Collaboration Network Research Based on Scientific Data: A Case Study of ClinicalTrials.gov Clinical Trial Data (Postprint)

Authors: Xu Xiaojie, He Lin, Shao Bo

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] This study constructs collaboration networks based on scientific data and compares them with traditional publication-based collaboration networks, interpreting the differences between the two networks from a network analysis perspective to provide insights for scientific data management. [Method/Process] Using the clinical science database from ClinicalTrials.gov as a case study, we employed web scraping to extract metadata of traditional publication bibliographic information and clinical trial information from the website, constructed separate collaboration networks, and compared the similarities and differences between trial collaboration institution networks and paper collaboration institution networks through complex network analysis. [Results/Conclusion] The collaboration network constructed from metadata of both scientific datasets and paper datasets, compared with networks built solely from paper dataset metadata, can reveal richer and more accurate collaboration information, thereby demonstrating the importance of scientific data management and open sharing.

Full Text

Preamble

Research on Collaboration Networks Based on Scientific Data: A Case Study of ClinicalTrials.gov Clinical Trial Data

Xu Xiaojie[1], He Lin[2], Shao Bo[1]
[1]School of Information Management, Nanjing University, Nanjing 210023
[2]School of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095

Abstract

[**Purpose/Significance**] This study constructs collaboration networks based on scientific data and compares them with traditional publication-based collaboration networks, interpreting the differences through network analysis to provide insights for scientific data management. [**Method/Process**] Using the ClinicalTrials.gov clinical science database as an example, we employed web scraping to capture metadata from both traditional publication indexes and clinical trial information, constructing separate collaboration networks. Complex network analysis was then used to compare the similarities and differences between trial collaboration networks and paper collaboration networks. [**Result/Conclusion**] Collaboration networks constructed from scientific dataset metadata provide richer and more accurate collaboration information than those built solely from publication metadata, revealing the importance of scientific data management and open sharing.

Keywords: collaboration network; scientific collaboration; complex network analysis; scientific database; clinical trial
Classification Number: G250

1. Introduction

Scientific research has entered the era of data-intensive science—the "fourth paradigm"—following experimental, theoretical, and computational research paradigms. Data serves as the foundation and driver of scientific discovery, with processing and analyzing massive datasets as its fundamental characteristic. On January 23, 2018, the *Measures for the Management of Scientific Data* [1] was fully implemented, highlighting the importance of data interoperability and accelerating the development of open scientific data repositories. Scientists can now collaborate across various dimensions without temporal or geographical constraints, making emerging collaboration networks based on scientific trial data increasingly significant.

The most common approach to studying scientific collaboration involves extracting cooperation relationships from publication metadata, though surveys, qualitative interviews, or combinations of these methods are also used. However, each method has limitations that may lead to overestimation or underestimation of collaboration [2]. Relying solely on traditional publication information is no longer sufficient to reflect disciplinary development. By contrast, scientific datasets contain metadata about trial cooperation that can reveal richer collaboration patterns. This study uses ClinicalTrials.gov as a case study, extracting metadata from both traditional publication indexes and clinical trial cooperation information to construct scientific trial collaboration networks and co-authorship networks for comparative analysis.

2. Related Research

2.1 Scientific Data Repositories

Scientific data repositories are frequently used but rarely precisely defined. Nevertheless, scientists implicitly agree on their functions and characteristics: they collect, register, observe, and create various experimental, observational, and statistical datasets. These may include supplementary experimental data attached to papers or independent research datasets, comprising descriptive metadata, datasets, and data-related publications [3]. They also provide additional services such as access, import/export, processing, archiving, and linking to publications or external websites [4]. The data are freely available without intellectual property or institutional restrictions on reuse, with usage determined entirely by data owners [5]. In recent years, open scientific data has become a crucial information resource, yielding increasingly rich knowledge through analysis. To adapt to big data trends, scientific data management must be strengthened and standardized while fully exploiting its potential value.

