DSI DATA SCIENCE AND INFORMETRICS
VOL.2, NO.4, NOV. 2022

RESEARCH ARTICLES

Analyzing the interdisciplinarity of Big Data research

Siluo Yang, Yue Zhang, Xiangxiang Shu, Wenjuan Xu

School of Information Management, Wuhan University, Wuhan, China

ABSTRACT

Big Data is a hot topic at present, and many disciplines are engaged in it. However, the relationship and status of these disciplines in Big Data research are still not well understood. Based on bibliometric data from the Web of Science, this study analyzes the interdisciplinarity of Big Data research. It focuses on its spatial and temporal distribution characteristics by combining traditional bibliometrics, interdisciplinarity indicators, and social network analysis. We found that: 1) In recent years, the research topics of Big Data include the Internet of Things, Syndromic Surveillance, Knowledge Management, Industry 4.0, etc. Most Big Data research comes from China, the United States, and India. 2) Big Data research involved 141 disciplines, but the disciplinary distribution is unbalanced, among which the diversity of disciplines is constantly improving; The balance of disciplines is on the rise, and the situation that is dominated by a few disciplines has been improved; The disparity of disciplines increases and researchers tend to cite the disciplines that are quite different from their own disciplines. 3) Computer Science and Engineering are the two most important disciplines in the Big Data field. According to the collaboration degree, the interdisciplinary collaboration network can be divided into five communities: Community 1, represented by Computer Science and Engineering; Community 2, represented by Business & Economics; Community 3, represented by Science & Technology; Community 4, represented by Materials Science; Community 5, represented by Health Care Sciences & Services. The collaboration between a few key disciplines has changed, and the whole network is still expanding in Big Data research.

KEYWORDS

Big Data; Interdisciplinarity; Social network analysis; Discipline variety

1 INTRODUCTION

1.1 Introduction to Big Data and Big Data Research

TechAmerica Foundation defines Big Data as: "Big Data is a term that describes a large number of high-speed, complex and variable data, which requires advanced technologies and processes to capture, store, distribute, manage and analyze information" (TechAmerica Foundation's Federal Big Data Commission, 2012). Big Data can also be defined as a vast data set with increasingly diverse and complex structures (Iqbal et al., 2020), characterized by mass, rapidity, diversification, and low-value density (Furht & Villanustre, 2016). However, the term has now been extended. Big Data is worthless in a vacuum, and its potential value will only be released when it is used to promote decision-making (Gandomi & Haider, 2015).

Now Big Data refers not only to large data volume but also to our increasing ability to analyze and interpret those data, including data collection, data storage and management, data processing and analysis, data privacy and security, etc. (Hulsen & Jamuar, 2019; Borjigin & Zhang, 2022). The technologies involved mainly include extensible storage systems, distributed file systems, databases, cloud computing, data mining tools, technologies, etc. (Oussous et al., 2018). Big Data applications involve many sectors, bringing unprecedented convenience to our life, such as in the fields of health care (Kankanhalli et al., 2016), smart grids, government systems (Stoianov et al., 2015), logistics systems, etc. In recent years, some scholars have explored more applications of Big Data, such as Big Data management in the mining industry (Qi, 2020), Big Data and the ethical framework of smart city (Chang, 2021), Big Data analysis of social media (Ghani et al., 2019), application of Big Data in the emotional analysis of tourism (Alaei at al., 2019).

As a multidisciplinary and interdisciplinary research field, Big Data research requires cooperation from various disciplines (Chen et al., 2014; Borjigin et al., 2021). Scholars from different disciplines have contributed their efforts and created the glory of Big Data. In recent years, The research on Big Data mainly focuses on the following four aspects: Firstly, it focuses on the technology of Big Data itself, concentrating on the improvement and innovation of technologies, methods, and tools related to Big Data. For example, Taleb et al. (2018) proposed a quality evaluation model to deal with the quality of unstructured Big Data; Sun and Wang (2017) put forward a possible mathematical theory as the basis of Big Data research. Secondly, it focuses on the application of Big Data in different industries; for example, Alhussain (2018) introduced how to use Big Data tools and methods to analyze medical Big Data; Caesarius and Hohenthal (2018) explored how can ordinary enterprises adopt Big Data technology to make changes and pointed out the challenges they will face. Thirdly, it focuses on opportunities and challenges brought by the development of Big Data. For example, Ashabi et al. (2020) conducted research on the current challenges and future development of Big Data; Gupta and Rohil (2020) pointed out that Big Data has hidden dangers such as privacy leakage, and he proposed some critical solutions to problems related to Big Data security and privacy. Fourthly, it focuses on standards and policymaking in the Big Data industry. Jia and Jia (2019) put forward the data model of information construction to promote the construction and implementation of information construction in colleges.

There are many interdisciplinary studies in the field of Big Data, but very few studies are about the interdisciplinarity of Big Data research. "Interdisciplinary research" refers to the practical activities involving two or more disciplines that cross the boundaries of known disciplines (Liu, 1993), such as the interdisciplinary research of Big Data and wireless channels (Zhang, 2016). However, "research of interdisciplinarity" takes interdisciplinary research as research subjects and conducts research about the characteristics of interdisciplinary research through bibliometric methods (Xiong & Fu, 2021; Chen et al., 2021). There are already some studies about the interdisciplinarity of Big Data research. For example, Hu and Zhang (2017; 2018) concluded that by 2015, there are 109 disciplines involved in Big Data research. Big Data research is highly interdisciplinary, involving multiple disciplines, but unevenly distributed. Using bibliometrics and visualization tools, Lv and Wang (2019) compared and analyzed the interdisciplinary development of Chinese and American Big Data research from 2009 to 2016. Zhang et al. (2018) put forward a method of extracting subject classification based on the address of co-author institutions to measure the interdisciplinary degree between differ-

ent institutions in scientific collaboration. Research on interdisciplinarity has been going on steadily. Jang et al. (2018) concluded that qualitative or quantitative research methods can be used to analyze interdisciplinarity. The limitation of the qualitative approach is that only small-scale case studies can be carried out, while the quantitative method can be used in interdisciplinarity research from three angles: authorship, subject matter, and citation literature. Bibliometric analysis, citation analysis, network analysis, and other methods can be chosen for analysis in quantitative research.

1.2 Objectives of this study

At present, only a few articles are about the interdisciplinarity of Big Data research, and the timeliness of these articles is not enough. Most of these studies focus on exploring the scope of disciplines involved in Big Data research and the core disciplines of Big Data. Most conclusions are general, without clarifying the collaboration pattern between disciplines and the space-time characteristics of interdisciplinarity in Big Data research. In this paper, quantitative research methods will be used to analyze the interdisciplinarity of Big Data research from three aspects: traditional bibliometrics, single index, and social network, combined with previous scholars' experience and methods, to get better results. Especially, this paper addresses the following problems.

1. What disciplines are involved in Big Data research? What are the main topics and countries in the field of Big Data?

2. What are the characteristics of discipline variety, balance, and disparity in Big Data research?

3. From the perspective of a co-discipline network, what is the relationship between disciplines involved in Big Data research?

2 Data and Methods

2.1 Data acquisition and processing

Many records contain "Big Data" in abstracts and author keywords only as general research background. We restrict title and author keyword to "Big Data" to ensure that the retrieved records are related to Big Data as much as possible. At the same time, the types of documents are restricted to "Article", "Proceedings," and "Review." The time range is from 2007 to 2021, with 14,081 documents. Full records and references of these documents are exported from the core collection of Web of Science in plain text format. To avoid the deviation caused by frequent updating of the database, all searches and data downloads were completed on January 10th, 2022.

In this study, we select the SC as the subject classification method for papers for social network analysis. In Web of Science, the SC indicates the research direction, called the Subject Category, while the WC indicates the category defined by Web of Science, which is called the Web of Science Category. When searching on the Web of Science, we know that the SC of all papers belonging to the same journal is the same, and so does the WC. Generally speaking, the SC, used as a discipline category, is an accurate and straightforward analysis unit to describe the disciplines involved in the research field. It can be proven by previous studies, like the research of Rafols and Meyer (2010) and the study of Taskin and Aydinoglu (2015).

Since each paper belongs to one or more disciplines, a co-discipline network can be built according to the co-occurrence relationship between disciplines. The nodes in the network represent disciplines. The number of occurrences of disciplines determines the size of nodes, and the thickness of lines between nodes is determined by the number of co-occurrences between these two different disciplines. According to the co-discipline network in the field of Big Data, we can know what disciplines are involved in Big Data research, and the relationship between these disciplines.

2.2 Methods and tools

To analyze and visualize the spatial and temporal distribution characteristics of the Big Data research. The tools mainly include SCI2, Pajek, VOSViewer, and WC19.exe. Among them, WC19.exe is used to calculate the indexes of discipline variety, balance, and disparity, SCI2 and Pajek are applied when making network analysis, and VOSViewer is used as a visualization tool.

Steps of social network analysis: firstly, convert the text data exported from the core collection of Web of Science into ISI format data, then import it into SCI2; secondly, analyze the Subject Category field and get the co-discipline network (network data file) in Big Data research; thirdly, time slice the network, and remove the isolated nodes at the same time; the fourth step is to export the co-discipline network from SCI2, and then import it into Pajek for social network analysis, such as cluster analysis and analyzing the degree centrality, betweenness centrality and closeness centrality of the network; finally, export the network from Pajek to VOSViewer for visualization.

Steps to calculate the indicators of discipline variety, balance, and disparity: Step 1, divide the text data derived from the core collection of Web of science by year, and form data sets in five time periods of 2007-2012, 2013-2015, 2016-2018, 2019-2021 and all years. Step 2, change the name of the data to something that WC19.exe can recognize. Step 3, analyze the data and export the result table by WC19.exe which includes Rao-Stirling diversity, True diversity, DIV, DIV*, Gini-index, Simpson, Shannon entropy, H(max), Shannon, Variety, Disparity, Perc. H(max), N of WCs, N of WC u, etc. Step 4, select appropriate indicators.

