COMP0123 COMPLEX NETWORK AND WEB

Analysis of Scientific Collaboration Trends in Papers with Fewer Authors

Zhou Yang[a]

Department of Computer Science, University College London

ARTICLE HISTORY

Compiled January 11, 2020

ABSTRACT

In this report, we mainly analyze the scientific collaboration trends in papers with fewer authors. Unlike the other work, we do not investigate how a collaboration network evolves. We analyze papers published in different year separately instead to identify the changes in scholar collaboration patterns. We counted the percentage of computer-related papers with different numbers of authors from 1950 to 2017. We found a very clear conclusion: the proportion of papers with more collaborators is increasing while the proportion of papers with fewer collaborators is decreasing. This trend is expected to continue. To explore the changes in the cooperation pattern in detail, we calculated the degree distribution of more than 100 scientific cooperation networks. We can conclude that even if only considering papers with fewer authors, scholars tend to collaborate with more people. Through calculation and analysis of the rich club coefficient, we cannot conclude how cooperation patterns change among low-degree nodes, but we can draw the following inference: The cooperation between rich nodes is becoming more frequent.

KEYWORDS

Complex Network; Scientific Collaboration Network; Rich Club Phenomena

Contents

1. Introduction

1.1. Background: Increasing Scientific Collaboration

It is a common sense that scientific collaboration has been increasing. There are many reasons for such increase. While the Internet connects the world, it also connects scholars and makes it more convenient for them to collaborate. More and more tools are available for scholars, e.g. Overleaf for writing drafts and Github for contributing to codebase. Besides, much scientific research nowadays is interdisciplinary and requires researchers from different domains to collaborate.

On the other side, the number of researchers (e.g. PhDs) is increasing, and many of them want to apply for competitive positions. As an optional measurement for the performance of scientists, the number of papers published can increase via scientific collaboration at a relatively low cost. Scholars often tend to participate in the work of different groups to make their resume more outstanding. This kind of collaboration is also double-way: you join my papers, and I should join your other work. This is a reciprocal relationship if not consider other factors. The young scholars hope to collaborate with famous scientists to boost their reputation. The scientist of fame can also produce more research achievements. We do not judge whether such motivation is fair, but if we view scientific collaboration network as a market, tending to collaboration more seems to be a dominant strategy. Besides, the success of scholars is not only decided by the number of papers and citations. The networks that scholars are in also influence whether scholars are successful or not. These factors all contribute to the fast increase in scientific collaboration.

1.2. Research Questions of this Report

The main research object in this report is the patterns of trends in scientific collaboration. Many researchers view the collaboration relationship as Simple Graph. Nodes in the graph represent scholars. If two scholars co-published a paper, there is an edge between corresponding nodes. With regarding this problem, the most classic paper is written by Newman. Newman analysed the collaboration patterns of scholars from different domains and studied some measurements, including degree distribution, betweenness, giant component and so on. He found the typical pattern existing in various science fields and such patterns also exist in current collaboration networks [1].

However, this research work usually did not focus on the co-authorship of papers with different numbers of authors. A sample question could be: do the power-law phenomena still exist if we only consider the papers with authors less than 4? The idea to analyse collaboration trends in papers with fewer authors originates in our thoughts that since academic cooperation is increasing, for example, the proportion of papers with more collaborators is rising, the papers fewer collaborators with must have also changed. We can know through simple data analysis that the percentage of

papers with fewer collaborators has been decreasing. But knowledge about the changes in collaboration patterns requires an in-depth analysis of the collaborative network of these papers. This is exactly the research focus of this report.

Another reason we want to focus on fewer collaborator papers is that there are many papers with a very high number of collaborators. Consider a paper with 50 collaborators. When we put it into a cooperative network, there will be 50 nodes with a degree of at least 49 in the network. Even if one of the authors is a student who has published only one paper, he will be a node with a degree of 49 and a clustering coefficient of 1 in the network, which looks like a super academic star. These phenomena are all we try to avoid.

In this report, we mainly investigate in the following research questions:

(1) How has the proportion of papers with different numbers of authors changed over these decades?
(2) For papers with less than 6 collaborators, how do the indicators (such as Power Law) in the collaboration network change?