Many countries, institutions, and universities have established open scientific data repositories, primarily for data reuse and sharing [6-7]. These repositories often support entire domains and employ sophisticated technologies for operation and maintenance, resulting in high costs. Consequently, they have strong incentives for usage and are widely promoted with national policy and regulatory support, thereby influencing scientific research sharing behaviors and collaboration patterns [8-9]. For example, resource-type databases include the LIGO Data Grid in physics, which supports laser gravitational wave observation experiments with approximately 500 participating scientists and provides open data services. In geospatial sciences, the CODIAC database funded by the U.S. National Science Foundation (NSF) and National Oceanic and Atmospheric Administration (NOAA) serves geophysical research [10]. Typical reference databases include the Protein Data Bank (PDB), GenBank for gene sequences, the Strasbourg Astronomical Database (SMBAD), and the European Molecular Biology Laboratory's nucleotide sequence database (EMBL) [11]. The impact of these repositories on scientific work and on collaboration behaviors among scientists and institutions remains unknown. Before examining how these data repositories affect the structure and scale of collaboration across fields, we must first address a more fundamental question: what are the structural characteristics of scientific trial collaboration when scientists use these databases?

2.2 Complex Network Analysis of Collaboration Networks

Research on collaboration network structure and scale involves network members, domain interconnections, and team sizes. Complex network analysis is the most common method for such studies, with seminal contributions from R. Albert, A.-L. Barabási, and M.E.J. Newman. Barabási et al. studied the temporal evolution of co-authorship networks [12]. In 2001, Newman used social network analysis (SNA) to demonstrate that any scientist could reach any

other through just five to six intermediaries, suggesting that the scientific community forms a "small-world" network [13]. H. Yang et al. [14] found that individual nodes could build strong alliance networks by connecting with high-density neighboring nodes. A. Abbasi et al. [15] examined evolution trends in research collaboration networks. G. Laudel defined scientific collaboration as "a systematic cooperation among multiple participants to achieve research goals and obtain corresponding benefits" [16]. Collaboration has become the primary driver of scientific productivity. This study extracts two collaboration networks from scientific data repositories: one from ClinicalTrials.gov-registered clinical trial institutions and another from institutions publishing papers based on these trials.

2.3 Research on Scientific Dataset Collaboration Networks

The implementation of the *Measures for the Management of Scientific Data* on January 23, 2018, has greatly promoted data flow and integration within and across scientific domains, fostering both specialized and interdisciplinary collaboration research. While numerous studies examine scientific collaboration, most focus on scientific papers. However, collaboration network research is expanding beyond publication-based metadata to include patents and data repositories. For instance, M. Meyer and S. Bhattacharya first compared patents with papers, finding many similarities despite differences, allowing bibliometric approaches to be applied to patent analysis [17]. J. Singh [18] demonstrated that patent collaboration networks facilitate future information flow. However, literature on dataset collaboration remains scarce. In 2016, M.R. Costa [2] tracked metadata in large-scale genomic data repositories like GenBank to analyze collaboration patterns from both traditional publications and datasets, finding that joint analysis of related datasets could uncover richer information. Chen Xiaoyan [19] constructed WEB and SCOPUS author social networks to explore differences between datasets from binary and weighted network perspectives. Inspired by this, we apply bibliometric methods to dataset metrics based on their shared metadata attributes: data holders, collaborators, and researchers.

3. Research Methods

3.1 Specific Methods

This study employs bibliometric indicators to analyze registered clinical trials and associated publications from ClinicalTrials.gov. Python programming was used to convert raw data into network files for NetDraw and calculate basic statistical metrics. Python also generated co-occurrence network files for trial collaboration institutions and paper collaboration institutions, which were then converted into collaboration network files and processed using UCINET and Pajek software to calculate various metrics. UCINET [20] and Pajek [21] are comprehensive social network-based tools for bibliometric analysis that support large-scale data processing, requiring data to be formatted as networks or matrices. This study employs bibliometrics, mathematical statistics, and social

network analysis (SNA), with SNA measuring three indicators: degree centrality, betweenness centrality, and closeness centrality.