3 Results and Discussions

This part mainly shows the research situation in Big Data from macro, meso, and micro levels. Firstly, we analyze the growing trend in the number of research papers in Big Data and extract the involved disciplines based on the SC. Then, according to the Citation Topics classification system and co-authorship network of countries, we grasp the distribution of research topics and the collaboration among countries in Big Data research from a macro level to ensure that a comprehensive and correct basic understanding of Big Data research can be established first. Secondly, by analyzing the variety, balance, and disparity of disciplines in the field of Big Data research, we know the distribution characteristics of these disciplines, such as whether the number of disciplines is monotonous or diverse, whether the contribution of each discipline is uniform or unbalanced, and whether the distance between citing and cited disciplines is far or near. We summarize the relationships among disciplines involved in Big Data research from an intermediate level, mainly describing the whole discipline network. Last but not least, we obtain the key disciplines in the co-discipline network by analyzing the network indicators such as degree centrality, betweenness centrality, and

closeness centrality. Then, we analyze the characteristics and evolution process of each collaboration community in the co-discipline network. These summarize the relationships among disciplines involved in Big Data research from the micro-level, mainly to depict key nodes and important cooperative groups in the network. These three parts respectively answer the questions raised in the rationale for this study.

3.1 Overview of Big Data research

3.1.1 The growth trend of research papers in Big Data and the distribution of disciplines

Figure 1 shows the growth trend of the number of disciplines and papers in Big Data research from 2012 to 2021. It can be seen that the published papers have been growing continuously, reaching the highest level in 2018, and then starting to decline slowly, which is consistent with the findings of Qiu Junping (Qiu, 2021). In his research, he concluded that it was an initial exploration period in Big Data during 2008-2011, while the period from 2012 to 2020 was a period of rapid growth. The growth trend is in line with the facts. The early 20th century was a turning point in Big Data research. In China, during this period, the Ministry of Science and Technology listed Big Data in the "973" basic research plan, and the Ministry of Industry and Information Technology also listed four technologies of Big Data as critical objects in the 12th Five-Year Plan. Big Data has gradually become a research hotspot for scholars in various fields. In foreign countries, the successful convening of the Big Data World Forum (BDWF) in 2011 and the official publication of Big Data written by British expert Viktor Mayer in 2013 both promoted and stimulated the research upsurge of Big Data (Wang, 2017).

Research in Big Data covers a wide range of disciplines. According to the statistics from 2007 to 2021, there are 141 disciplines involved in Big Data research. Only the data from 2012 to 2021 are shown in this section due to the small number of papers published before 2012. Before 2017, the number of disciplines involved in Big Data had been increasing and gradually stabilized from 2017 to 2021, basically around 102, which shows that the research in Big Data has stabilized after the outbreak and the development of the whole field is relatively mature. Table 1 lists 20 disciplines involved with more than 150 occurrences in Big Data research, ranked in the frequency of occurrence. To be more specific, Computer Science, Engineering, Telecommunications, and Business & Economics are the most important disciplines in the field of Big Data. More than 1,000 papers are published, accounting for 64% of the total occurrences of all disciplines. Social Sciences - Other Topics, Automation & Control Systems, Science & Technology - Other Topics, Information Science & Library Science, Environmental Sciences & Ecology, Education & Educational Research, Operations Research & Management Science rank high, with more than 400 papers published. Among the total occurrences of all disciplines, these 11 disciplines account for 78%, indicating the uneven distribution of disciplines in Big Data research. In the study of Hu and Zhang (2017), up to 2015, the leading disciplines in Big Data research are Computer Science, Engineering, Telecommunications, and Business & Economics, among which Computer Science and Engineering account for 55.67% of the total occurrences of disciplines. The result shows that the contributions of disciplines in Big Data research are still unbalanced from 2015 to 2021.

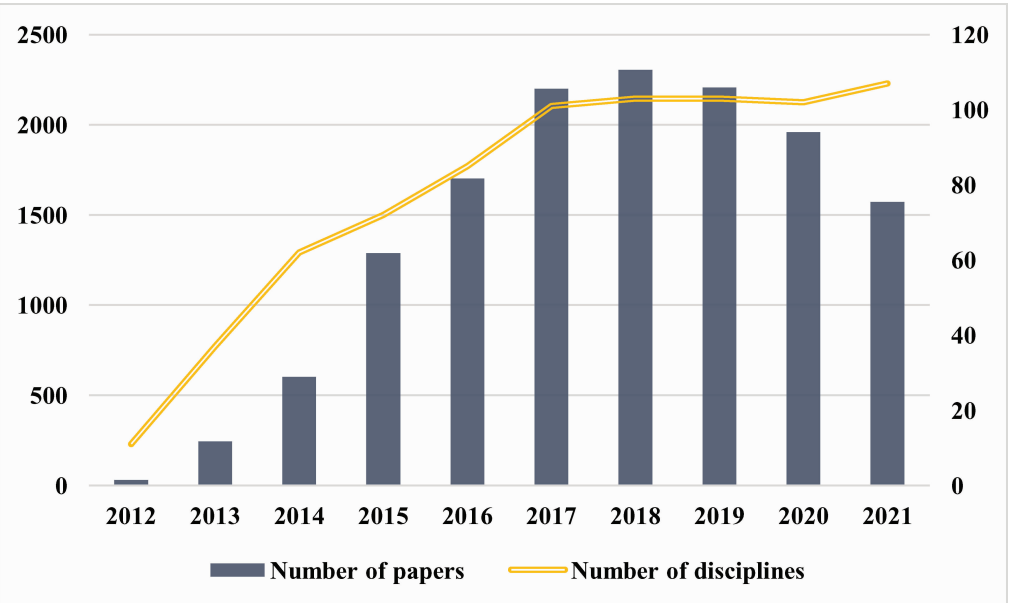

Figure 1 Number of disciplines and papers in Big Data research (2012-2021)

Table 1 Top 20 disciplines involved in Big Data research (2007-2021)

Rank	Discipline	The number of occurrences
1	Computer Science	7935
2	Engineering	4099
3	Telecommunications	1637
4	Business & Economics	1174
5	Social Sciences – Other Topics	511
6	Automation & Control Systems	501
7	Science & Technology – Other Topics	498
8	Information Science & Library Science	498
9	Environmental Sciences & Ecology	489
10	Education & Educational Research	479
11	Operations Research & Management Science	428
12	Mathematics	246
13	Public Administration	230
14	Remote Sensing	225
15	Energy & Fuels	209
16	Geology	207
17	Materials Science	186
18	Medical Informatics	174
19	Transportation	170
20	Health Care Sciences & Services	153

3.1.2 Distribution of research topics in Big Data

Most existing classification models are periodical-level classification systems, while Incites has developed a classification algorithm based on citations--Citation Topics. It is an article-level classification system focusing on citations among documents. The intensity of these citation relationships will bring related documents together to form discrete related document clusters. Citation Topics builds a three-level hierarchy of macro, meso, and micro topics, including 10 macro topics, 326 meso topics, and 2444 micro topics (An & Xiao, 2021). In this section, using the Citation Topics classification system, the topics of related papers in Big Data research are classified into three levels: macro, meso, and micro, to reveal the distribution of main topics in Big Data research step by step.

Macro topic distribution is based on Citation Topics. Research papers in Big Data are mainly distributed in three topics: Electrical Engineering, Electronics & Computer Science, Social Sciences, and Clinical & Life Sciences, as shown in Figure 2. Specifically, Social Sciences and Clinical & Life Science are the main application fields in the field of Big Data. Electrical Engineering, Electronics & Computer Science are the main driving force for the development of Big Data. More than half of the research in Big Data belongs to them, indicating a highly uneven distribution of topics. However, after 2018, the number of papers on this topic began to decrease, which is not only related to the year-on-year decrease in the number of documents published in Big Data research but also because more research in Big Data began to tilt toward other topics.

Meso topic is distribution based on Citation Topics. As shown in Figure 2, Distributed & Real Time Computing, Knowledge Engineering & Presentation, and Artificial Intelligence & Machine Learning are the three most essential meso topics in Big Data research. These three topics began to rise around 2012 and peaked around 2017, after which their popularity slowly declined. Telecommunications, Management, Security Systems, Virology-Tropical Diseases, Design & Manufacturing, and Transportation are hot topics in recent years.

Micro topic distribution based on Citation Topics. Figure 4 shows the micro topic distribution (top 20) in Big Data research based on Citation Topics classification from 2007 to 2021, which is a finer-grained representation than meso topic and macro topic. The micro topic is named according to the most important author keywords by algorithm tools, with the so-called "importance" determined by the number of occurrences and co-occurrences. Since the term Cloud Computing emerged in 2012, it has been the hottest topic in Big Data, reaching its climax in 2017. While the popularity of Cloud Computing has been declining since 2018, it remains a critical topic in Big Data. In addition, the Internet Of Things, Syndromic Surveillance, Knowledge Management, and Industry 4.0 have become hot topics in Big Data in recent years.

Qiu (2021) used the LDA model to cluster the literature on Big Data. He found that the research hotspots of Big Data in China mainly focus on the application level, that is, library services, smart city construction and intelligent urban transportation, education, e-commerce, network marketing, etc., which is consistent with the conclusion of this section. Looking at the topic distribution at three levels in Big Data research, it is evident that the development and breakthrough of Big Data technology and the application of Big Data are still constant topics despite the changing research hotpots. In addition, the differences in the number of papers among research topics are also shrinking, which means that the distribution balance among topics is constantly improving.