1.3. Summary of Results

We extract the data needed from a larger dataset [9] containing all the computer science papers published before 2018. More than 100 scientific collaboration networks were constructed. We used netowrks [10] computed degree distribution, average clustering coefficient and rich-club coefficient for each of them. We also provide a very detailed Jupyter Notebook for people who are interested in reproducing our results. We use Gephi [11] to visual one network and display the figures in Appendix. We compare the results and answer the research questions above. We find that

(1) The proportion of papers with more collaborators has been increasing, and the proportion of papers with fewer collaborators is decreasing.
(2) Even if we only consider papers with fewer ($\leq$ 5) authors, scholars tend to collaborate with more people.
(3) The cooperation between Rich nodes is becoming more frequent

We also have to confess that our conclusion is empirical and lacks of strict mathematical approval.

1.4. What I Have Learned from the Project

Thanks for this coursework, I get deeper understanding of main concepts and conclusions in complex network field. I also get valuable practical experience in literature review, Gephi usage, data collection and data processing. Personally speaking, this topic, scientific collaboration is quite important to me since I want to be a PhD after MSc. I could apply the conclusion to my real academic life.

1.5. Structure of the Report

The structure of this report is as follows: In Section 2, we mainly discuss the research work related to this question. In Section 3, we introduce how to extract and pre-process the data. In the fourth section, we show the results of the calculations and analyze the changes in the cooperation pattern. We conclude this report in the final section.

2. Literature Survey

In this section, we mainly introduce the research papers related to our topic, including common statistical properties of scientific collaboration networks, domains that are analyzed and work on evolution of collaboration network.

Newman first studied the structure of scientific collaboration networks in different research fields. He showed that these collaboration networks form small worlds [1] . In this paper, many indicators were investigated, including: number of authors, mean papers per author, number of collaborators (node degree), the giant component and average degrees of separation. Then Newman analyzed more detailed statistical properties of scientific co-authorship networks [2] , including size of giant component, closeness, betweenness and clustering coefficient.

Many papers were inspired by Newmans initial work. People started to analyze the co-authorship in different domains. Hou et al. studied the structure of scientific collaboration networks in Scientometrics [3] . Tomassini et al. paid attention to genetic programming collaboration network [4] . Some researchers used countries to separate the network, e.g. collaboration network in Turkey [5]. So one kind of follow-up work centres in applying complex network theory into different science domains.

Another important research stream is to identify how the scientific collaboration networks evolve. Barabsi et al. [6] are the first to infer the dynamic and the structural mechanisms that govern the evolution and topology of this complex system. Tomassini [7] studied the genetic programming network and found that degree distribution tends to stabilize toward an exponentially truncated power-law. Huang et al analyze different fields in computer science and found that major observations are that the database community is the best connected while the AI community is the most assortative. Lara-Cabrera et at [8] chose an interdisciplinary domain in computer science as research objects and also observed sub-linear preferential attachment for new nodes.

What we have done in this report has some connections with evolution of scientific collaboration networks. But we are not studying one network, no matter static one or dynamic one. We view the collaboration networks of each year separately to identify the changes in human behaviors.

3. Data and Methodology

3.1. Data Preparation

In this section, we are going to introduce how we extract and pre-process our data.

