Scientific collaborations within a university: From the viewpoint of complex networks

Weiquan Dong, Yizhen Yu and Xingang Wang
School of Physics and Information Technology, Shaanxi Normal University, Xi'an 710062, China
E-mail: wangxg@snnu.edu.cn

Abstract. While the topological features of scientific collaboration networks have been well explored in literature, the mechanisms by which the researchers are assembled remain not very clear. Here, by a 6-year publication data of the Shaanxi Normal University (SNNU), we construct the collaboration network and, combining with the additional information of the research staffs obtained through survey and interviews (including department affiliation, research direction, seniority, acquaintance, etc), investigate the key factors that promote scientific collaborations. It is found that, although of a relatively small size, the constructed network possesses still the general features of large-scale complex networks, including the community structures, the heterogeneous degree and weight distributions, the preferential attachment of the new links, etc. In exploring the assembly mechanisms of the collaboration patterns, it is further revealed that among the many factors influencing scientific collaborations, the closeness of the research directions has a clear advantage over the others. The findings provide insights into the formation of the collaboration patterns within universities, and might helpful to the network modeling and research management in practice.

1. Introduction

Bibliometric studies over the past two decades have shown a rapid increase in the number of co-authored papers in every scientific discipline, as well as within and across different countries and geographic areas. Indeed, with the increased complexity of the scientific problems (e.g., the brain research) and experimental facilities (e.g., the Large Hadron Collider), the demanding for large-scale, interdisciplinary scientific collaborations is more urgent than ever [1, 2, 3]. To understand the patterns of scientific collaboration and their impacts on knowledge production and discipline evolution, scientific collaboration itself has became an active research topic in the past decades [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18], and has attracted considerable attentions in a variety of disciplines such as information science, management science, psychology, sociology, philosophy, to name just a few. Recently, stimulated by the significant progresses achieved in network science [19, 20, 21], a new powerful method, namely the collaboration (co-authorship) complex network, has been proposed and employed in analyzing the scientific collaborations [4, 5]. In the general collaboration networks, authors are represented by nodes, and two authors are connected by a link if they have co-authored a paper. Thanks to the well-recorded electronic bibliographic datasets, large-scale collaboration networks have been reliably constructed and systematically investigated for a variety of disciplines in recent years, in which many novel properties of scientific collaborations have been disclosed, e.g., the small-world and scale-free
(rich-club) features [22], the community (cluster) structures [5], the associativity property [23], the robust-yet-fragile feature [24], the abnormal weight distribution [13], the formation and evolution of research teams [14], etc.

Although collaboration networks provide a powerful tool in unveiling the patterns of scientific collaborations, the efficiency of this method has been largely limited by the simplified node (author) and link (collaboration) properties. To be specific, many intrinsic properties of the network nodes (e.g., the gender and age of the authors) and links (e.g., the acquaintanceship between the authors) have been overlooked in literature [25]. This limitation is rooted in the format of the available bibliometric datasets, which record only the information of the author name and affiliation. As scientific collaborations among researchers are largely affected by their personal attributes [14, 15, 26], to have a deeper understanding on the assembly mechanism of collaboration patterns, it is therefore necessary to take into account the author attributes as much as possible. Motivated by this, in the present work we incorporate the method of collaboration network with the conventional methods (survey and interview) [27], and investigate the collaboration patterns within a university. Specifically, by a 6-year bibliometric dataset of SNNU, we construct the scientific collaboration network, and, based on a survey on the personal attributes for all the recorded researchers (including discipline, research interest/direction, working experience, date of enrollment, affiliation, acquaintanceship, etc), explore the key factors that influence scientific collaborations inside SNNU. An interesting finding is that, among the investigated factors, the research direction has a dominant role in influencing the collaborations than others. This finding reflects to a certain degree the current situation of scientific collaboration inside the China universities, which might be helpful for the research management and policy design in universities of the similar level.