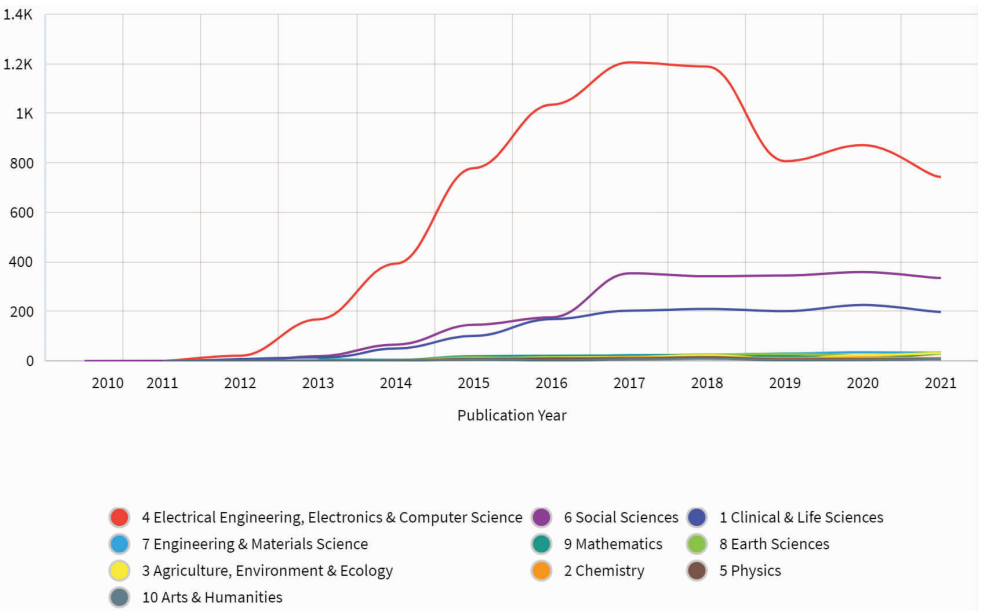

Figure 2 Annual paper number change of macro themes related to Big Data based on Citation Topics classification from 2007 to 2021

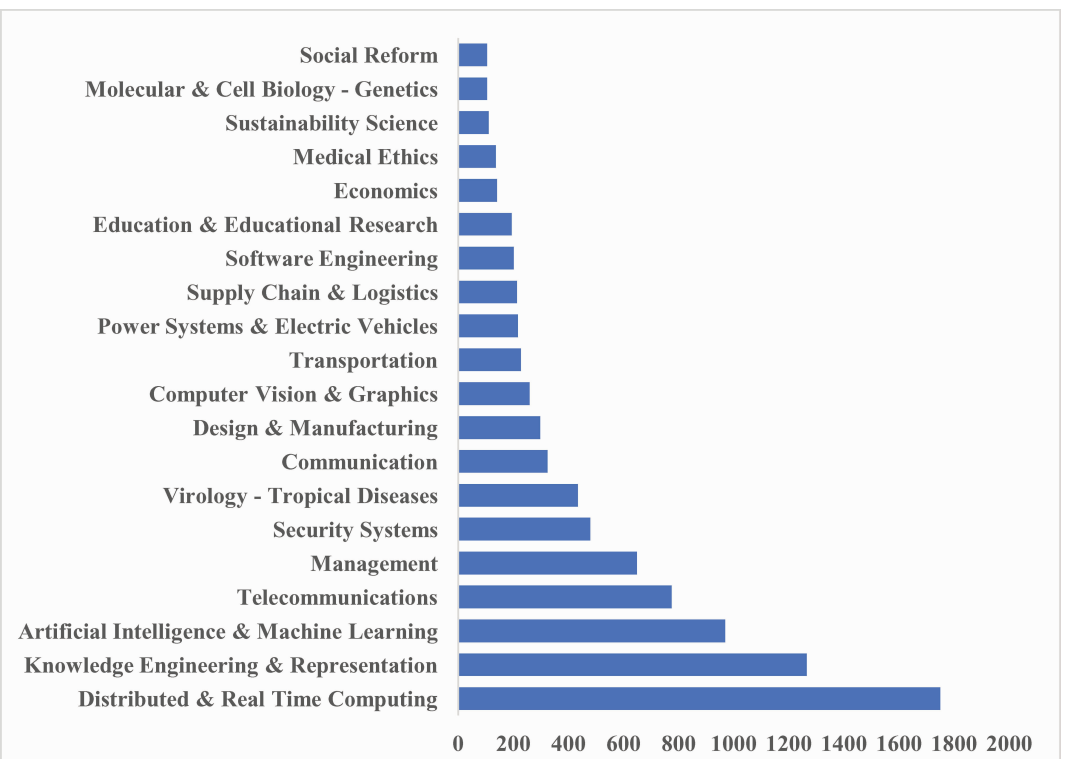

Figure 3 Top 20 meso themes related to Big Data based on Citation Topics classification from 2007-2021

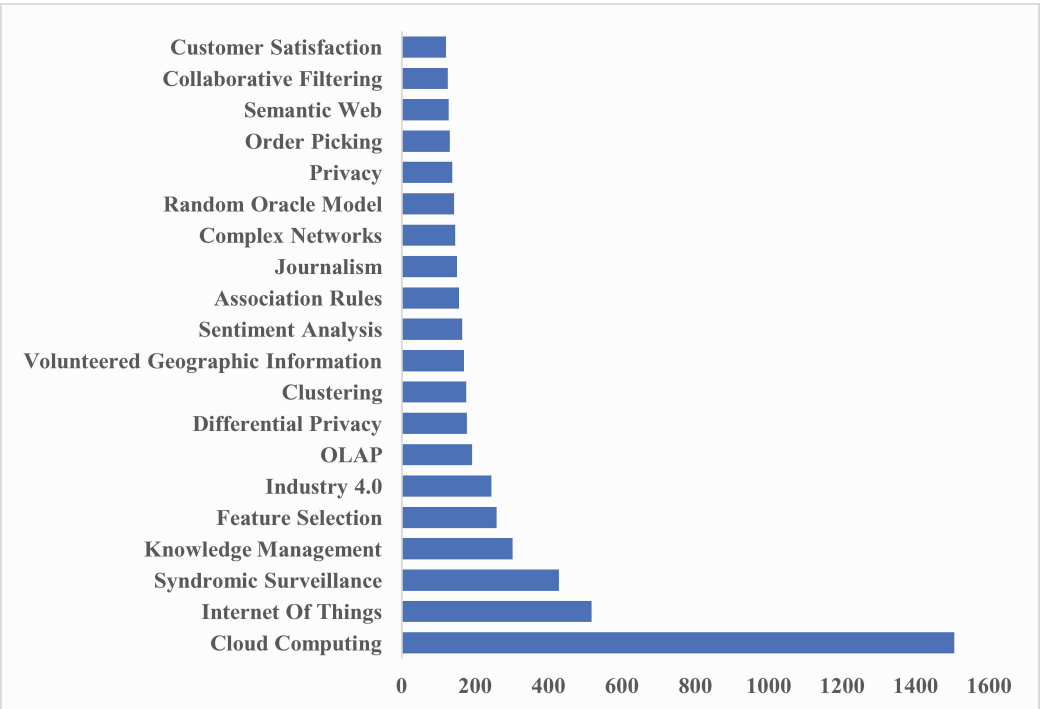

Figure 4 Top 20 micro themes related to Big Data based on Citation Topics classification from 2007-2021

3.1.3 Collaboration among countries in Big Data research

To understand the collaboration among countries in Big Data research, we make a network of co-authorship of countries in Big Data research from 2007 to 2021 and set a minimum number of papers published at 10, with 80 countries meeting the requirements. Table 2 shows the top 10 countries in Big Data research from 2007 to 2021, and Figure 5 is an overlay visualization of co-authorship of countries in Big Data research from 2007 to 2021. In Figure 5, nodes represent countries with a corresponding number of occurrences. The more occurrences, the larger the node. The connection between nodes indicates that the two countries have cooperated in publishing posts. The more articles published in cooperation, the thicker the line between the two nodes. Years from far to near are indicated from blue to yellow. The closer the color is to yellow, the later the publication time.

As the top three countries in terms of publications from 2007 to 2021, there is a significant disparity in the number of papers published between China, the United States, and India. The top three countries accounted for 60.7% of all research. At the same time, China took possession of 34.6%, which shows a highly uneven distribution of the number of papers published among countries in Big Data research. In addition, more than 400 articles have been published in England, Australia, South Korea, Italy, Spain, Canada, and Germany. China works most closely with the United States in Big Data research, with 484 articles co-authored, accounting for about 10% of the articles published in China and about 20% of the articles published in the United States, followed by 168 articles with Australia, 120 articles with England and 114 pieces with Canada. Besides, Saudi Arabia, Pakistan, Iran, Brazil, United Arab Emirates, Switzerland, and other countries have also been active in Big Data research in recent

years.

Over time, the scale of the international collaboration network is constantly expanding and gradually transformed into a multi-dominant mode: China, the United States, India, and England occupy the core position in the international collaboration network and play the role of a bridge; Chinese authors are highly dominant in the global collaboration network. Using social network analysis and scientific map, Lv (2021) revealed the collaboration mode of countries in Big Data research and drew more profound conclusions: global Big Data research focuses on domestic cooperation, especially intra-institutional collaboration, with a low proportion of collaboration between countries and institutions. International collaboration and cross-institutional collaboration are possible directions to promote the leap-forward development in Big Data.