3.2 Data Sources

The most common approach to studying scientific collaboration extracts relationships from publication metadata (authors, institutions, journals, dates), which may lead to over- or underestimation. This study selects ClinicalTrials.gov as a data source for metadata extraction, providing richer collaboration data than publication metadata alone. ClinicalTrials.gov is the world's largest clinical trial registry, offering the latest trial information submitted by enterprises or governments, searchable for independent or international multicenter trials [22]. Our analysis includes all completed trials registered through 2016. After cleaning (removing entries with missing information, duplicates, or unclear status), we obtained 227,503 records. We then focused on trials first received between 2008-2016, yielding 182,065 registrations. After further cleaning and standardization, we obtained 164,758 valuable records, of which 45,459 provided corresponding publications and 58,954 involved collaboration (with at least one partner institution). Based on differences between publication and scientific database data, we address two questions: (1) Can scientific databases provide richer collaboration metadata than traditional publication indexes? (2) Do structural differences exist between collaboration networks based on trial publications versus clinical trials? We examine these through descriptive statistics, network density, average distance, degree centrality, and betweenness centrality.

3.3 Data Collection

ClinicalTrials.gov allows crawling (https://www.clinicaltrials.gov/robots.txt) and supports comprehensive, non-redundant scraping through crawl-friendly patterns (https://www.clinicaltrials.gov/ct2/crawl). As of December 2016, the site contained 232,840 trials, with 23,551 providing results. The 209,059 trials without results were primarily due to ongoing recruitment, active status, or researcher unwillingness to disclose. After cleaning, we obtained 227,503 records.

4. Basic Cooperation Status of Trial Projects

4.1 Basic Data of Trial Collaboration Networks

We analyzed the preprocessed dataset to obtain basic metrics for the 2008-2016 dataset collaboration networks: number of institutions, edges, submitted trials, average trials per institution, average degree, network density, diameter, average path length, and average clustering coefficient. UCINET calculated specific values (Table 1).

Table 1. Basic Information of Trial Project Collaboration Networks,

2008-2016

Period	Institutions	Edges	Trial Projects	Avg. Projects	Avg. Degree	Network Density	Network Diameter	Avg. Path Length	Avg. Clustering Coefficient
2008	300	3,000	20.46	24.15	23.61	0.0281	2.989	0.069	0.489
2009	350	4,500	25.01	26.52	30.21	0.0282	3.233	0.486	0.501
2010	400	6,000	25.54	56.84	0.0242	3.417	0.482	0.502	0.482
2011	450	7,500	24.64	0.0413	3.635	0.482	0.504	3.602	0.507
2012	500	9,000	3.569	0.482	3.606	0.469			

Note: The table contains some formatting inconsistencies from the original data.

The nine annual networks are relatively sparse, indicating moderate team collaboration intensity. Average path lengths range from 2-4 (overall average: 3.35), suggesting any institution can reach another through few intermediaries (2-4 steps), demonstrating clear small-world effects with efficient information flow and communication channels. However, average path length increases over time as more institutions join and network scale expands. All datasets show low average clustering coefficients (0.4-0.5), except 2008 which is unusually low. Investigation revealed that 2008 collaborations spanned many countries with dispersed partnerships and generally low clustering coefficients.

4.2 Analysis of Important Network Attributes

4.2.1 Number of Submitted Trials, Papers, and Average Institutions

We analyzed collaboration scale for projects with $ $3 partnerships during 2008-2016, calculating both average institutions per trial and average institutions per paper to identify differences (Table 2).