2. Topological properties

The bibliometric dataset used in the present work is based on the journal publications recorded by the Institute for Scientific Information (ISI) during the period from January 2009 to December 2014. To focus on only the high-quality collaborations among research staffs working on natural science within SNNU, we filter the bibliometric dataset by the following requirements: 1) SNNU is appeared as the first address; 2) The paper is published in journals of natural science; 3) The journals are listed in the 1st and 2nd divisions of the Journal Citation Reports [28]; and 4) The paper is co-authored by at least two staffs from SNNU (i.e., excluding the solo publications and papers co-authored by the supervisor and his/her students). After such a filtering, there are totally 1645 papers left in the dataset, which are collaborated by \( N = 503 \) staffs distributed in \( M = 16 \) different departments in the university. (During the investigated period, the journal list is slightly modified year by year. In obtaining the dataset, we include a journal into the list once it appears in the 1st or 2nd divisions during the period.)

Having obtained the bibliometric dataset, we next construct the scientific collaboration network following the conventional method of network construction [4, 5]. In constructing the collaboration network, we represent each staff by a node, and two staffs are connected by a link if they have co-authored a paper. We have in total \( L = 935 \) links in the network. The weight of the link, \( w_{ij} \), is defined as the number of co-authored papers between staffs \( i \) and \( j \). The nodes are colored according to the departments, i.e., staffs affiliated with the same department are of the same color. The topological structure of the collaboration network is presented in Fig. 1. Unlike the general complex networks where nodes are reachable to each other (e.g., the Internet) [19, 20], here it is seen that the collaboration network is fragmental and scattering – a feature commonly observed in scientific collaboration networks [11, 25]. In particular, the network is constituted by a giant cluster and many small motifs. The giant cluster contains 401 nodes and 842 links, with staffs coming from 12 different departments. The disconnected motifs are of various sizes, ranging from 2 to 11.
Figure 1. (Color online) The structure and topological properties of the scientific collaboration network. (a) The network structure, which consists of $N = 503$ nodes and $L = 935$ links. The node size is plotted as proportional to the node degree, and the edge thickness is plotted as proportional to the link weight (the number of co-authored papers). Nodes are colored according to the departmental information of the staffs. The network is made up of a giant cluster (which contains 401 nodes and 842 links), 12 small motifs (of size between 3 and 11), and 17 node-pairs. (b) The degree distribution follows roughly the power-law scaling, $p(k) \propto k^\gamma$, with the fitted exponent $\gamma \approx -1.8$. (c) The weight distribution follows roughly the power-law scaling, $p(w) \propto w^\alpha$, with the fitted exponent $\alpha \approx -1.6$. (d) The variation of the average-nearest-neighbor-degree, $k_{nn}$, as a function of $k$, with the fitted slope $\beta \approx 1.4 \times 10^{-2}$.

We go on to explore the topological properties of the network. Previous studies have shown that realistic complex networks normally possess the small-world and scale-free features [22]. As the network is fragmental, in investigating the small-world feature we focus on only the giant cluster. The numerical results show that the giant cluster has the averaged diameter $d \approx 6.4$ and the clustering coefficient $c \approx 0.55$. Clearly, the giant cluster possesses the small-world feature. To investigate the scale-free feature of the network, we plot in Fig. 1(b) the distribution of the node degrees, $k$. It is seen that the degrees are of heterogeneous distribution. Particularly, the degree distribution follows roughly a power-law scaling within the range $k \in [1, 38]$, i.e., $p(k) \propto k^\gamma$, with the fitted exponent $\gamma \approx -1.8$. Besides the node degrees, the distribution of the link weights is also heterogeneous, as depicted in Fig. 1(c). Specifically, the weight distribution follows roughly the power-law scaling $p(w) \propto w^\alpha$ within the range $w \in [1, 90]$, with the fitted exponent $\alpha \approx -1.6$. For scientific collaboration networks, another commonly interested topological property is assortativity (assortative mixing), i.e., the propensity that nodes of the similar degrees are connected to each other [23]. To explore this property, we measure the average-nearest-neighbor-degree, $k_{nn}$, for node of degree $k$ in the network, and plot the variation of $k_{nn}$ as a function of $k$ [29]. If $k_{nn}$ is increasing (decreasing) with $k$, then the
network is classified as assortative (disassortative) [23]. The results are presented in Fig. 1(d). It is seen that, with the increase of $k$, the value of $k_{nn}$ is fluctuating wildly, showing no clear trend of increase or decrease. (By a linear fitting of the data, in fact we have $k_{nn} \propto \beta k$, with the slope $\beta \approx 1.4 \times 10^{-2}$.) Fig. 1(d) thus indicates that the collaboration network of SNNU has nearly the neutral assortativity, which is different from the general collaboration networks studied in literature (which are normally assortative) [23, 29]. The neutral assortativity might be attributed to the limited size (resulting in the large fluctuations) and sparse connections (resulting in the fragmental structure) of the SNNU collaboration network.