Table 2 The number of papers of the top 10 countries in Big Data research (2007-2021)

Country	Documents	Proportion
China	4870	34.6%
USA	2469	17.5%
India	1231	8.7%
England	780	5.5%
Australia	590	4.2%
South Korea	550	3.9%
Italy	547	3.9%
Spain	542	3.8%
Canada	489	3.5%
Germany	416	3.0%

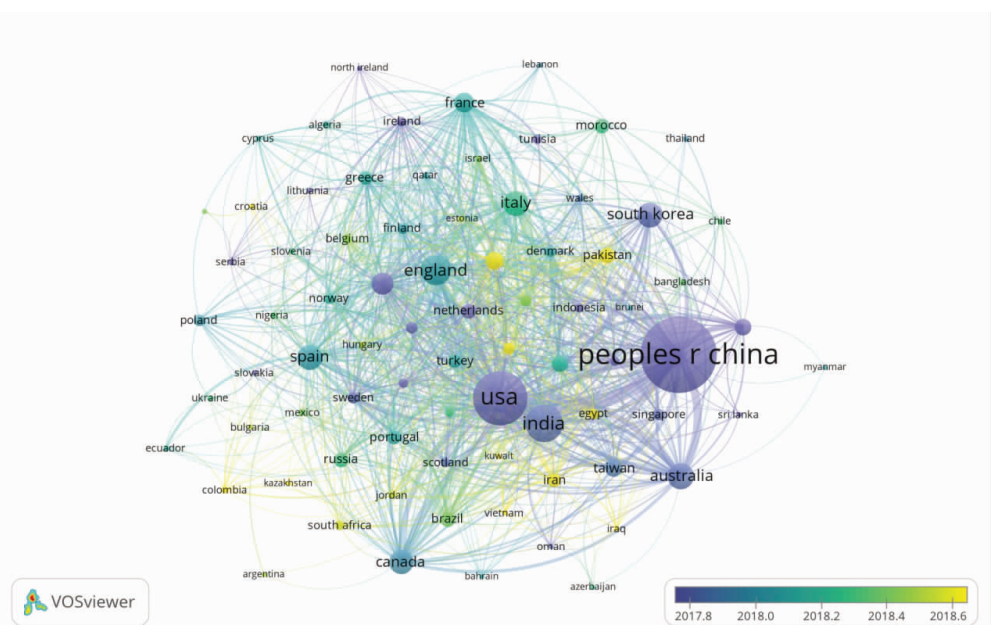

Figure 5 Overlay visualization of co-authorship of countries in Big Data research (2007-2021)

3.2 Three –dimensional index evaluation of discipline variety, balance, and disparity

3.2.1 Variety

The definition of interdisciplinary emphasizes "integrating the knowledge of two or more disciplines", so the variety index measures whether the disciplines involved in interdisciplinary research are various. Table 3 is the result of analyzing the variety in the field of Big Data using WC19.exe, a tool measuring interdisciplinarity created by Professor Loet Leydesdorff. Simpson, DIV, and Variety are three indicators used to measure the variety of research. The researcher's papers are divided into five stages from the time dimension. As can be seen from the analysis results, the variety of Big Data research is constantly improving with time, and the three indicators all show an upward trend. Specifically, the Simpson index was 0.89 in 2007-2012, rising to 0.95 in 2019-2021. The DIV index increased from 0.003 in 2007-2012 to 0.12 in 2019-2021, and the variety index increased from 0.10 to 0.86, which shows that with the deepening of research in the field of Big Data, more and more disciplines are involved, and the variety index naturally increased. These also show that the coverage of Big Data research is constantly expanding, and many researchers in other disciplines have begun to apply Big Data to interdisciplinary research. Looking at All years, the Variety index accounts for 0.9 in all years, indicating that the disciplines involved in the research of the Big Data field account for 90% of all disciplines in WCS. The results of the analysis fully demonstrate the variety of disciplines in the study of the Big Data field. It can be seen that interdisciplinary research is quite common in Big Data research, and most disciplines have cross research with Big Data.

Table 3 Simpson, DIV and Variety index over time

Year	Simpson	DIV	Variety
2007–2012	0.89	0.003	0.10
2013–2015	0.91	0.05	0.56
2016–2018	0.93	0.09	0.79
2019–2021	0.95	0.12	0.86
All years	0.94	0.11	0.90

3.2.2 Balance

The balance index measures the difference between disciplines involved in a research field and reflects the contribution of various disciplines. For the interdisciplinarity of a field, the higher the balance index, the more uniform the distribution of disciplines, and the stronger the interdisciplinary characteristics. Figure 6 is the result of analyzing the balance index of Big Data research by using WC19.exe. Professor Loet Leydesdorff used the Gini index to measure the balance of the interdisciplinary research. The value range of Gini index is 0-1. The closer the result is to 1, the more uniform the distribution of disciplines. Therefore, according to Figure 6, the Gini index of Big Data was 0.96 from 2007 to 2021, and it became 0.86 from 2019 to 2021, which shows that the uniformity of discipline distribution in the Big Data field is on the rise. With the deepening of research in the Big Data field, the participation of various disciplines and the uniformity of disciplines are rising. During 2007-2012, the research on Big Data might be limited to Computer Science, with few

research results from other disciplines and a large gap between disciplines. With the deepening of interdisciplinary research, the variety of disciplines is increasing, the influence of main disciplines is constantly improving, and the cooperation with other disciplines is also increasing, which shows that the interdisciplinary research in the field of Big Data is continuously becoming more balanced. On the whole, however, the Gini index is very close to 1, which shows that the distribution in Big Data is very uneven, with the main disciplines still dominating and the interdisciplinary with other disciplines only serving as auxiliary research.

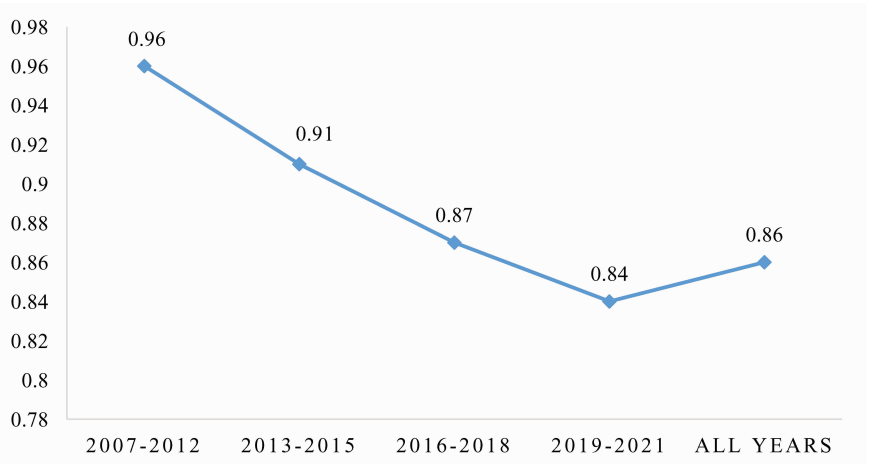

Figure 6 Gini index

3.2.3 Disparity

The disparity is the degree to which the categories of the elements are different. The disparity index measures the situation in which researchers cite the disciplines quite different from their own disciplines when conducting research. The discipline of big data research is computer science. Researchers cite the more disciplines that are very different from computer science, the more pronounced the interdisciplinary characteristics are. The intersection of disciplines is reflected by measuring the degree of overlap in the scope of research involved in the discipline over a continuous period. The stronger the dissimilarity, the more pronounced the interdisciplinary characteristics. Table 4 is the result of analyzing the disparity in the field of Big Data using WC19.exe. According to table 4, it can be concluded that the disparity in Big Data research is gradually increasing over time, which was 0.79 in 2007-2012 and 0.90 in 2019-2021. This shows that with the deepening of research in Big Data, researchers tend to cite the disciplines which are more different from Big Data fields for cross-disciplinary research.

Table 4 Disparity index over time

Year	Disparity index
2007–2012	0.79
2013–2015	0.88
2016–2018	0.90
2019–2021	0.90
All years	0.90

3.3 Network analysis of interdisciplinary collaboration in Big Data research

3.3.1 Overall network characteristics

Table 5 shows the annual basic information of the co-discipline network in Big Data research from 2012 to 2021 (excluding isolated nodes). The "overall" refers to the total value from 2007 to 2021. Since the data before 2012 are too small to be used for network analysis, only the data after 2012 are considered here. Before 2017, the nodes and lines of the co-discipline network are constantly increasing, which shows that the scale of the network is expanding every year. The scale of the network gradually stabilizes from 2017 to 2021, which shows that the co-discipline network has taken shape. The average degree of nodes in the network has been kept at around 5, with little change.

Table 5 Descriptive statistics of interdisciplinary collaboration networks (2012-2021)

Year	Number of nodes	Number of lines	Average degree
2012	9	19	1.7778
2013	33	79	2.6667
2014	51	162	4.2745
2015	62	203	4.4839
2016	72	279	5.6944
2017	86	328	5.5814
2018	87	341	5.7931
2019	86	311	5.186
2020	89	339	5.573
2021	86	310	5.1628
overall	132	732	9.0606

Figure 7 shows the evolution of network indicators of the co-discipline network in Big Data research from 2012 to 2021, and the "overall" refers to the total value from 2007 to 2021. Network density is the ratio of the number of edges existing in the network to the upper limit of the number of edges that can be accommodated, which is used to describe the closeness of interconnection between nodes in the network (Zhu & Li, 2008). The larger the value, the closer the connection. The maximum density that can be found in the existing network is 0.5. The density of the co-discipline network in Big Data research is just more than 0.05, which indicates that the density of the co-discipline network in Big Data research is low. The network clustering coefficient is used to describe the probability that two adjacent nodes of a certain node are also adjacent to each other (Yan & Ding, 2010), which is often used to indicate the degree of node aggregation. The higher the clustering coefficient, the more likely two nodes will be divided into a cluster. The clustering coefficient of the co-discipline network in Big Data research has been kept at around 0.3, which means that the nodes in the network are likely to be divided into clusters. Louvain Method is adopted in the community division of Big Data research. It can be seen that the number of community divisions fluctuates significantly from 2012 to 2021, but all of them are around 8. Among them, in the community division supported by all data from 2007 to 2021, the co-discipline network of Big Data research is divided into 5 communities.