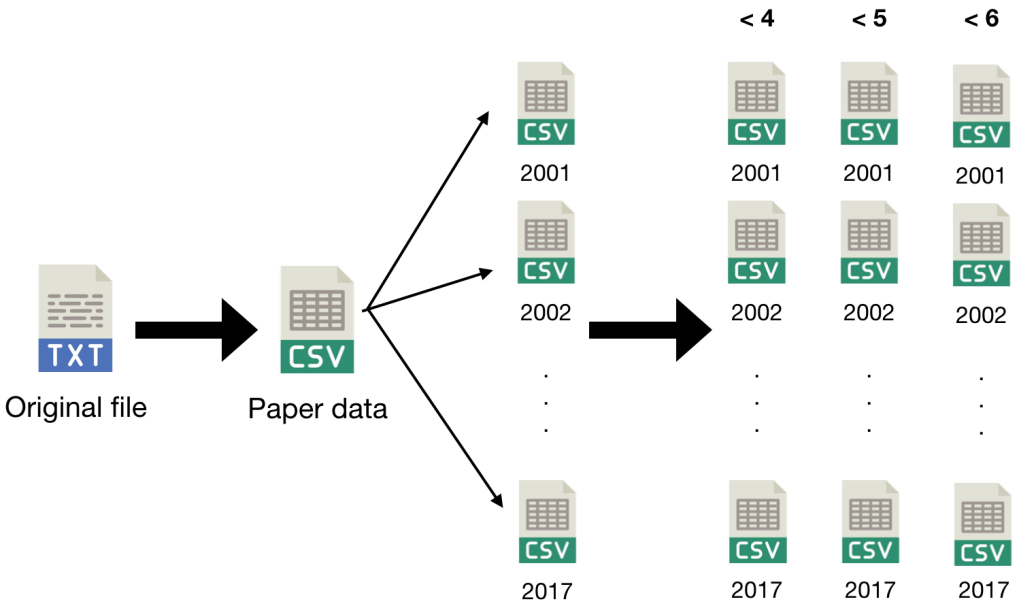

Figure 1. The process to extract and divide data

The original data we use is Citation Network Dataset [9] . The citation data is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. The dataset contains more than 4 million papers published within 1890 and 2018. Each paper and each author will have an ID to identify them uniquely. This dataset also provides rich metadata, such as the title of the paper, the name of the conference, the publisher, etc., but we do not need these irrelevant data. All we need is the publication year and its author ID.

Our hypothesis for constructing a network is that if four authors co-author a paper, then we assume that the four authors know each other and they form a fully connected graph. We do not rule out some papers with multiple versions, but this has not affected our analysis much. Because in our network, the cooperative relationship between scholars is dual. Even if the same paper was included twice for some reason, it had no effect.

Our goal is to extract from the raw data collaboration networks formed by all papers with fewer authors published in a particular year. To make our research more reliable, we constructed three sets of data: a collection of papers with less than 4 collaborators, a collection of papers with fewer than 5 collaborators, and a collection of papers with

fewer than 6 collaborators. The specific extraction process is as follows:

(1) Convert original data (.txt file) to a csv file. Each row represents a paper, including paperID, publication year, and IDs of all authors.
(2) Separate the csv file in Step 1 to multiple csv files according to publication year.
(3) For the file of each year, extract three csv file:
    (a) Set of papers with less than 4 authors
    (b) Set of papers with less than 5 authors
    (c) Set of papers with less than 6 authors
(4) In the last step, we get $3 * year\_range$ csv files. We convert these files to network format.

Although the original data source provides paper data published between 1890 and 2018, we only choose papers from 1975 to 2017 as our research object since many old papers might not be contained in the dataset and affect the data analysis results.

3.2. Methodology

In this section, we briefly describe how we analyze this data, as well as related tools and third-party libraries.

We use Python's third-party library networkx [10] to calculate the relevant properties in scientific collaboration networks. Networkx provides rich APIs to calculate network attributes such as degree distribution, rich-club coefficient, betweenness centrality and more. It also provides simple network visualization functions, but it is not convenient to plot the properties, so we wrote Python scripts to visualize data with the matplotlib library.

Another commonly used tool in the field of complex networks is Gephi [11]. Gephi is a very powerful tool with an easy-to-use interface. Not only can users calculate attributes with just a few clicks, but Gephi can also plot this data well. But our research object is 120 networks. Gephi does not provide a suitable function for comparing attributes in different networks, so we chose to write our own script to process the data. We provide a very detailed Jupyter Notebook for people who are interested in reproducing our results.

4. Analysis of Trends in Network

4.1. Changes in the Proportion of Papers with Different Numbers of Authors

The first question we have to answer is, is academic cooperation increasing? We counted the number of papers in different years and the number of collaborators and their proportion in the total number of papers in that year. We counted the proportion of papers with different numbers of authors between 1950 and 2017 and plot the data

in Figure 2. The x-axis represents the year, and the y-axis represents the proportion. Data of papers with different numbers of authors are displayed in different colours.