In investigating scientific collaboration networks, a particularly interested topological property is the community structure [30], i.e., the network is partitioned into some communities, with the connectivity inside each community is much denser than between the communities. As community reflects the meso-scale structure of complex networks, it has been employed as an important approach to exploring the patterns of scientific collaborations in previous studies [4, 11, 15]. Here we choose the giant connective cluster in the network, and partition it into communities according to the algorithm proposed by Girvan and Newman [30]. In particular, we remove recursively the network links by a descending order of their betweenness, and monitor the variation of the network modularity, $Q$. [Roughly speaking, modularity measures the probability for a pair of connected nodes randomly chosen to be staying within the same community, which is varying between 0 (no community structure) and 1 (distinct community structures). Please refer Ref. [30] for the detail definition of $Q$]. By finding the maximum modularity, $Q_{\text{max}}$, we are able to identify the optimal partition of the communities. Numerically, we find that $Q_{\text{max}} \approx 0.9$, under which the network is partitioned into $M' = 21$ communities, with the size of the communities ranging from 3 to 62.

Do the identified communities reflect the realistic situation? In partitioning the staffs within a university into communities [30], a conventional approach in social science is to group the staffs according to their departmental affiliations. That is, staffs from the same department form a community. Apparently, the two types of communities, i.e., the collaboration and department communities, are different from each other. For instance, the giant cluster contains 21 communities, while the staffs are from only 12 departments. However, as scientific collaborations between researchers are largely influenced by their disciplines and acquaintanceship, the two types of communities do have certain degree of correlations. To quantify the correlation between the two types of communities, we introduce the community-correlation coefficient (CDC)

$$ r = \langle r_l \rangle = \left\langle \frac{n'_l}{n_l} \right\rangle $$

with $n_l$ the size of the $l$th collaboration community identified through the Girvan-Newman algorithm, and $n'_l$ the maximum number of nodes affiliated to the same department within this community. $\langle \ldots \rangle$ represents that the result is averaged over all $M'$ communities. Clearly, the larger $r$ is, the stronger the correlation between the two types of communities is, and the stronger the community structures influenced by the departmental affiliations will be. The calculated result shows that $r \approx 0.8$. This strong correlation is somewhat expectable, as scientific collaborations within the same discipline in general have the higher propensity than cross disciplines, and, more importantly, staffs within the same department are usually acquaintances and friends [1, 2]. We would like to note that, while staffs within the same department are more likely to collaborate with each other (as compared with the collaborations between staffs from different departments), the collaborations are not of the equal probability, i.e., there exist some collaboration patterns within the department. For instance, staffs in the physics department are organized into 4 communities. Our above analysis of community structures thus suggests that departmental affiliation is influential, but not crucial to the scientific collaborations.
3. Assembly mechanisms

To explore the key factors promoting scientific collaborations, we next study the collaboration patterns at the nodal level, based on the additional information of the staffs obtained through the conventional approaches of survey and interview. Specifically, we are going to analyze in detail: 1) the collaboration propensity of the newcomers, i.e., with whom the newcomers tend to collaborate; 2) the collaboration patterns within three specific, small-size departments; and 3) the collaboration pattern at the departmental level within SNNU.