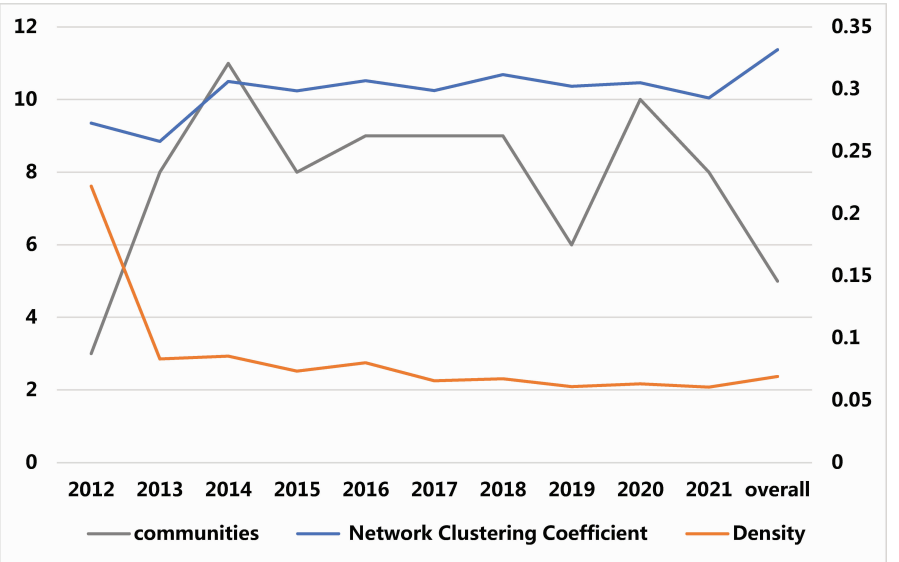

Figure 7 The evolution of network indicators of the largest component of interdisciplinary in Big Data research networks during 2012-2021 and all years : network indicators (right axis) and the number of communities (left axis)

3.3.2 Network characteristics of individual disciplines

Figures 8, 9, and 10 are the results obtained by analyzing the co-discipline network in Big Data research. As Hu (2017) has already explored the data before 2016 in his study, to ensure the timeliness and novelty of the study, we mainly calculate the degree centrality, betweenness centrality, and closeness centrality of the co-discipline network from 2016 to 2021. We list the top five disciplines and corresponding indicators (to show more clearly, Social Sciences-Other Topics and Science & Technology-Other Topics are replaced with Social Sciences and Science & Technology). Degree centrality, betweenness centrality, and closeness centrality are indicators to measure the position of nodes in the network, but they follow different standards. Degree centrality holds that if a node is connected with many nodes, the node is in a relatively central position in the network; Betweenness centrality holds that the more times a node appears on the shortest path of any two nodes, the more important it is. If the betweenness centrality of a node is high, it plays a key bridge role in the network. Closeness centrality believes that the less a node depends on other nodes when transmitting information, the more important it is. The correctness of these indicators has been verified in many studies (e.g., Sheng & Tang, 2022; Lv & Zhou, 2021).

By analyzing the degree centrality of each discipline in the co-discipline network in Big Data research from 2007 to 2021, the top five disciplines are Computer Science, Engineering, Business & Economics, Environmental Sciences & Ecology, and Social Sciences. From Figure 8, it can be seen that the degree centrality of Computer Science and Engineering has been far higher than that of other disciplines in the past six years, which shows that the collaboration between disciplines in Big Data research has always been dominated by these two disciplines. Business & Economics, Science & Technology, and Social Sciences also frequently appeared on the list. Education & Educational Research ranked in the top five for two consecutive years from 2017 to 2018, and environmental sciences & ecology ranked in the top five for three consecutive years from 2019 to 2021. Generally speaking, from 2016 to 2021, the

discipline ranking in Big Data research with degree centrality as the indicator changed little every year. The popular disciplines mainly include Computer Science, Engineering, Business & Economics, Science & Technology, and Social Sciences. The emerging hot discipline changed from Education & Educational Research to Environmental Sciences & Ecology. Compared with pre-2016, Automation and Control Systems are significantly less important in degree centrality, and Environmental Sciences & Ecology have become an important part of Big Data research. This also shows that scholars have been paying more and more attention to ecological and environmental issues in recent years.

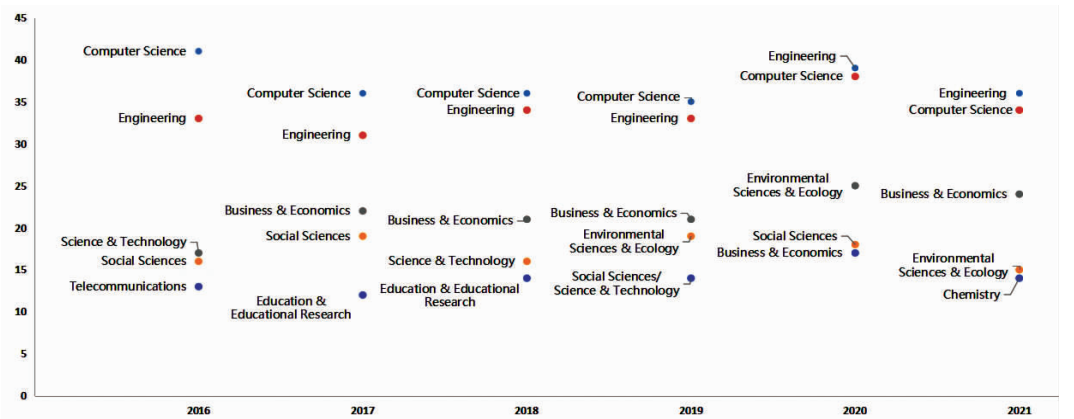

Figure 8 The degree centrality of top five disciplines (2016-2021)

Taking betweenness centrality as the analysis index, the top five disciplines in Big Data research from 2007 to 2021 are Engineering, Computer Science, Public, Environmental & Occupational Health, Environmental Sciences & Ecology, Neurosciences & Neurology. As can be seen from Figure 9, the betweenness centrality of Engineering and Computer Science has always been in the top two. It shows that the two disciplines play a vital and stable "bridge" role in the co-discipline network. Compared with degree centrality, the ranking of disciplines in Big Data taking betweenness centrality as the indicator from 2016 to 2021, except for the two disciplines of Engineering and Computer Science, changes greatly every year. The top ten disciplines with high betweenness centrality are more diverse but less focused.

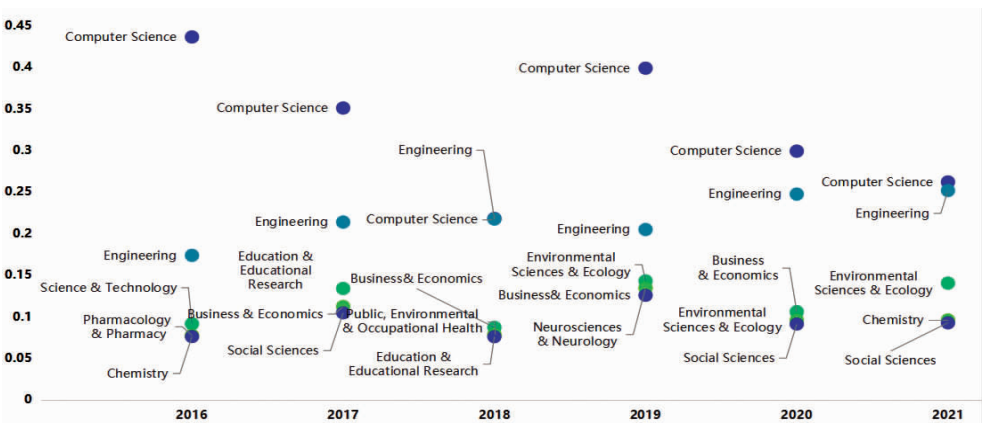

Figure 9 The betweenness centrality of top five disciplines (2016-2021)

By analyzing the closeness centrality of disciplines in Big Data research from 2007 to 2021, The top five disciplines are Computer Science, Engineering, Environmental Sciences & Ecology, Science & Technology, and Business & Economics. By observing Figure 10, we can conclude that the closeness centrality of Computer Science, Engineering, Business & Economics is always at the forefront, which shows that these three disciplines are particularly independent in the network. The closeness centrality of Environmental Sciences & Ecology also ranks in the top five from 2019 to 2021. Compared with pre-2016, the independence of Automation & Control Systems and Mathematics has declined, while that of Environmental Sciences & Ecology and Science & Technology has improved.