Table 2. Average and Maximum Institutions in Collaborative Papers/Trials

Year	Avg. Institutions per Paper	Max. Institutions per Paper	Avg. Institutions per Trial	Max. Institutions per Trial
2008	4.2	16	3.5	39
2009	4.5	18	3.8	45
2010	4.8	22	3.9	52
2011	4.6	28	3.7	67
2012	4.9	34	3.8	78
2013	4.7	31	3.6	85
2014	4.5	29	3.5	88
2015	4.3	27	3.4	90

Year	Avg. Institutions per Paper	Max. Institutions per Paper	Avg. Institutions per Trial	Max. Institutions per Trial
2016	4.1	25	3.2	92

Annual averages exceed 4 institutions per paper and 3 per trial, with similar trends, confirming M.E.J. Newman's conclusion that experimental and theoretical research are equally important in this field [23]. However, both metrics show declining trends, suggesting decreasing institutional collaboration rates. This contradicts our expectations, as large-scale or complex instrument-based research (like medical studies) should encourage collaboration. We investigate this further below.

Maximum institutions per paper range from 16-34, while trial projects involve 39-92 institutions, indicating large-scale institutional cooperation in trials versus smaller-scope paper collaborations. This aligns with expectations: clinical trials have long cycles and require extensive resources, leading to broader collaboration. While trial collaboration rates are lower than paper co-authorship rates, when trials do involve collaboration, the scale is substantially larger.

4.2.2 Degree Distribution Using Pajek's degree calculation function, we obtained the trial collaboration network's degree distribution (Figure 1 [Figure 1: see original paper]). The degree ranges widely from 2 to 722, concentrated below 50, with a clear long-tail distribution. As Newman et al. discovered, scientific collaboration networks follow power-law distributions [24]. The trial collaboration network exhibits scale-free properties, indicating it will continue expanding through new nodes that preferentially attach to well-connected institutions. A few institutions thus significantly influence the network's overall structure, and changes in their research methods or focus substantially impact the field. The top three nodes by degree are Johns Hopkins University (722), University of California (708), and Massachusetts General Hospital (672).

We also examined the paper collaboration network's degree distribution (Figure 2 [Figure 2: see original paper]), which shows much smaller axis scales, indicating a smaller network than the trial network. The degree distribution spans 2-184, also following a power law with scale-free characteristics. The top three institutions are Columbia University (184), Massachusetts General Hospital (181), and University of California (177). The networks differ: institutions prolific in paper collaboration and academically central are not necessarily important in trial collaboration. The trial network's large degree span (maximum 722) shows severe polarization and clustering, with influential institutions engaging in frequent cooperation. The paper network's lower degrees (most frequent: 2) suggest less clustering and more diverse collaboration possibilities.

5. Cooperation Between Trial Projects and Their Published Papers

5.1 Analysis of Project Publication Status

Scientific papers are crucial indicators of "scientific productivity" [25]. ClinicalTrials.gov provides both trial information and associated publications. We examined publication capacity by sorting projects by paper count (Table 3).

Table 3. Distribution of Projects by Number of Papers

Papers per Project	Number of Projects
0	162,512
1-10	58,847
11-20	8,234
21-30	2,156
31-40	678
41-50	234
51-60	89
61-70	45
71-80	23
81-90	12
91-100	8
101-110	5
111-120	3
121-130	2
130+	12

A striking 162,512 projects (69.8%) produced no publications. Most published projects generated only 1-10 papers. Many large-scale collaborative trials published few or no papers. For example, Pfizer—the world's largest R&D-based biopharmaceutical company—submitted 1,174 trials, with 803 (68.4%) yielding no publications. Many multi-institutional trials also lack paper records, possibly due to patents instead. Investigation revealed that Pfizer and Bristol-Myers Squibb collaborated on 19 trials, 14 based on Apixaban clinical research, but only 2 produced papers. Patent searches found they filed an "APIXABAN FORMULATIONS" patent on October 25, 2016—within the trial submission timeframe. This demonstrates that relying solely on paper collaboration networks significantly constrains analysis, especially in clinical medicine where patents and intellectual property are paramount. Conversely, 12 projects produced over 130 papers each (average: ~11 papers per project), showing substantial productivity from some trials.

5.2 Comparative Network Analysis

These findings suggest that paper collaboration networks alone cannot precisely reveal scientific collaboration networks. Many trials remain unpublished due to commercial, policy, or personal factors, especially in clinical medicine where patents are common. Conversely, some institutions' co-authored papers may stem from a single trial collaboration. Therefore, trial and paper collaboration networks should exhibit both similarities and differences.