In modeling the growth of collaboration networks, it has been commonly assumed that the new nodes (newcomers) are connected to the existing nodes (veterans) by the mechanism of preferential attachment \[6, 9, 10, 16\]. That is, the larger the degree of an existing node is, the higher the possibility for it to be connected by the newcomer will be. While preferential link attachment has been observed in the empirical data of many natural and man-made complex networks \[19\], it remains unknown whether it applies to the growth of scientific collaboration network, as the relevant data is unavailable from the electronic bibliographic datasets. This suspicion is reasonable, as it usually takes a certain period for a newcomer to familiarize the new academic environment and find the new collaborators, resulting in probably the non-preferential link attachment. According to the surveyed information, we identify from the dataset the veterans who joined SNNU before 2009, and plot in Fig. 2 the average number of newcomers (staffs joined SNNU between 2009 and 2014) they have collaborated with, \(n_k\), versus their degrees, \(k\). It is seen that, in accordance with the assumptions made in previous modeling studies, the value of \(n_k\) is indeed increasing with \(k\) by roughly a linear fashion. Fig. 2 provides (probably the first) empirical evidence showing the applicability of preferential attachment mechanism in modeling scientific collaboration networks.

![Figure 2](image_url)

**Figure 2.** The preferential attachment mechanism of the newcomers in collaborating with the veterans. \(k\): the degree of the veterans joined SNNU before 2009. \(n_k\): the averaged number of newcomer-neighbors joined SNNU between 2009 and 2014. The fitted results show that \(n_k \propto \eta k\), with \(\eta \approx 0.15\).
We continue to analyze the collaboration patterns within a specific department. Due to the limited resources, we have only surveyed three research-oriented departments inside SNNU, namely the departments “A”, “B”, and “C”. Department “A” is a national engineering laboratory, which was founded in 2008 and now has 41 incumbent staffs. According to our dataset, department “A” has totally published 96 co-authored papers, which involve $N_A = 35$ staffs and give $L_A = 54$ collaboration links. The topological structure of the collaboration network for department “A” is presented in Fig. 3(a). It is seen that the nodes are organized into roughly a star-structure. Specifically, there is a supper (central) node of degree $k = 22$ in the network, which contributes 61% of the total publications; while the degrees of the remaining (peripheral) nodes are less than $k = 6$. From the surveyed result, it is also learned that the super “node” is a senior researcher of broad research interest, with research directions covering almost all the directions of the peripheral “nodes”. Besides the star-structure component, there also exists an isolated node-pair and two isolated nodes in the network. By checking the research background of these staffs, it is revealed that these staffs are working on research directions not covered by the super “node”.

Department “B” is a key laboratory of the Ministry of Education of China, which was founded in 2011 and now has 21 incumbent staffs. During the investigated period, $N_B = 17$ staffs have co-authored 39 papers in total, giving $L_B = 26$ collaboration links. The topological structure of the collaboration network for department “B” is presented in Fig. 3(b). It is seen that, different from department “A”, the collaboration network of department “B” consists of 2 disconnected clusters, together with 5 isolated nodes. Comparing to department “A”, it is seen that the node degrees of department “B” are more homogenously distributed, with $k$ ranging from 1 to 7. By interviewing with staffs within department “B”, it is learned that this is an interdisciplinary department founded by integrating three institutes from different departments. Although staffs in department “B” have the same research interest, they are actually working on different research directions and topics. Fig. 3(b) thus suggests that departmental affiliation is not the necessary condition for promoting scientific collaborations.

Department “C” is a provincial key laboratory, which was founded in 2005 and now has 36 incumbent staffs. During the investigated period, $N_C = 34$ staffs have co-authored 139 papers in total, giving $L_C = 54$ collaboration links. The topological structure of the collaboration network for department “C” is presented in Fig. 3(c). In comparison with the networks of departments A and B [Fig. 3(a) and (b)], the network of department C is featured by the existence of a few of tightly connected hub nodes. More specifically, the four largest-degree nodes [of node indices 14 ($k = 12$), 15 ($k = 9$), 17 ($k = 7$), and 22 ($k = 6$) in Fig. 3(c)] are globally connected and form the core of the community, to which most of the remaining nodes can reach by only one link. By interviewing the staffs, it is learned that this laboratory has long been established (back to 1980s) before promoting to the provincial key laboratory, and the research interest of the staffs are highly correlated. In particular, the four hub “nodes” are of the same research direction, but working on different topics.