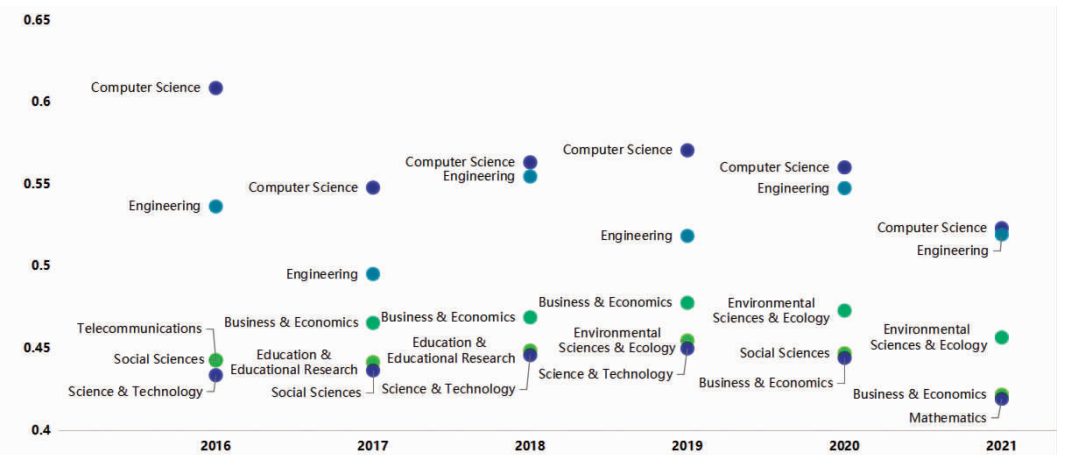

Figure 10 The closeness centrality of the top five disciplines (2016-2021)

Considering the above indicators comprehensively, Computer Science and Engineering are the two most important disciplines in Big Data research. Business & Economics, Science & Technology, Social Sciences, and Environmental Sciences & Ecology are also in a critical position. Environmental Sciences & Ecology, Public, Environmental & Occupational Health, and Neurosciences & Neurology are emerging disciplines in Big Data research in recent years. As can be seen from the figures, the differences in degree centrality, betweenness centrality, and closeness centrality among disciplines are gradually decreasing. It shows that the collaboration mode among disciplines in Big Data research has changed from the leading role of several important disciplines to more extensive and universal collaboration.

3.3.3 Interdisciplinary collaboration communities

In this study, the Louvain community discovery algorithm is used to divide the co-discipline network of Big Data into communities. After the community division, the nodes within the same community have a relatively strong correlation, while the correlation between nodes in different communities is weak. Under this division, disciplines with close collaboration will be divided into one community, so finally, discipline groups based on collaboration will be formed. Figure 7 shows the number of communities divided by the Louvain algorithm from 2012 to 2021. The number of communities is basically around 8, indicating that the collaboration among disciplines in Big Data research has become mature and stable. As Hu (2017) has analyzed the data before 2016 in his research, Figures 11 to 16 only show the community division of the co-discipline network in Big Data research every year from 2016 to 2021. Figure 17 shows the community division of the overall co-discipline network in Big

Data research from 2007 to 2021. In these diagrams, each node represents a discipline, and the color of the node represents the community to which the discipline belongs. The size of the node is determined by the number of occurrences of the discipline, and the thickness of the connection between the nodes is determined by the number of co-occurrences of the two disciplines. We mainly analyze the five crucial communities in the co-discipline network of Big Data research.

There are three central communities in the co-discipline network, namely, community 1, represented by Computer Science and Engineering; community 2, represented by Business & Economics; and community 3, represented by Science & Technology. These community representatives are the most important disciplines in the whole co-discipline network and also the skeleton of the entire network. They lead their respective communities, while other disciplines are in a relatively secondary position.

Community 1 is the core community in the whole co-discipline network. Its number of occurrences and co-occurrences is huge, and it is also a community with the most stable composition of discipline members. The disciplines belonging to the community mainly include Computer Science, Engineering, Telecommunications, Automation & Control Systems, and Operations Research & Management Science. As we can see, at the beginning of 2016, the number of disciplines' occurrences in community 1 was polarized, and Computer Science was the absolute leader. However, as time goes by, the number of occurrences of Engineering and Telecommunications is increasing, and the gap with Computer Science is narrowing. By 2021, we can find that Computer Science, Engineering, and Telecommunications within community 1 are in a three-pronged trend. Therefore, we can conclude that the two disciplines, Engineering and Telecommunications, developed rapidly from 2016 to 2021 and gradually became the core disciplines in Big Data research.

The importance and stability of community 2 are second only to that of community 1. Its disciplines mainly include Business & Economics, Social Sciences-Other Topics, Information Science & Library Science, and Education & Educational Research. Education & Educational Research belongs to community 1 in 2020 and 2021. It shows that Education & Educational Research has a good collaborative relationship with the members of community 1 and community 2 and has collaborated more closely with community 1 in recent two years. We can also find that Business & Economics is always in the leading position in the number of occurrences in Community 2. Still, from 2016 to 2021, the number of occurrences of Social Sciences and Information Science & Library Science has increased greatly. They have become the two most important disciplines in Community 2 besides Business & Economics.

The representative disciplines of community 3 include Science & Technology, Environmental Sciences & Ecology, and Public, Environmental & Occupational Health. From 2016 to 2021, the scale of community 3 continues to grow. By around 2020, the scale of Community 3 has gradually approached that of Community 2, and the stability of the community has steadily increased. At the beginning of 2016, the leading disciplines within Community 3 were Science & Technology, Environmental Sciences & Ecology, and Public, Environmental & Occupying Health. Since 2019, the number of occurrences of Environmental Sciences & Ecology has surpassed that of Science & Technology, becoming the most important discipline in Community 3.

Community 4 and Community 5 can't be regarded as the supporting communities of the co-discipline network in Big Data research, but their number of occurrences is increasing

from 2016 to 2021. In addition, the cooperation among disciplines within the community 4 and 5 has become mature and stable. Community 4 mainly comprises Materials Science, Instruments & Instrumentation, Chemistry, and Physics. We can find that the members of community 4 are always changing, for example, Materials Science belonged to community 1 in 2016, and Instruments & Instrumentation belonged to community 1 in 2017 and 2021. However, over time, the nodes within community 4 have changed from small and changeable at the beginning to large and few. The leading members of community 5 are Health Care Sciences & Services, Psychology, Neurosciences & Neurology, and Medical Informatics, which didn't become stable until 2020. Before that, disciplines in community 5 were attached to other communities. In addition to the above five communities, disciplines such as Remote Sensing, Geography, and Physical Geography also formed a community in 2021, but the collaboration among these disciplines is unstable.

Disciplines close to one another are more likely to be in the same community because they cost less to collaborate. Therefore, we find that each community represents one major research direction in Big Data research. Collaborations in community 1 are related to techniques in Big Data, which are always the center of the whole co-discipline network, connecting with and supporting other independent communities (Hu, 2017). Community 2 and 3 are also framework of the whole network, and collaborations in them are usually about applications of Big Data. We can find that the scale of communities is unbalanced, while the scale and stability of communities are improving yearly. The gap between communities is also shrinking. Social Science in Big Data was still in its infancy in 2016 (Hu & Zhang, 2018), but now it has become mature.

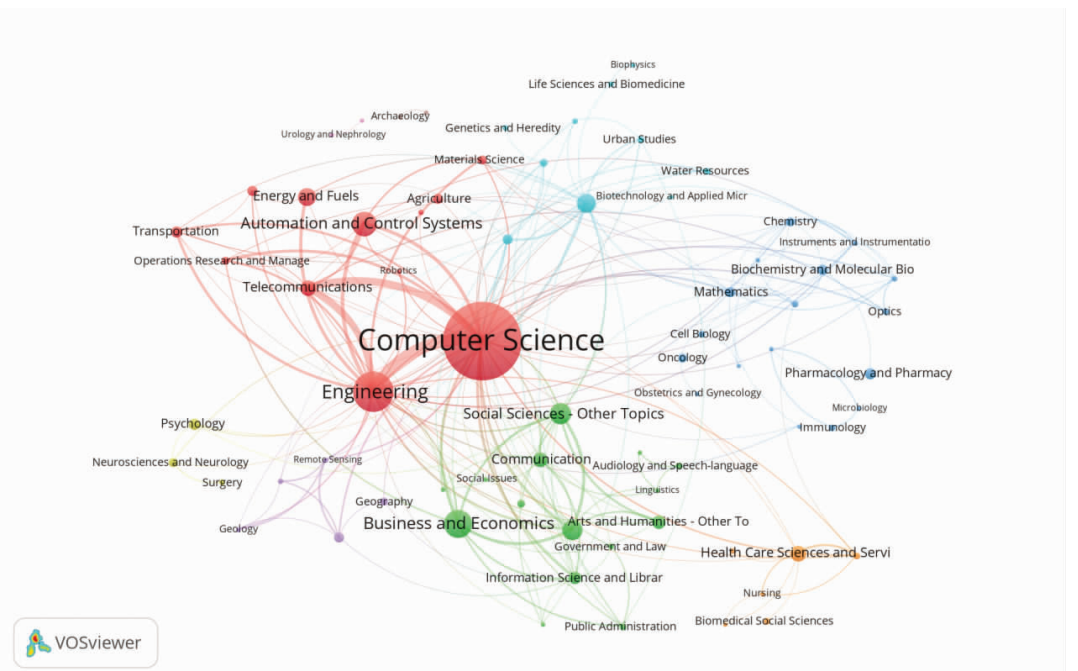

Figure 11 Interdisciplinary collaboration communities (2016)

