Zhou 2015, Li et al. 2017, Qiu & Wu, 2011).

Those methods can find actual teams in many cases, however, there are still some aspects can be improved. First, to the authors' best knowledge, when referring to closeness, only productivity is considered by the methods. The time perspective is rarely considered. However, an ordinarily recognized team should satisfy both output and time stability: a set of researchers that cooperate for considerable time duration and have considerable outputs. Second, although certain restrictions on network features (or, network statistics) are given, outliners may still exist in the result team. For example, in a considerably dense network, a team member may only connect to one other team member -- conflict with the intuition that team members should cooperate with several other members to some extent.

Focusing on the problems, in this paper, we propose a research team identification method based on relationship analysis. We first define weak/strong relationship by considering both cooperation length (years) and outputs (papers), and then define research team based on the relationships. Further, we propose an algorithm that efficiently goes through the co-author network and output the research teams. At last, we verify the method with teams identified

[1] This work is supported by the National Natural Science Foundation of China under grant (No. 71603252). Corresponding author: Yue Mingliang, yueml@whlib.ac.cn.

based on the co-author network constructed from bibliographic data of Computer Sciences from Science Citation Index Expanded (SCI-EXPANDED) database.

Research team: Definition

We aim to find such team that, in considerable time duration, each team member cooperates with several other team members and has considerable outputs based on the cooperation. We first induce the length and output restrictions in the definition of Strong/weak Relationship as follows.

Suppose we have a set of authors and their publications. Let $T_{ij}=\{t_1,t_2,\ldots,t_n\}$ be the set of publication years of the papers co-authored by author A_i and author A_j. Let $T_{\max}$ and $T_{\min}$ be the latest year and the earliest year in T_{ij} respectively. Then *Time Stable* is defined as follows.

Definition 1 Time Stable Given two thresholds $\lambda>0,\gamma\in(0,1)$, the cooperation between A_i and A_j is *Time Stabile*, if $\|T_{ij}\|\geq\lambda,\frac{\|T_{ij}\|}{T_{\max}-T_{\min}+1}\geq\gamma$, where $\|\circ\|$ denotes the cardinal number of a set.

λ here is to convey the meaning that two authors should co-author for at least λ (say 3) years, whereas γ requires that the authors should cooperate for at least γ (say 0.3) times of their cooperative career. For example, suppose $T_{ij}=\{2002,2003,2011\}$, then we can say A_i and A_j satisfy the condition of time stable if we set $\lambda=3,\gamma=0.3$, since $\|T_{ij}\|$ =3 and $\frac{3}{2011-2002+1}=0.3$.

Further, let P_i and P_j be the set of publications of A_i and A_j respectively, P_{ij} be the set of publications co-authored by A_i and A_j. We define *Output Stable* as:

Definition 2 Output Stable Given two thresholds $\tau>0,\xi\in(0,1)$, the cooperation between A_i and A_j is *Output Stabile*, if $\|P_{ij}\|\geq\tau,\frac{\|P_{ij}\|}{\max(\|P_i\|,\|P_j\|)}\geq\xi$, where $\max(\cdot,\cdot)$ returns the larger one of two numbers.

From the definition we can see, two authors are said to satisfy output stable if they co-authored at least τ papers, and the number of co-authored papers should greater than ξ times of the number of publications of either author. For example, suppose A_i published 10 papers, A_j published 15 papers, and they co-authored 6 papers, then we can say A_i and A_j satisfy the condition of output stable if we set $\tau=5,\gamma=0.3$, since $6>5,\frac{6}{\max(10,15)}=0.4>0.3$.

Definition 3 Strong/Weak Relationship Given two authors A_i and A_j who have co-authored with each other, we say A_i has Strong Relationship with A_j, if the cooperation between A_i and A_j is both Time Stable and Output Stable. On the opposite, we say A_i has Weak Relationship with A_j if the cooperation does not satisfy Strong Relationship.