Our above analysis on the formation of collaboration patterns points to the fact that, among the many influenceable factors, research direction has the dominating role in affecting scientific collaborations. More specifically, the closer the research directions between two staffs is, the higher the possibility for them to collaborate with each other will be. This is evidenced in the collaboration patterns of department “A” [Fig. 3(a), where the research directions of the peripheral “nodes” are different from each other, but are all overlapped with the super “node”], department “B” [Fig. 3(b), where the research directions for nodes within the same cluster is highly centralized, but are different among the clusters], and department “C” [Fig. 3(c), where all “nodes” are working on different topics of the same research direction].

To explore further the crucial role of research direction on collaboration, we finally explore the collaboration patterns at the department level, i.e., combining nodes of the same color into a new
Figure 3. (Color online) (a) The collaboration network of department “A”, which consists of \( N_A = 35 \) nodes and \( L_A = 54 \) links. The majority of the nodes are contained in the star-structure. (b) The collaboration network of department “B”, which contains \( N_B = 17 \) nodes and \( L_B = 26 \) links. The network consists of 2 disconnected clusters. (c) The collaboration network of department “C”, which contains \( N_C = 34 \) nodes and \( L_B = 54 \) links. The four largest-degree nodes are globally connected. (d) The collaboration network at the department level, which consists of 16 nodes and 34 links.

Figure 3. (Color online) (a) The collaboration network of department “A”, which consists of \( N_A = 35 \) nodes and \( L_A = 54 \) links. The majority of the nodes are contained in the star-structure. (b) The collaboration network of department “B”, which contains \( N_B = 17 \) nodes and \( L_B = 26 \) links. The network consists of 2 disconnected clusters. (c) The collaboration network of department “C”, which contains \( N_C = 34 \) nodes and \( L_B = 54 \) links. The four largest-degree nodes are globally connected. (d) The collaboration network at the department level, which consists of 16 nodes and 34 links.

node and integrating the links accordingly. In such a way, the original network is compressed into a small-size network of only \( M = 16 \) nodes and \( L = 16 \) weighted links. (For a better presentation, only links of weight larger than 5 are kept). The network structure is presented in Fig. 3(d). Several features are observed in this departmental collaboration network, which manifests from another aspect the importance of research direction on scientific collaborations. Firstly, despite the large number of recorded publications, the network is still not globally connected, implying the difficulty of interdisciplinary collaborations (resembling the feature of department “B”). Secondly, the hub (non-hub) nodes stand for departments of fundamental (modern) science, e.g., the largest-degree \( (k_1 = 7) \) and second largest-degree \( (k_2 = 6) \) nodes symbolize, respectively, the chemistry (node 1) and biology (node 2) departments. This feature is similar to that observed in department “A”, where the super “node” is of broad research interest and the peripheral nodes are of focused research interest. Thirdly, the connection between nodes
1 (the chemistry department) and 5 (the department of material science) is extremely strong (totally 95 co-authored papers), despite the fact that they belong to different disciplines. By checking the history of these two departments, it is found that node 5 was separated from node 1 in 2011, and many staffs still continue their collaborations since then. This finding, together with the finding of CDC in Sec. II, confirms the fact that scientific collaborations are influenced but not governed by the department affiliations.

4. Discussion and conclusion
The present work is featured by considering additional information about the network nodes and links. While collaboration networks have been extensively investigated in literature, the existing studies are mostly focusing on the network topological properties (e.g., degree distribution, small-world effect, and community structures) and less attention has been paid to the detailed information of the network nodes and links (e.g., affiliation, seniority, research direction, and acquaintanceship). In the present work, we pick up these missed information by the conventional methods used in social science (i.e., survey and interview), and revisit the assembling mechanism of collaboration patterns. With the additional information, we are able to analyze individually the factors that affect collaboration and, more importantly, identify the key elements, providing therefore a deeper understanding on the mechanism of scientific collaboration. We would like to note that all the findings are based on the specific publication dataset of SNNU, and it remains unknown to us whether the same results stand for other universities and organizations. In our further works, we shall keep on updating the SNNU publication dataset (e.g., extending the publication period), and, in the meantime, try to justify our findings by more datasets from other universities national- and world-wide. It is our hope that the findings, and also the new approach of collaboration analysis proposed in the present work, would draw attention from researchers in the relevant fields, and, by “collaborating” with each other, we together could unveil the origin of scientific collaboration inside universities. (All data used in the present work can be obtained from the authors upon request for research purposes.)