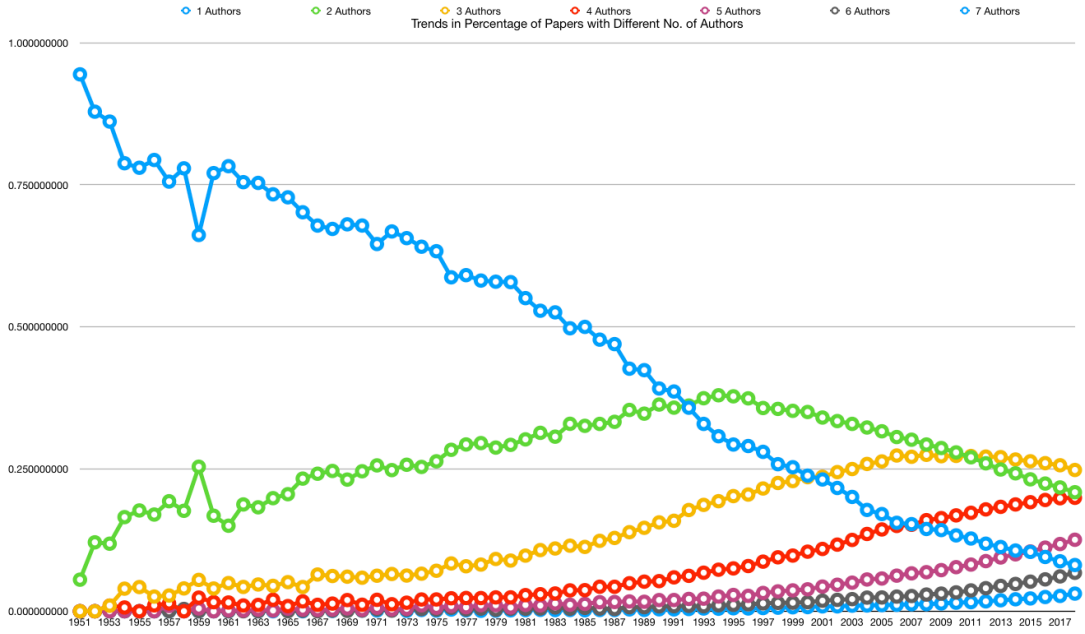

Figure 2. Trends in the percentage of papers with different numbers of authors, from 1950 to 2017.

xx

We can assert from this picture that cooperative academic papers in the computer field have been increasing since 1950. It can be seen from the figure that around 1950, the proportion of papers without collaborators was very high(close to 90%). But until 2017, the percentage of independent authors' papers dropped almost uniformly, to almost 7%.

Corresponding to this is the continuous growth of multi-collaborative papers. The papers of the dual collaborators continued to grow until they peaked (35%) around 1995 and then began to decline. Although the years to the peak were different, the percentage change in the papers of the three collaborators showed a similar pattern: it grew to around 2007 and then began to decline. From the figure, we can also observe that the proportion of the papers of the four collaborators also peaked in 2017 (slope 0), and it is likely that the proportion of the papers of the four collaborators will start to decline in the future.

The decline in the proportion of these papers with fewer authors also means an increase in the proportion of papers with more authors. As shown in the figure, the proportion of papers with 5, 6, 7 collaborators is increasing (the slope of the curve is also increasing).

Although we can't predict the end of growth, this chart very clearly illustrates that academic cooperation in the computer field has been increasing and will continue to

increase.

4.2. Statistical Properties of Collaboration Networks Categorized by Number of Authors

From a macro perspective, academic cooperation has been growing. But let's take a look at how academic cooperation networks have changed. We mainly focus on the following properties:

- Degree Distribution
- Clustering Coefficient
- Rich-Club

4.2.1. Degree Distribution

We want to examine how the degree distribution of these networks looks like. We calculated the degree distribution of cooperation networks composed of different papers (number of authors$\leq$ 3, 4, 5) from 1975 to 2017. For more clear visualization, we only illustrate 6 years of network data in the following figure. We have observed significant power-law phenomena for any year.