Let co-authorship network denote the network with authors as nodes, Strong/Weak Relationships as links, then we have:

Definition 4 Research Team Given a set of authors $R=\{A_1,A_2,\ldots,A_t\}$, two positive integers M and N, R is a Research Team, if 1) $\forall A_i, A_j \in R$, A_i is reachable to A_j within the co-authorship network with regard to R; 2) $\forall A_i \in R, \exists R^{SW} \subseteq R, \left\|R^{SW}\right\| \geq N$, then $\forall A_j \in R^{SW}$, A_i has strong or weak relationship with A_j; *3*) $\forall A_i \in R$, $\exists R^S \subseteq R^{SW}, \left\|R^S\right\| \geq M$, then $\forall A_j \in R^S$, A_i has strong relationship with A_j.

Following definition 4, we can know that every member in a team at least has co-authored with N other team members, and has strong relationship with M other members. Next, we introduce the team identification algorithm, so that research teams can be efficiently acquired from large set of authors.

Team Identification

Suppose that the input data are tuples formatted as <*paperID*; *authorID List*; *Publication Year*>, e.g., <P_1; A_1, A_2, A_3; 2010>, which means author A_1, A_2 and A_3 co-authored paper P_1 at year 2010. Please note that the information can be very easily acquired based on bibliographic data (downloaded from SCIE database). The only problem is that authors should be carefully recognized before authorID can be assigned. We will discuss data collection and refinement process in the next section.

Algorithm 1 Team Identification

Abstract **Input:** List of Tuples $\langle paperID; authorIDList; PublicationYear\rangle$
Output: Research Teams

```
1: for each tuple in the input Tuples do
2:     Record each ⟨authorID_i; authorID_j; PublicationYears⟩
3:     Mantain a globel list of authors AUs
4: end for
5: for each author authorID_i do
6:     for each coauthor (of authorID_i), authorID_j do
7:         Determine ⟨authorID_i; authorID_j; RealtionType⟩ and Record in set RELs
8:     end for
9: end for
10: NET ← (AUs,RELs)
11: for each author authroID_i in NET do
12:     subNET ← traversal NET starting from authroID_i
13:     NET ← NET − subNET
14:     repeat
15:         for each author authroID_j in subNET do
16:             if authroID_j does not satisfy Definition 4, Condition 1 - 3 then
17:                 Exclude authroID_j and its corresponding realtionships from subNET
18:             end if
19:         end for
20:     until No change is made on subNET
21:     Record subNET in set TEAMs
22: end for
23: return TEAMs
```

The main idea of our algorithm is that, we first record the cooperation years of each and every pair of authors that have cooperation (Line 1-4, **Publication Years** is a vector of years that may have duplicate elements); then determine the relationship types of them based on Definition 3 (Line 5-9). After that, the co-author network is naturally constructed, by considering authors as network nodes and (weak or strong) relationships as network edges (Line 10). Then, starting from any author, we can get the largest connected sub-network that containing the author, using network traversal algorithm such as depth-first traversal algorithm (Awerbuch 1985) (Line 12). At last, we repeat the node (and edge) cutting process on the sub-network until the network remains unchanged to get a research team. While during

edge cutting, the algorithm traversal the sub-network and determine the node (author) satisfy Definition 4 Condition 1 and 2 or not (Line 14-20). And the sub-networks that at last survive the test are the output research teams[2].

Case Study

In this section, we verify the proposed method with teams identified based on the co-author network constructed from bibliographic data of Computer Sciences from Science Citation Index Expanded (SCI-EXPANDED) database.

Data

We use SU="computer science" and PY=2008-2017 as search strategies to search and download bibliographic data from SCI-EXPANDED database. The data relates to 344098 research papers. We use data later than year 2008 since after 2008 authors and their address (institutions) are carefully matched in the database. This property can greatly benefit the data refinement process.