Summarizing up, based on a 6-year publication dataset of SNNU, we have constructed the scientific collaboration network of the university, and, combining with the staff additional information obtained through survey and interviews, explored in detail the assembly mechanism of various collaboration patterns. It is found that, among the many factors influencing scientific collaboration, the closeness of the research directions has a clear advantage over the others. The finding sheds new light on the mechanism of scientific collaborations, and might helpful to the policymaking and management of research activities in universities.

Acknowledgments
This work was supported by the National Natural Science Foundation of China under Grant No. 11375109 and by the Fundamental Research Funds for the Central Universities under Grant No. GK201601001.

References
[1] Guimerà R, Uzzi, B, Spiro J and Nunes Amaral L A 2005 Science 308 697.
[2] Wuchty S, Jones B F and Uzzi B (2007) Science 316 1036.
[3] Sonnenwald D H (2007) Annual Review of Information Science and Technology 41 643.
[4] Newman M E J (2001) Proc. Natl Acad. Sci. USA 98 404.
[5] Newman M E J (2001) Phys. Rev. E 64 016132.
[6] Li M, Wu J, Wang D, Zhou T, Di Z and Fan Y (2007) Physica A 375 355.
[7] Hsu J W and Huang D W (2009) Phys. Rev. E 80 057101.
[8] Lee D, Goh K I, Kahng B and Kim D (2010) Phys. Rev. E 82 026112.
[9] Zou S R, Xu Y Y, Xu X L, Wang J and He D R (2011) Chinese Physics Letters 28 018902.
[10] Durugbo C, Hutabarat W, Tiwari A and Alcock J R (2011) Information Sciences 181 3143.
[11] Martin T, Ball B, Karrer B and Newman M E J (2013) Phys. Rev. E 88 012814.
[12] Uddin S, Hossain L and Rasmussen K (2013) PLoS ONE 8 e57546.
[13] Ke Q and Ahn Y Y (2014) Phys. Rev. E 90 032804.
[14] Milojevic S (2014) Proc. Natl Acad. Sci. USA 111 3984.
[15] Sun X, Lin H, Xu K and Ding K (2015) Scientometrics 104 43.
[16] Liu J, Li Y, Ruan Z, Fu G, Chen X, Sadiq R and Deng Y (2015) Physica A 419 29.
[17] Zhou J, Zeng A, Fan Y and Di Z (2015) Scientometrics 106 805.
[18] Shen Z, Yang L, Pei J, Li M, Wu C, Bao J, Wei T, Di Z, Rousseau R and Wu J (2016) Journal of Informetrics 10 82.
[19] Albert R and Barabasi A L (2002) Rev. Mod. Phys. 74 47.
[20] Newman M E J (2003) Society for Industrial and Applied Mathematics Review 45 167.
[21] Fang J Q, Bi Q and Li Y (2007) Front. Phys. 2 109.
[22] Barrat A, Barthelemy M, Pastor-Satorras R and Vespignani A (2004) Proc. Natl Acad. Sci. USA 101 3747.
[23] Newman M E J (2002) Phys. Rev. Lett. 89 208701.
[24] Liu X F, Xu X K, Small M and Tse C K (2011) PLoS ONE 6 e26271.
[25] Cainelli G, Maggioni M A, Uberti T E and Felice A D (2015) Scientometrics 102 673.
[26] Pepe A (2011) Journal of the American Society for Information Science and Technology, 62 2121.
[27] Wasserman S and Faust K (1994) Social network analysis: Methods and applications (Cambridge University Press, Cambridge, England)
[28] http://www.fenqubiao.com/
[29] Catanzaro M, Caldarelli G and Pietronero L (2004) Phys. Rev. E 70 037101.
[30] Girvan M and Newman M E J (2002) Proc. Natl Acad. Sci. USA 99 7821.