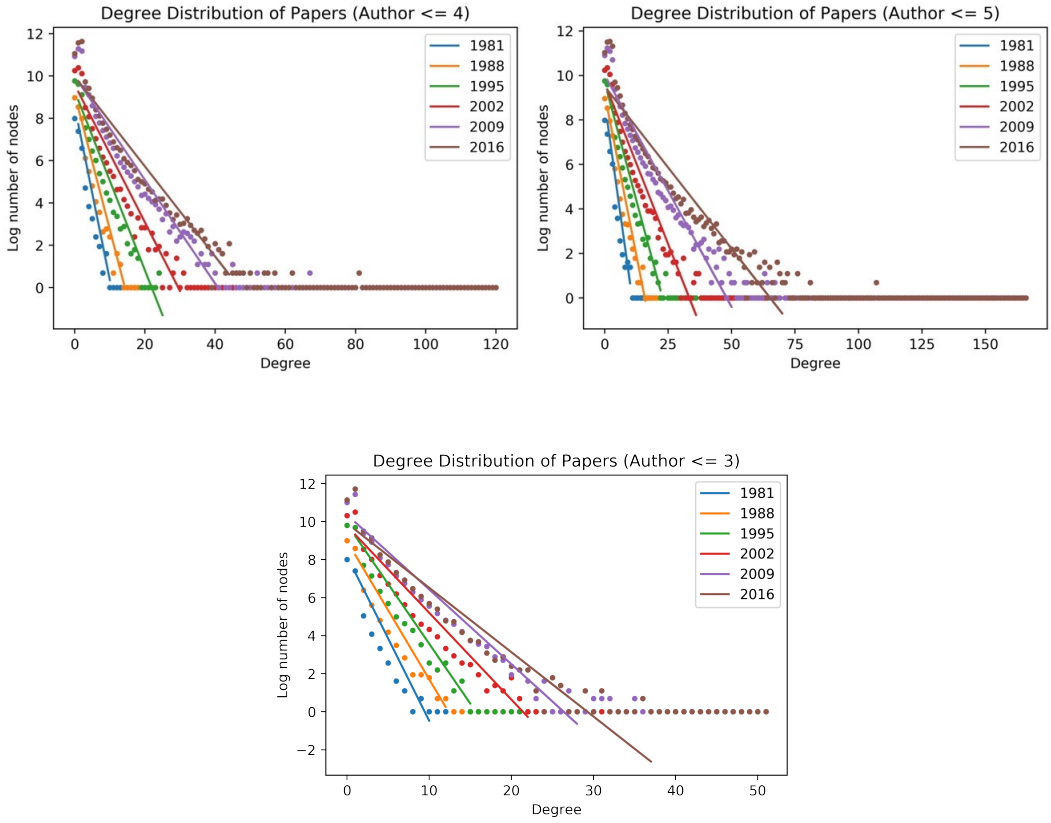

Figure 3. Degree Distribution, from 1975 to 2017.

We fitted these data using univariate linear regression and observed an interesting phenomenon: despite fluctuations, the slope of these curves generally increased with increasing year. If you compare slopes of different kinds of papers, you can always observe that the slope increases with the number of authors. If we look at Figure 1 and Figure 2 together, we can find that there is no significant correlation between the proportion of (number of authors $\leq$ 3, 4, 5)papers and the slope of power-law phenomena.

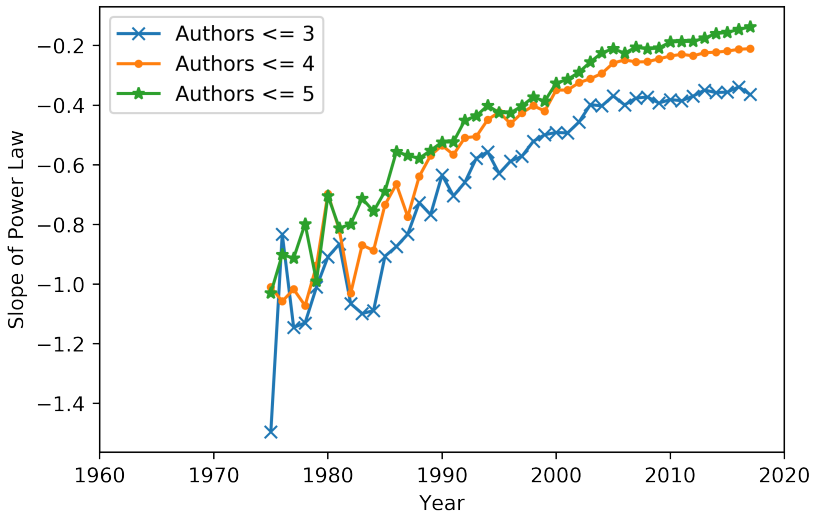

Figure 4. Trends of the power-law slope

We believe that the gentler slope indicates that scholars are more inclined to cooperate with more people. Let us give a simple example to illustrate.

We only consider the network composed of papers with 3 or fewer authors. If researchers are only working with fixed people, no matter how many papers they publish, their only neighbours are each other.

The degree of each point in this network is usually 2 or 3, and the network will be filled with a large number of components that are not connected. If we plot the degree distribution, we get a very steep straight line.

If they start to work with different people, then more people will have a higher degree; in other words, the curve will become flattered. We can, therefore, point to a change in the pattern of cooperation: scholars tend to collaborate with more people.

4.2.2. Average Clustering Coefficient

We also calculate the average clustering coefficient of these networks. The clustering coefficient is defined as:

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \tag{1}$$

k_i is the degree of node i, and e_i is the number of edges among i's neighbours. It measures the connectivity among a node's neighbours. For example, in social network, how your friends know each other. Evidence shows that in various types of real world networks, especially social network structures, clusters tend to form between nodes. The average clustering coefficient is defined as:

$$C = \frac{1}{n} \sum_{i=1}^{n} C_i \tag{2}$$

We plot the trends in average clustering coefficient in Figure 5.