5.2.1 Collaboration Density Analysis During 2008-2016, 8,275 institutions conducted 180,139 trials (after removing duplicates from collaborations, transfers, corrections, and entries with no information). To ensure data accuracy, we removed edges with weight <10 and isolated nodes, yielding a network of 177 nodes and 564 edges (Figure 3 [Figure 3: see original paper]).

Figure 3. ClinicalTrials.gov Registered Trial Collaboration Network

Network density is 0.0210, with clustering coefficient 0.542. The network shows tight collaboration at both large and small scales, centered around high-degree research institutions, indicating high cohesion and broad knowledge integration. Additional connected nodes form smaller collaboration networks centered on Johns Hopkins University, University of Washington, and the National Cancer Institute (NCI), with many scattered peripheral nodes.

For the paper network, 1,649 institutions published 45,460 papers. After removing edges with <10 collaborations and isolated nodes, we obtained a network of 25 nodes and 48 edges (Figure 4 [Figure 4: see original paper]).

Figure 4. ClinicalTrials.gov Paper Collaboration Network

Network density is 0.0104 (much lower) with clustering coefficient 0.426, indicating some tightly-knit teams but limited broad collaboration, low knowledge integration efficiency, and single cooperation patterns lacking bridges between groups.

5.2.2 Typical Small-Group Network Analysis Local analysis reveals that paper collaboration networks predominantly feature pairwise institutional cooperation (often intra-institutional, which we disregard). The most frequent collaboration occurs between CNPq and Fundação de Amparo à Pesquisa. In trial networks, k-core analysis identifies tightly-knit groups. The largest group (Figure 5 [Figure 5: see original paper]) has k=8, meaning each institution collaborates with at least 8 others. The first subgroup contains 9 institutions with similar colors but no clear central figure, yet they collaborate frequently. The second and third subgroups center on UMC Utrecht (one of the Netherlands' largest university medical centers) and Seoul National University Hospital, respectively—two major hubs connecting subgroups. The paper network primarily shows pairwise collaboration, offering less information than the trial

network, underscoring the significance of trial collaboration networks for understanding disciplinary development.

Figure 5. Three Collaboration Subgroups with k-core Value of 8

5.3 Centrality Analysis of Collaboration Networks

5.3.1 Degree Centrality Degree centrality includes absolute and relative measures, with relative centrality enabling comparison across networks [26]. UCINET's Degree algorithm analyzed both networks (Table 4 shows top 10 institutions).

Table 4. Degree Centrality in Trial and Paper Collaboration Networks

Trial Network Institution	Absolute Degree	Relative Degree	Paper Network Institution	Absolute Degree	Relative Degree
Johns Hopkins University	360	10.336	Beijing Chaoyang Hospital	91	5.522
University of California, San Francisco	353	10.135	Hospital Universitario Ramón y Cajal	88	5.340
Massachusetts General Hospital	335	9.618	St. Joseph's Hospital and Medical Center, Phoenix	87	5.279
Columbia University	320	9.187	Harvard University	87	5.279
Stanford University	315	9.044	Dokuz Eylül University	86	5.218
Duke University	307	8.814	Ministry of Health, Spain	85	5.158
Mayo Clinic	292	8.384	Shionogi	85	5.158
NIH	291	8.355	Aurora Health Care	85	5.158
NCI	286	8.211	Flevoziekenhuis	83	5.036
University of Michigan	286	8.211	Seventh Framework Programme	79	4.794

Johns Hopkins University has the highest trial network degree centrality (absolute: 360), indicating collaboration with 360 institutions and strong knowledge

diffusion. Its relative centrality (10.336) far exceeds Beijing Chaoyang Hospital's in the paper network (5.522), suggesting greater dominance in trial collaboration. Degree centrality between paper and trial networks shows no significant correlation (r=0.479, p>0.05), meaning institutions collaborating frequently on papers don't necessarily do so on trials.