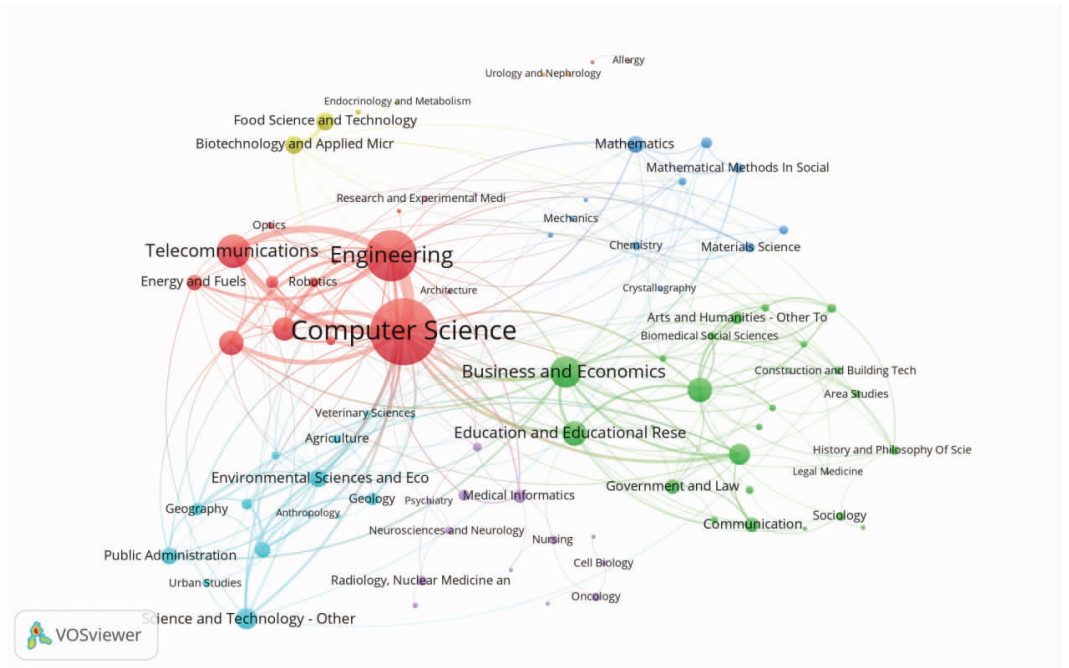

Figure 12 Interdisciplinary collaboration communities (2017)

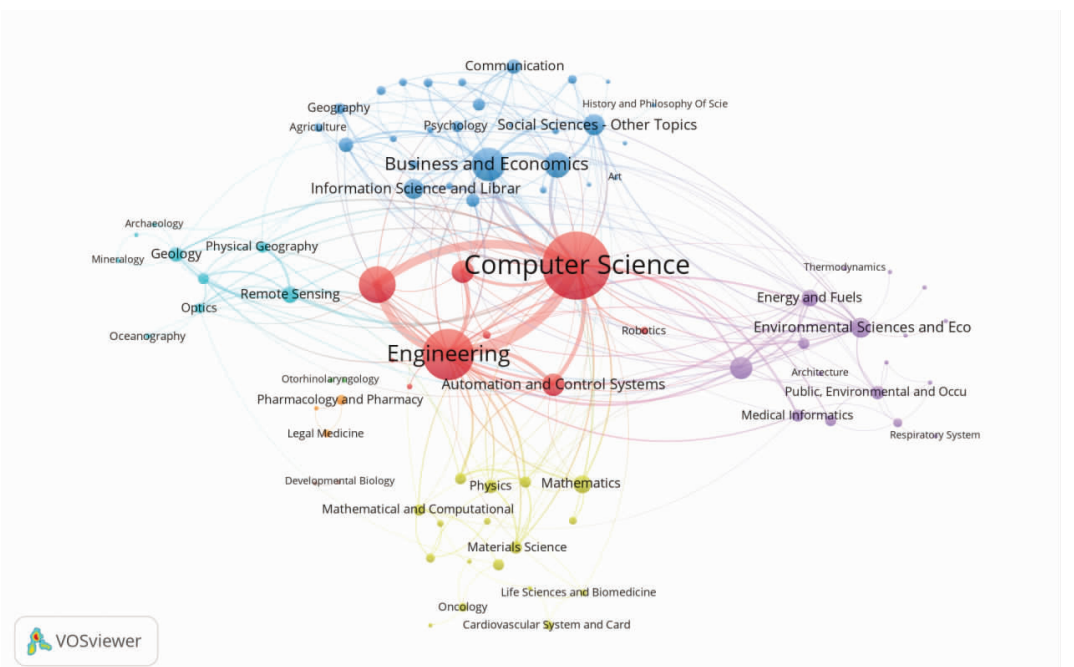

Figure 13 Interdisciplinary collaboration communities (2018)

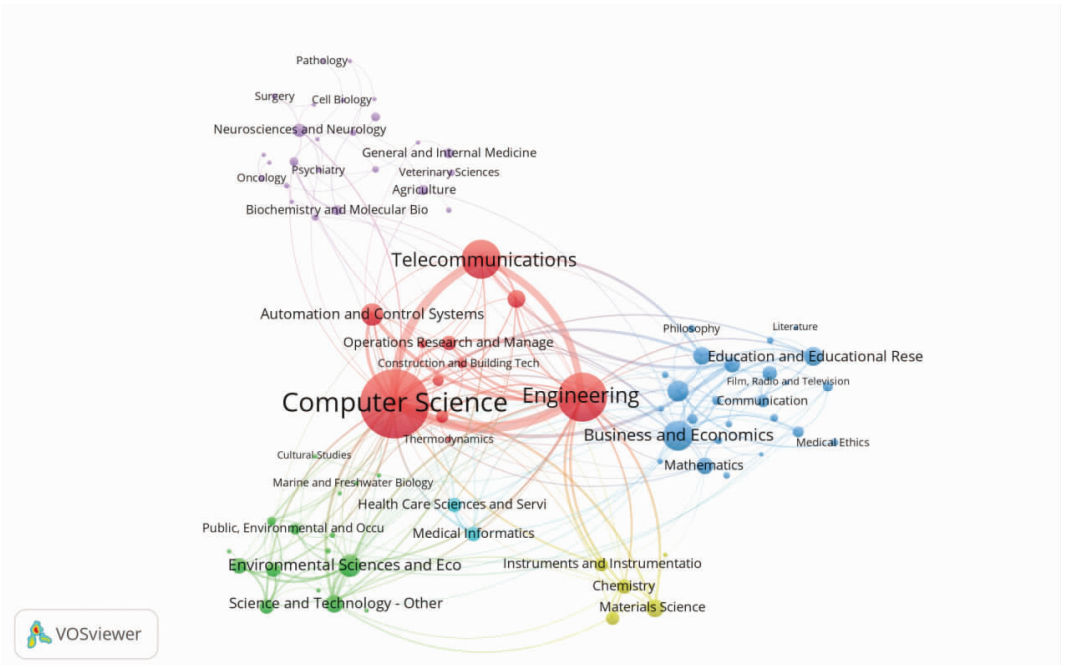

Figure 14 Interdisciplinary collaboration communities (2019)

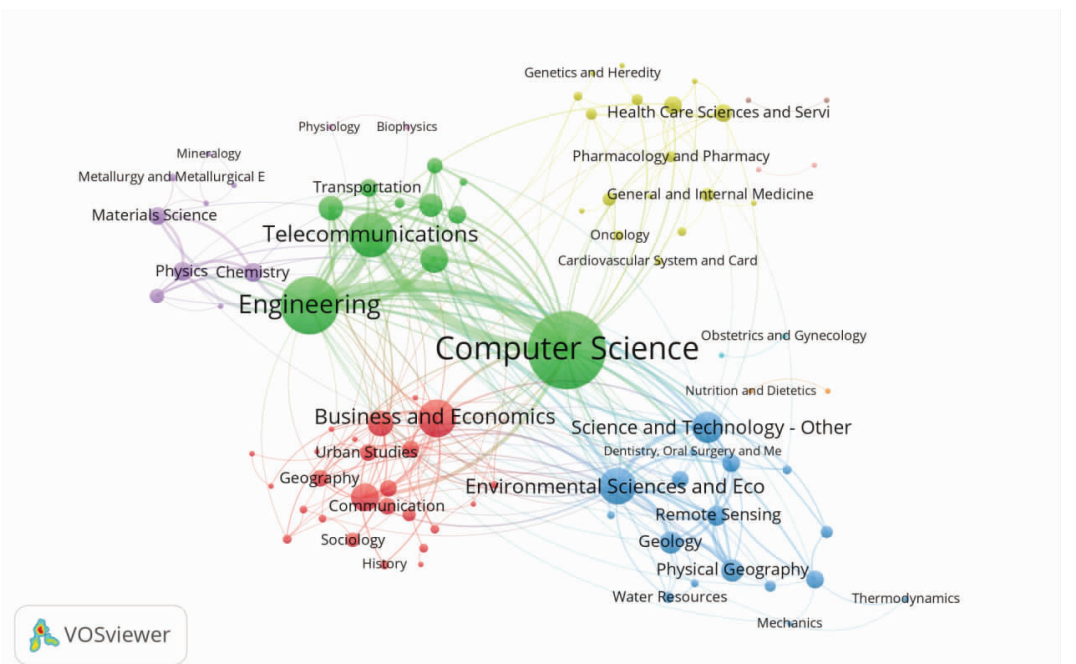

Figure 15 Interdisciplinary collaboration communities (2020)

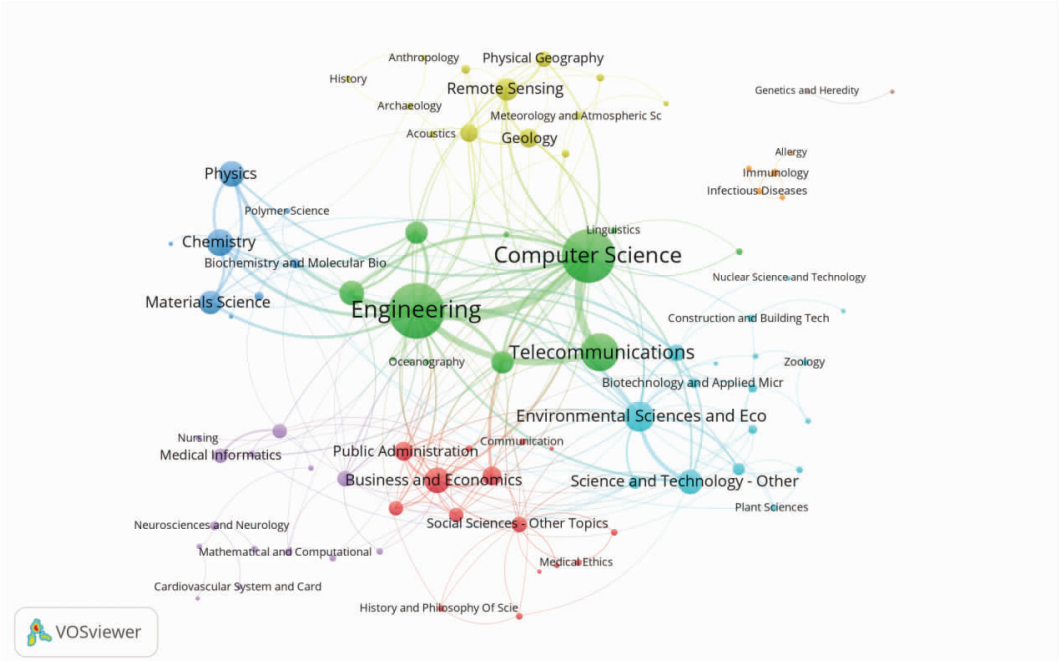

Figure 16 Interdisciplinary collaboration communities (2021)