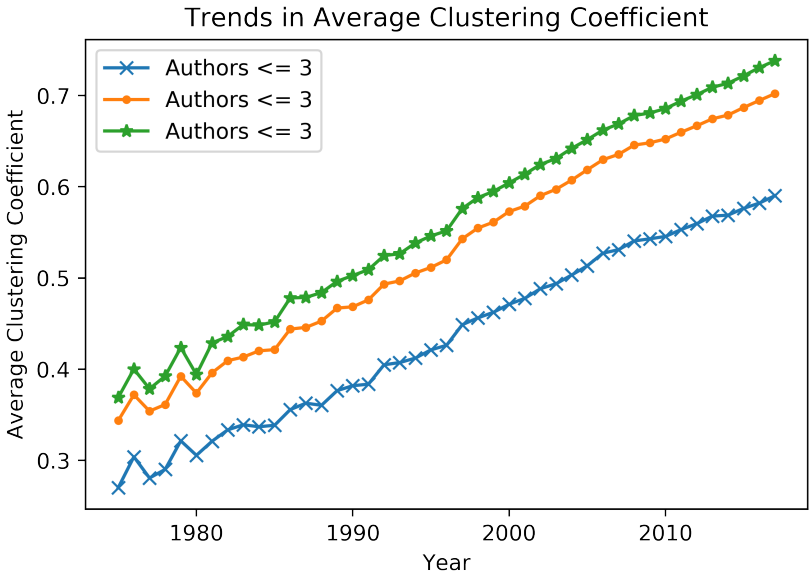

Figure 5. Trends of the average clustering coefficient

From the figure, we can clearly see that the average clustering coefficient all increases with the year for each network composed of papers with different numbers of collaborators

4.2.3. Rich-Club Coefficient

We study the Rich-club phenomenon in the scientific cooperation network to analyze how the superstars in the academic field cooperate with each other.

Rich-club phenomenon is first discovered by Zhou [12] in the Internet topology. The rich nodes are a small number of nodes with large numbers of links and are very well connected to each other. In the original paper, the rich-club coefficient was defined as:

$$\phi(r) = \frac{2E(r)}{r(r-1)} \tag{3}$$

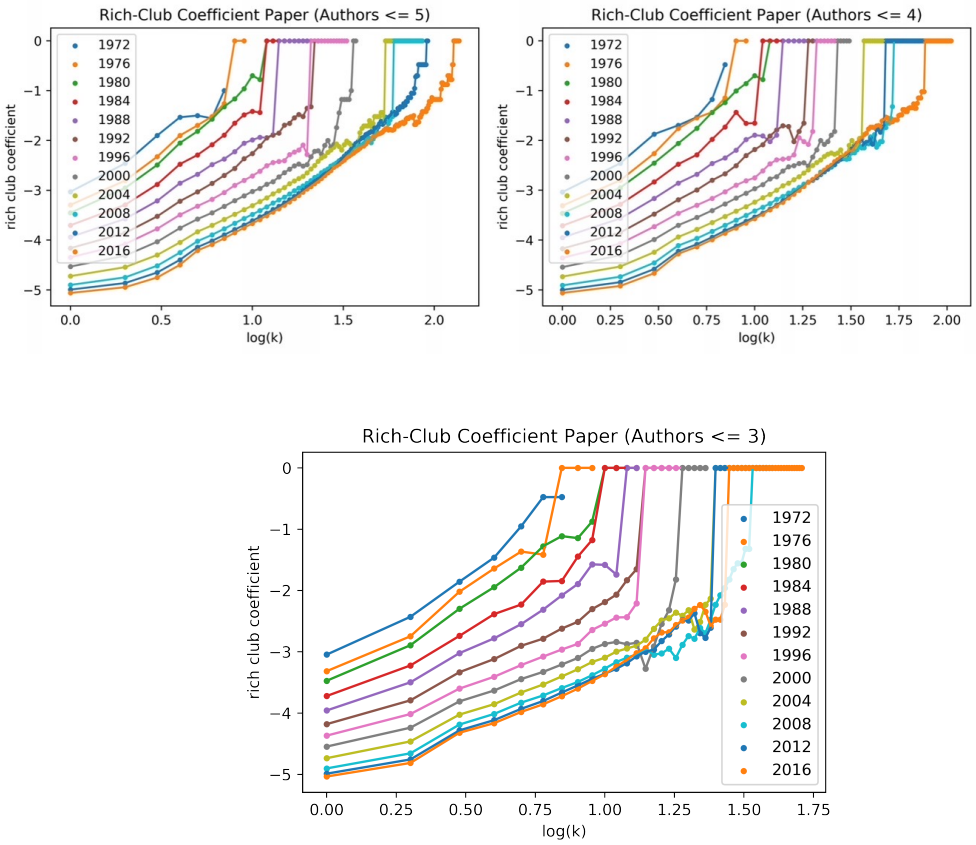

Figure 6. Trends in the rich-club coefficient, from 1975 to 2017.