Data refinement

Authors should be carefully recognized before teams can be identified. Two problems are commonly encountered in author recognition: homonymy (two authors with the same name) and synonymy (the existence of different variations on an author's name) (Antonio, Carlos & Félix, 2010). Same as Antonio et al., we adopt the idea given by the SCImago group to obviate those difficulties: avoiding homonymy by combining author and institution and synonymy by combining author and paper.

Results

After data collection and refinement, the teams are identified using the proposed algorithm. The experimental setting are: $\lambda = 3, \gamma = 0.3, \tau = 3, \xi = 0.3, M = 1, N = 3$. The algorithm finds 121 research teams that involve 688 authors. Each team contains 3-30 members, with average 5.69 members. The average team density is 0.88[3]. The statistics (on team size and density) signify that the identified research teams meet the consensus of a traditional research team that members cooperate tightly with each other in their academic career. It may not be intuitive that only 121 teams (with 688 authors) are found in more than 300,000 papers. We will refer to this problem latter in the next section. Now we give some survey results of teams whose information were verified through the Internet.

During the verification, many authors' homepages are hard to find. Only their personal pages in academic social networks like LinkedIn can be found. This kind of pages cannot provide information that is reliable and plentiful enough for verification. For some other authors, even their formal homepages can be found; their team information is not included in the page[4]. After investigation, five teams whose members' physical relations can be clearly recognized are exemplified here for demonstration (Many other teams can also be recognized however will not be demonstrated here due to the space limitation).

[2] Please note that during the edge cutting, the original connected sub-network may split into many disconnected sub- networks (teams).

[3] Network Density is defined as d=2L/N(N-1), where N is node number and L is edge number. $d \in (0,1)$, the closer d is to 1, the more connections among network nodes. Edges in this paper are weak or strong relationships.

[4] Please refer to http://www.utko.feec.vutbr.cz/~herencsar/ for this situation. The team members recognized by algorithm are Herencsar, Norbert, Vrba, Kamil, Jerabek, Jan, Sotner, Roman, Koton, Jaroslav. In the page, the team members' names actually appear -- in "List of Publications" -- many times. However, we still think it is the bibliographic kind of team with no clear team information.

Table 1. Investigation results of the identified teams

	Member	Institution	URL
TEAM 1	Xu, Jiang	Hong Kong University of Science and Technology	1) http://www.ece.ust.hk/~eexu/BDSL.html 2) http://www.ece.ust.hk/ece.php/profile/facultydetail/eeweiz
	Nikdast, Mahdi	Hong Kong University of Science and Technology	
	Zhang, Wei	Hong Kong University of Science and Technology	
	Wu, Xiaowen	Hong Kong University of Science and Technology	
	Ye, Yaoyao	Hong Kong University of Science and Technology	
	Wang, Xuan	Hong Kong University of Science and Technology	
	Wang, Zhehui	Hong Kong University of Science and Technology	
	Wang, Zhe	Hong Kong University of Science and Technology	
	Yang, Huazhong	Tsinghua University	
	Wang, Yu	Tsinghua University	
TEAM 2	Liao, Jianxin	Beijing University of Posts and Telecommunications	1) http://int.bupt.edu.cn/content/content.php?p=6_28_83 2) http://int.bupt.edu.cn/content/content.php?p=5_14_53 3) http://www.bupt.edu.cn/content/content.php?p=0_15_784
	Wang, Jingyu	Beijing University of Posts and Telecommunications	
	Qi, Qi	Beijing University of Posts and Telecommunications	
	Zhu, Xiaomin	Beijing University of Posts and Telecommunications	
	Li, Tonghong	Technical University of Madrid	
TEAM 3	Puliafito, Antonio	University of Messina	1) http://www.unime.it/it/ateneo/amministrazione/struttura/100350 2) http://www.unime.it/it/persona/salvatore-distefano 3) https://www4.ceda.polimi.it/manifesti/manifesti/controller/ricerche/RicercaPerDocentiPublic.do?evn_didattica=evento&k_doc=387991&aa=2013&lang=EN&jaf_currentWFID=main
	Bruneo, Dario	University of Messina	
	Longo, Francesco	University of Messina	
	Scarpa, Marco	University of Messina	
	Distefano, Salvatore	Polytechnic University of Milan	
TEAM 4	Zhu, Wenwu	Tsinghua University	1) http://www.tsinghua.edu.cn/publish/cs/4616/index_3.html 2) http://www.tsinghua.edu.cn/publish/cs/4853/2016/20161114091240133840490/20161114091240133840490_.html
	Wang, Zhi	Tsinghua University	
	Sun, Lifeng	Tsinghua University	
	Yang, Shiqiang	Tsinghua University	
	Cui, Peng	Tsinghua University	
TEAM 5	Chen, Tianshi	Chinese Academy of Sciences	1) http://novel.ict.ac.cn/tchen/ 2) http://novel.ict.ac.cn/ychen/index_cn.html 3) http://sse.ustc.edu.cn/pages/spage.php?type=0&pageid=1164)http://cs.ustc.edu.cn/szdw/fjs/201402/t20140216_190095.html
	Chen, Yunji	Chinese Academy of Sciences	
	Li, Ling	Chinese Academy of Sciences	
	Wang, Chao	University of Science and Technology of China	
	Li, Xi	University of Science and Technology of China	
	Zhou, Xuehai	University of Science and Technology of China	