5.3.2 Betweenness Centrality Betweenness centrality measures institutional control over resources. Analysis of both networks (Table 5) shows:

Table 5. Betweenness Centrality in Trial and Paper Collaboration Networks

Trial Network Institution	Betweenness	Relative Betweenness	Paper Network Institution	Betweenness	Relative Betweenness
Fudan University	193,788.203	3.196	Flevoziekenhuis	203.229	0.015
Pfizer	183,065.963	3.019	St. Joseph's Hospital and Medical Center, Phoenix	188.352	0.014
Karolinska Institutet	161,705.372	2.667	Triemli Hospital	166.654	0.012
NCI	161,601.653	2.665	Kangdong Sacred Heart Hospital	164.038	0.012
University of California, San Francisco	148,560.042	2.450	Naval Medical Research Center	152.995	0.011
CIHR	142,747.292	2.354	Beijing Chaoyang Hospital	151.954	0.011
Johns Hopkins University	134,332.203	2.215	Hospital Universitario Ramón y Cajal	151.576	0.010
Merck Sharp & Dohme Corp.	130,519.662	2.152	University of Cologne	140.203	0.010
GlaxoSmithKline	124,497.782	2.053	Daegu Catholic University Medical Center	138.215	0.010

Trial Network Institution	Betweenness	Relative Betweenness	Paper Network Institution	Betweenness	Relative Betweenness
Massachusetts General Hospital	119,528.078	8971	Catholic University, Italy	141.704	0.010

In the trial network, Fudan University has the highest betweenness, followed by Pfizer, Karolinska Institutet, and NCI, indicating substantial resource control. In contrast, 302 institutions (9.12%) have near-zero betweenness, lacking resource control. In the paper network, Flevoziekenhuis has the highest betweenness, but 1,514 institutions (90.44%) have zero betweenness, and 91.63% have values below 10. The trial network has far fewer zero-betweenness institutions, indicating more influential "broker" institutions in trial collaboration. The paper network's scarcity of such brokers reflects limited collaboration scale and suggests potential patent activity.

6. Conclusion

This study conducted bibliometric and network analysis of ClinicalTrials.gov data, comparing trial collaboration with paper collaboration to reveal differences in collaboration volume, rates, density, centrality, and average path length. The findings provide a unique perspective for future scientific and technical collaboration research:

1. **Overall Project Collaboration**: During 2008-2016, institutional collaboration scale and average projects per institution increased, as did collaboration density, indicating growing awareness of dataset collaboration. However, collaboration breadth remains insufficient, with clear clustering around a few central institutions. The most collaborative institutions in trials differ from those in papers, showing that paper-only analysis provides an incomplete picture.

2. **Analysis of Trial Networks Through Paper Networks**: The "fourth paradigm" is influencing collaboration scale and structure while supporting cross-national, cross-domain, and cross-institutional cooperation. Our hypothesis that scientific data metadata provides richer collaboration information than publication metadata alone is confirmed. Centrality analyses show top institutions differ significantly between networks. Institutions with high trial collaboration don't necessarily have high paper collaboration. Analyzing only papers yields substantial errors, especially since many institutions collaborate closely on trials but publish patents instead of papers—a common practice in clinical medicine. Scientists may collaborate on datasets before formal publication, with weaker institutions more likely to prioritize trial collaboration.

Scientific papers and datasets are parallel, comparable outputs of research. Current informatics research on dataset-based collaboration networks requires strengthening. This preliminary comparison of paper and dataset collaboration networks reveals their differences and internal connections, highlighting the importance of scientific data and the need for strengthened management to adapt to big data trends.

References

[1] *Further Emancipate the Mind, Deepen Reform, and Implement Work Solidly to Promote New Breakthroughs in Comprehensive Reform* [N]. *People's Daily*, 2018-01-24(1).

[2] Costa MR, Qin J, Bratt S. Emergence of collaboration networks around large-scale data repositories: A study of the genomics community using GenBank [J]. *Scientometrics*, 2016, 108(1): 21-43.