In this equation, r is a node's position in a sorted list of decreasing degrees. Modified definition replace rank r with node degree k. For each degree k, the rich-club coefficient is the ratio of the number of actual to the number of potential edges for nodes with degree greater than k [10].

$$\phi(k) = \frac{2E(k)}{k(k-1)} \tag{4}$$

In our report we use the latter one mainly because the networkx library also uses this definition.

We calculated the Rich-club coefficient of the scientific cooperation network composed of different types (number of authors $\leq$ 3, 4, 5) papers from 1975 to 2017. For more clear visualization, we illustrate a network rich-club coefficient every four years in Figure 6.

From these figures, we can observe that in the collection of papers composed of 3, 4 or 5 authors, the phenomenon of rich clubs exists, and they showed a similar pattern of change.

(1) When the degree of nodes is relatively small, the rich-club coefficient of nodes

with the same degree decreases as the year increases.

(2) As the years increase, the difference in the rich-club coefficient decreases.

(3) In the ten years since 2008, the rich man's club coefficient curve has been very close.

We can explain such changes. We think the first point is mainly due to the increase in the number of papers and authors. Because this means that the number of nodes with a degree greater than k has increased dramatically. $k(k-1)$ grows much faster than $2E(k)$ when k is relative small. So the rich club coefficient of the nodes with the same degree decreases as the year increases. The difference in coefficient between adjacent years will also become smaller as the year increases, which is observed as point 2.

But more interestingly, for the nodes with a relatively high degree, the opposite has even happened. It means that for nodes with a higher degree, in other words, the rich nodes, $2E(k)$ is growing faster. From our simple analysis, we can't get the change of the cooperation pattern between low-level nodes, but we can draw the following inferences: **Cooperation between the rich nodes is becoming more frequent.**

5. Conclusion and Discussion

In this report, we mainly analyze the scientific collaboration trends in papers with fewer authors. Unlike the other work, we do not investigate how a collaboration network evolves. We analyze papers published in different year separately instead, which can better reflect the changing trend of academic cooperation behaviours in each year. Besides, we do not consider all the papers but divide papers according to the numbers of collaborators. Research on this particular subset is indeed temporarily lacking. We want to verify some common properties in scientific collaboration networks, such as whether Power law and Rich-Club phenomena are related to the number of collaborators in papers.

We counted the percentage of computer-related papers with different numbers of authors from 1950 to 2017. We found a very clear conclusion: the proportion of papers with more collaborators is increasing, and the proportion of papers with fewer collaborators is decreasing, and this trend will continue.

To explore the changes in the pattern of cooperation in more detail, we calculated the degree distribution of 120 scientific cooperation networks. We can conclude that even if only considering papers with fewer authors, scholars tend to collaborate with more people.

Through calculation and analysis of the Rich club coefficient, we cannot judge how cooperation patterns change among low-degree nodes, but we can draw the following inference: The cooperation between Rich nodes is becoming more frequent.

References

[1] M. E. J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci., vol. 98, no. 2, pp. 404409, Jan. 2001, doi: 10.1073/pnas.98.2.404

[2] M. E. J. Newman, Who Is the Best Connected Scientist?A Study of Scientific Coauthorship Networks, in Complex Networks, E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, Eds. Berlin, Heidelberg: Springer, 2004, pp. 337370.

[3] H. Hou, H. Kretschmer, and Z. Liu, The structure of scientific collaboration networks in Scientometrics, Scientometrics, vol. 75, no. 2, pp. 189202, Dec. 2007, doi: 10.1007/s11192-007-1771-3.

[4] M. Tomassini, L. Luthi, M. Giacobini, and W. B. Langdon, The structure of the genetic programming collaboration network, Genet. Program. Evolvable Mach., vol. 8, no. 1, pp. 97103, Mar. 2007, doi: 10.1007/s10710-006-9018-2.

[5] A. avuolu and . Trker, Scientific collaboration network of Turkey, Chaos Solitons Fractals, vol. 57, pp. 918, Dec. 2013, doi: 10.1016/j.chaos.2013.07.022.