Figure 1: Relations among the team members

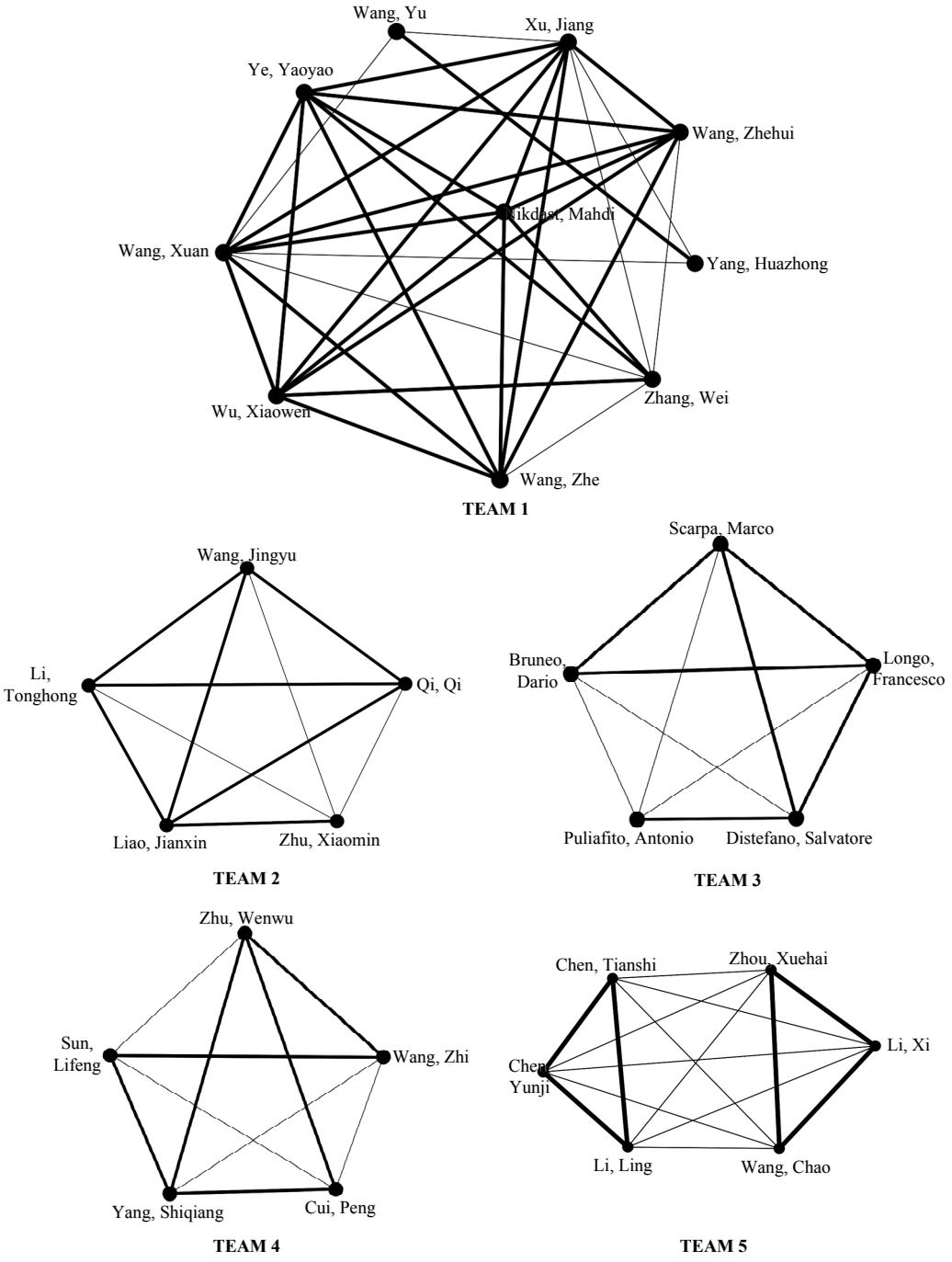

Table 1 gives the 5 teams with their member information, and the URLs that containing the information. Fig. 1 demonstrates the relations among the team members, with thick line representing strong relationship and thin line weak relationship. From the URLs we can see: **1)** for TEAM 1, almost all the members are listed in Xu, Jiang's homepage, "Group Members" section -- with the absence of Zhang, Wei and Yang, Huazhong. **2)** For TEAM 2-4, most team members are listed in the same faculty introduction page of the corresponding institution's official website. Only Li, Tonghong (TEAM 2), Distefano, Salvatore (TEAM 3) and Wang, Zhi (TEAM 4) are absence. **3)** For those who are in absence in TEAM 1-4, Zhang, Wei works

at Department of Electronic & Computer Engineering, HKUST -- same as Xu, Jiang (TEAM 1, URL 2); Li, Tonghong was graduated from Beijing University of Posts and Telecommunications and has tight cooperation with Liao, Jianxin (TEAM 2, URL 2 and 3); Distefano, Salvatore works (worked) at both University of Messina and Polytechnic University of Milan (TEAM 3, URL 2 and 3); Wang, Zhi was Prof. Yang, Shiqiang's PhD student (TEAM 4, URL 2)[5]. **4)** For TEAM 5, Chen, Tianshi; Chen, Yunji and Li, Ling are from the a team of Chinese Academy of Sciences, Wang, Chao; Li, Xi and Zhou, Xuehai are from the other team of University of Science and Technology of China. As we can see in Fig. 1, TEAM 5, if we consider on the team level, only weak relationships exist between the two (sub-)teams. That is, the situation can be easily avoided by adopting a post-verification process to the algorithm that removes all the weak links to see whether the team is still connected. If it is, then all the members belong to one team. If not, each connected component that fulfils Definition 4 is an independent team[6].

From the case study we can see that the algorithm can properly find research teams with members actually cooperated (are cooperating) with each other in their academic careers.

Discussion and Conclusion

It is easy to see that parameters used in the algorithm can influence the result teams in a very great manner. In general, the larger λ, γ, τ and ξ are, the less strong relationships are (and in turn the less teams can be found). Meanwhile, the larger M and N are, the more connections a team should have, the less teams can be found. That is exactly why only 121 teams are found in our case study: the bibliographic data only covers 10 years; it is a bit hard for authors to cooperate in 3 different years. Besides, as we can see, the average team density is as high as 0.88, which makes the qualified teams become even less. One can control the parameters to get the kind of teams they need. Actually, if we only set τ and leave other parameters as 0, the proposed method then degenerates to the traditional (mere) co-author based method: only numbers of co-authorships is considered. Under such setting, the team number will increase dramatically, and almost all the authors can be included into a certain team.