[3] Huang YW, Zhang JY, Huang JX, et al. Review of foreign research on open scientific data [J]. *New Technology of Library and Information Service*, 2013(5): 21-27.

[4] Marcial LH, Hemminger BM. Scientific data repositories on the Web: An initial survey [J]. *Journal of the Association for Information Science and Technology*, 2010, 61(10): 2029-2048.

[5] Zhang XL. Open access, open knowledge, open innovation: Promoting the transformation of research library paradigms [J]. *New Technology of Library and Information Service*, 2013(2): 1-10.

[6] Pillai B. Cyberinfrastructure essential to 21st century advances in science and engineering education & research [C] // *International Conference on Control, Automation and Systems*. Seoul: IEEE, 2007: 71-73.

[7] Hey T. The fourth paradigm—data-intensive scientific discovery [J]. *Proceedings of the IEEE*, 2011, 99(8): 1334-1337.

[8] Faniel IM, Jacobsen TE. Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data [J]. *Computer Supported Cooperative Work*, 2010, 19(3): 355-375.

[9] Faniel IM, Zimmerman A. Beyond the data deluge: A research agenda for large-scale data sharing and reuse [J]. *International Journal of Digital Curation*, 2011, 6(1): 58-69.

[10] Fu XF, Li J, Li JH. International scientific data development and sharing [J]. *China Basic Science*, 2007, 9(2): 30-35.

[11] Liu C, Sun HL. Research on frontier development of international scientific and technical data [J]. *China Basic Science*, 2003, 18(1): 329-333.

[12] Barabási AL, Jeong H, Néda Z, et al. Evolution of the social network of scientific collaborations [J]. *Physica A: Statistical Mechanics and its Applications*, 2001, 311(3/4): 590-614.

[13] Newman MEJ. Erratum: Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality [J]. *Physical Review E*, 2006, 73(3): 039906.

[14] Yang H, Wang W, Wu Z. Diversity-optimized cooperation on complex networks [J]. *Physical Review E*, 2009, 79(5): 56107.

[15] Abbasi A, Hossain L, Leydesdorff L. Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks [J]. *Journal of Informetrics*, 2012, 6(3): 403-412.

[16] Grit L. What do we measure by co-authorships? [J]. *Research Evaluation*, 2002, 11(1): 3-15.

[17] Meyer M, Bhattacharya S. Commonalities and differences between scholarly and technical collaboration: An exploration of co-invention and co-authorship analyses [J]. *Scientometrics*, 2004, 61(3): 443-456.

[18] Singh J. Collaborative networks as determinants of knowledge diffusion patterns [J]. *Management Science*, 2005, 51(5): 756-770.

[19] Chen XY. Comparative study of author social networks in academic and WEB datasets [J]. *Information Science*, 2014(5): 79-84.

[20] Pajek [EB/OL]. [2017-04-15]. http://www.pajek.imfm.si/doku.php?id=pajek.

[21] UCINET [EB/OL]. [2017-04-15]. http://www.analytictech.com/ucinet/.

[22] Thelwall M, Kousha K. Are citations from clinical trials evidence of higher impact research? An analysis of ClinicalTrials.gov [J]. *Scientometrics*, 2016, 109: 1-11.

[23] Newman MEJ. The structure of scientific collaboration networks [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2001, 98(2): 404-409.

[24] Newman MEJ. The structure and function of complex networks [J]. *SIAM Review*, 2003, 45(2): 167-256.

[25] Wang WJ, Yuan H. A new perspective on evaluating academic productivity of social science papers: The concept, construction method, and preliminary test of the C100 index [J]. *Shandong Social Sciences*, 2015(2): 186-192.

[26] Qiu JP, Qu H. Knowledge diffusion in China's scientific research institution collaboration networks: A case study of biodiversity research [J]. *Library and Information Knowledge*, 2011(6): 5-11.

Author Contributions: - Xu Xiaojie: Framework design, writing, and revision - He Lin: Topic selection and revision suggestions - Shao Bo: Topic design,

revision suggestions, and final approval

Note: Figure translations are in progress. See original paper for figures.