[6] A. L. Barabsi, H. Jeong, Z. Nda, E. Ravasz, A. Schubert, and T. Vicsek, Evolution of the social network of scientific collaborations, Phys. Stat. Mech. Its Appl., vol. 311, no. 3, pp. 590614, Aug. 2002, doi: 10.1016/S0378-4371(02)00736-7.

[7] M. Tomassini and L. Luthi, Empirical analysis of the evolution of a scientific collaboration network, Phys. Stat. Mech. Its Appl., vol. 385, no. 2, pp. 750764, Nov. 2007, doi: 10.1016/j.physa.2007.07.028.

[8] R. Lara-Cabrera, C. Cotta, and A. J. Fernndez-Leiva, An analysis of the structure and evolution of the scientific collaboration network of computer intelligence in games, Phys. Stat. Mech. Its Appl., vol. 395, pp. 523536, Feb. 2014, doi: 10.1016/j.physa.2013.10.036.

[9] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998.

[10] NetworkX NetworkX. [Online]. Available: https://networkx.github.io/.

[11] Gephi - The Open Graph Viz Platform. [Online]. Available: https://gephi.org/.

[12] Shi Zhou and R. J. Mondragon, The rich-club phenomenon in the Internet topology, IEEE Commun. Lett., vol. 8, no. 3, pp. 180182, Mar. 2004, doi: 10.1109/LCOMM.2004.823426.

Appendix: Network Visulization (Year 2020, Authors $\leq$ 5)

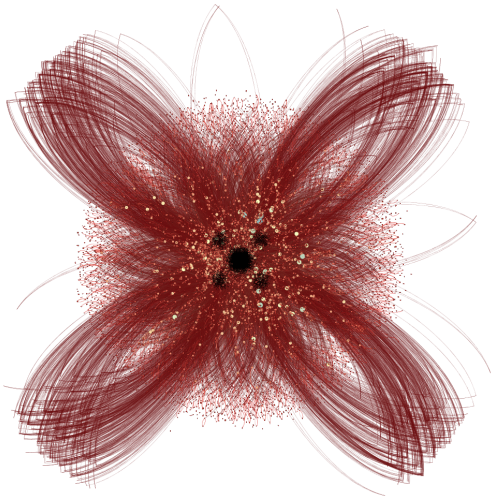

Figure 7. Network Visualization 1

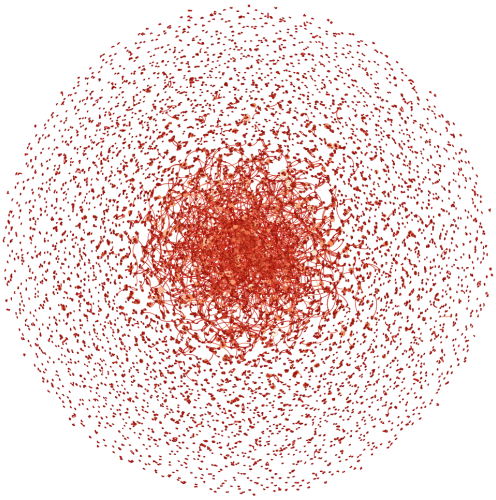

Figure 8. Network Visualization 2

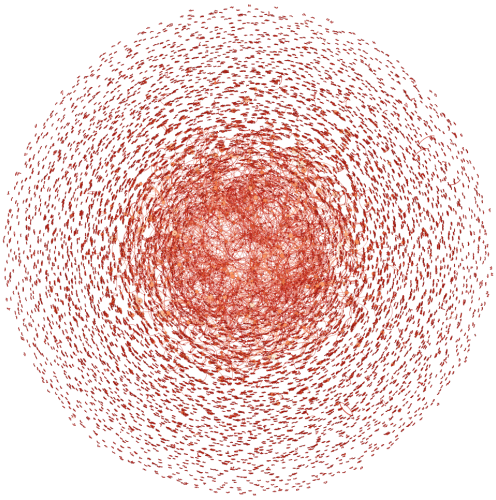

Figure 9. Network Visualization 3

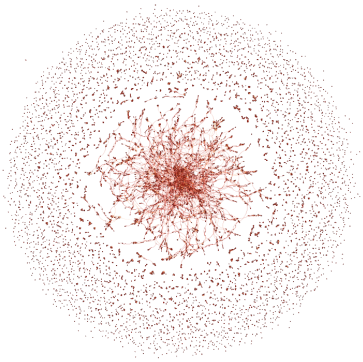

Figure 10. Network Visualization 4