The other thing is that, the members in a research team may need to have the same (or similar) research interests. The problem does not exist in this paper since all the papers are come from Computer Sciences. However, the proposed scheme can be easily extended to involve research interests, by adding a relationship type that describing the similarity on research interests between authors, and requiring each pair of authors in a team should be connected by the relationship (with similarity larger than a given threshold). Besides adding conditions, the restriction defined in this paper can also be modified (strengthened or relaxed). For example, suppose we have adequate bibliographic data, we can further strengthen the time and output stable, by requiring the (least) consecutive cooperation years should be longer than γ times of their cooperative career, and at least having p papers published on each of the cooperation year.

[5] Chinese names: Liao, Jianxin (廖建新), Wang, Jingyu (王敬宇), Qi, Qi (戚 琦), Zhu, Xiaomin (朱晓民), Zhu, Wenwu (朱文武), Sun, Lifeng (孙立峰), Yang, Shiqiang (杨士强), Cui, Peng (崔鹏), Chen, Tianshi (陈天石), Chen, Yunji (陈云霁), Li Ling (李玲), Wang, Chao (王超), Li, Xi (李曦), Zhou, Xuehai(周学海)

[6] By doing so, the two members from Tsinghua University, Yang, Huazhong and Wang, Yu, should be removed from TEAM 1. The remove of Yang, Huazhong is straightforward, since he is not mentioned in Xu, Jiang's homepage. For Wang, Yu, the investigation shows that he is the Postdoc fellow of Prof. Xu, Jiang, and has relatively less output with Xu, .i.e., it is reasonable to exclude him from the team.

In conclusion, in this paper we propose a research team identification method that can efficiently go through the co-author network and output the research teams. The result teams are verified through Internet and proven to be real. The formalization makes the scheme easily extended to involve other restrictions like members having the same research interests.

References

Awerbuch, B. (1985). A new distributed depth-first-search algorithm. Information Processing Letters, 20(3): 147-150.

Börner, K., Dall'Asta, L., Ke, W., & Vespignani, A. (2010). Studying the emerging global brain: analyzing and visualizing the impact of co‐authorship teams. Complexity, 10(4): 57-67.

Calero, C., Buter, R., Valdés, C. C., & Noyons, E. (2006). How to identify research groups using publication analysis: an example in the field of nanotechnology. Scientometrics, 66(2): 365-376.

Li G., Li C. & Li X. (2014) The Identification of Research Teams Based on Social Network Analysis. Library and Information Service, 58(7): 63-70, 82.

Li, G., Liu, M., Wu, Q. & Mao, J. (2017) A Research of Characters and Identifications of Roles Among Research Groups Based on the Bow-Tie Model. Library and Information Service, 61(5): 87-94.

Perianes-Rodríguez, A., Olmeda-Gómez, C., & Moya-Anegón, F. (2010). Detecting, identifying and visualizing research groups in co-authorship networks. Scientometrics, 82(2): 307-319.

Qiu, J. & Wu, C. (2011) Study on the Co-Author Relationship of Informetrics Based on Social Network Analysis. Document, Information & Knowledge, 6: 12-17.

Ren, N. & Zhou, J. (2015) The Discovery and Evaluation of Research Team Under the Mode of Weighted Co-Author Network. New Technology of Library and Information Service, 9: 68-75.

Shen G., Huang S. & Wang D. (2013) On the Scientific Research Teams Identification Method Taking Co-authorship of Collaboration as the Source Data. New Technology of Library and Information Service, 1: 57-62.

Yan, E., & Ding, Y. (2010). Applying centrality measures to impact analysis: a coauthorship network analysis. Journal of the Association for Information Science & Technology, 60(10): 2107-2118.