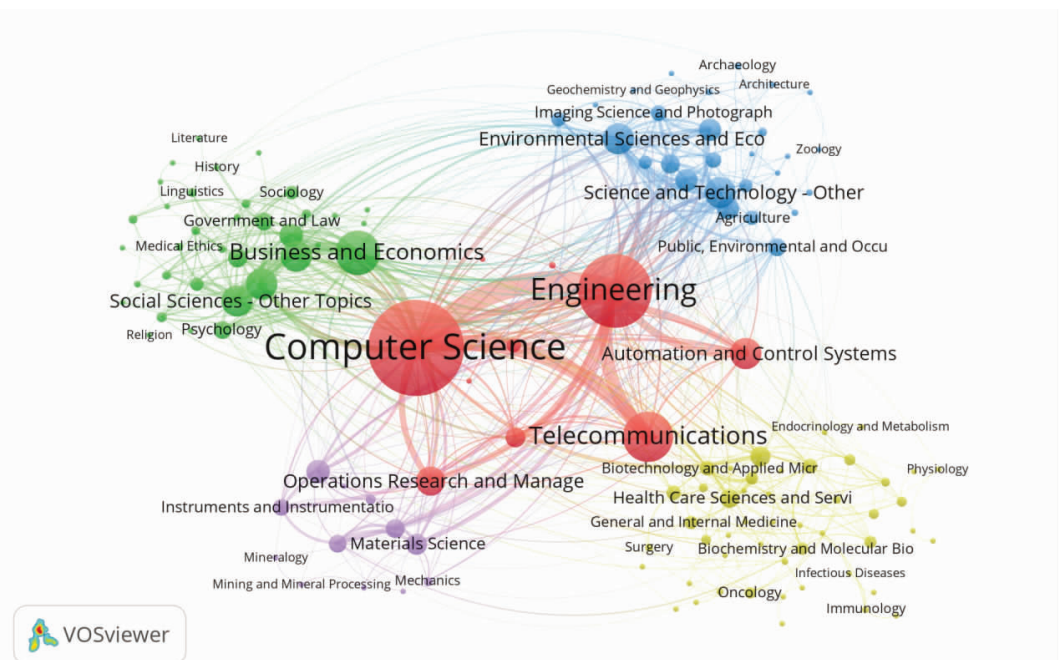

Figure 17 Interdisciplinary collaboration communities (2007- 2021)

4 Conclusions

4.1 Summary

This study explores the interdisciplinarity of Big Data research. Firstly, we summarize Big Data research by traditional bibliometrics. Secondly, we analyze the interdisciplinarity of Big Data research from three-dimensional indicators of discipline variety, balance, and disparity. Finally, we study the interdisciplinary collaboration in Big Data research from the social network analysis.

According to the number of published papers from 2007 to 2021, we can know that Big Data research reached its peak in 2018 and then began to decline slowly. Nevertheless, Big Data research is still a critical topic at present. In recent years, the micro-topics in Big Data research mainly focus on the Internet of Things, Syndromic Surveillance, Knowledge Management, Industry 4.0, and so on. From 2007 to 2021, China, the United States, and India were the top 3 countries in terms of the number of papers, and the collaboration between China and the United States was the closest. Big Data research involves a total of 141 disciplines. Based on the number of papers published, Computer Science, Engineering, Telecommunications, and Business & Economics are the four most important disciplines, with more than 1,000 published papers in each discipline. The four disciplines account for 64% of the total occurrences of all disciplines, which shows that the distribution of disciplines in Big Data research is highly uneven.

From the perspective of variety, balance, and disparity, it can be concluded that the interdisciplinary characteristics of the Big Data field are becoming more and more apparent. The variety index is constantly improving, and more and more disciplines are getting involved in Big Data research. Then the discipline balance is continually rising, the participation of each discipline is gradually deepening, and discipline uniformity is on the rise. Regarding discipline disparity, researchers in the Big Data field are more inclined to cite the disciplines that are quite different from their disciplines. Generally speaking, the interdisciplinary of the Big Data field is increasing.

By analyzing the co-discipline network of Big Data, we can know that before 2017, the nodes and lines of the discipline network are constantly increasing, which shows that the scale of the network is expanding. At the same time, the scale of the network tends to be stable gradually from 2017 to 2021, which shows that the co-discipline network has taken shape. Then, from the three indicators of degree centrality, betweenness centrality, and closeness centrality, we find the critical nodes in the co-discipline network of Big Data and finally draw a conclusion: the two most important disciplines in Big Data research are Computer Science and Engineering. Disciplines like Business & Economics, Science & Technology, Social Sciences, and Environmental Sciences & Ecology are also essential. Among them, Environmental Sciences & Ecology, Public, Environmental & Occupational Health, and Neurosciences & Neurology are emerging disciplines of Big Data in recent years. Finally, focusing on the five communities in the network, Communities 1, 2, and 3 are the most important communities in the network. The collaboration among disciplines within them is relatively stable, while communities 4 and 5 are slowly forming, developing, and stabilizing in recent years. Besides, other communities are still in their infancy.

Generally speaking, the research in Big Data has entered a relatively stable stage, and the number of disciplines involved has gradually become stable. By analyzing the whole

co-discipline network, we can know that more and more disciplines are involved in Big Data research, the variety of disciplines is constantly improving, and the balance and disparity between disciplines are constantly rising. Although the discipline balance in Big Data research has been continuously improved, the fact that Big Data research is still led by several major disciplines will not change in the short term. Computer Science and Engineering are still the main undisputed contributors. With the deepening of research, the collaborative communities in the co-discipline network are slowly expanding and maturing. At the same time, with the addition of new disciplines, there will be more collaboration among disciplines and more new collaborative communities.

4.2 Implications

Firstly, Big Data is an emerging field. This study provides a clear and comprehensive understanding of the collaboration pattern and distribution characteristics of interdisciplinarity in Big Data research. Secondly, this study points out five communities of interdisciplinary collaboration in the Big Data field. For example, Community 1 is represented by Computer Science and Engineering. These can provide valuable references for people to understand the field of big data in depth. Furthermore, this study introduces not only the characteristics of discipline variety, balance, and disparity but also the perspective of a co-discipline network. The methods and framework can serve as a reference for interdisciplinarity studies in the future.

4.3 Limitations and prospects

This study analyzes the interdisciplinarity of Big Data research from three aspects, the results are intuitive, and the conclusions are meaningful. However, this study only obtained the characteristic of the interdisciplinarity of Big Data research by using descriptive statistics and is weak in theoretical contribution. It could not precisely measure the interdisciplinarity from the semantic perspective, nor did it analyze the interdisciplinary knowledge flow of Big Data. In the future, these aspects deserve further exploration: first, pay more attention to the characteristics of the dynamic development of interdisciplinary; second, introduce new methods such as artificial intelligence and text mining techniques; third, explore the knowledge flow and influence factor in Big Data research.

Acknowledgements

This article is an outcome of the major project-The Research on the Construction of A Cloud Platform for Science and Education Evaluation and Intelligent Service Based on Big Data (No. 19ZDA348), supported by National Social Science Foundation of China

References

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment analysis in tourism: Capitalizing on big data. *Journal of Travel Research, 58* (2), 175–191.

Alhussain, T. (2018). Medical big data analysis using big data tools and methods. *Journal of Medical Imaging and Health Informatics, 8* (4), 793–795.

An, P., Xiao, X., Guo, H., Yan, D., & Li, J. (2021). Big earth data research topic evolution and influence analysis. *Bulletin of Chinese Academy of Science, 36* (8), 973–988. doi:10.16418/j.issn.1000–3045.20210807001.

Ashabi, A., Sahibuddin, S. B., & Haghighi, M. S. (2020, April). Big data: Current challenges and future scope. In *2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)* (pp. 131–134). IEEE.

Borjigin, C., Zhang, C., Sun, Z., & Yi, N. (2021). Theoretical data science: Bridging the gap between domain–general and domain–specific studies. *Data Science and Informetrics, 1* (1), 1–28.

Borjigin, C., & Zhang, C. (2022). Data science: Trends, perspectives, and prospects. *Data Science and Informetrics, 2* (3), 1–21.

Caesarius, L. M., & Hohenthal, J. (2018). Searching for big data: How incumbents explore a possible adoption of big data technologies. *Scandinavian Journal of Management, 34* (2), 129–140.

Chang, V. (2021). An ethical framework for big data and smart cities. *Technological Forecasting and Social Change, 165*, 120559.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications, 19* (2), 171–209.

Chen, S., Qiu, J., & Yu, B. (2021). China´s research contribution in big data. *Data Science and Informetrics, 1* (3), 1–13.

Tech America Foundation´s Federal Big Data Commission. (2012). Demystifying big data: A practical guide to transforming the business of Government.

Furht, B., & Villanustre, F. (2016). Introduction to big data. In *Big data technologies and applications* (pp. 3–11). Springer, Cham.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management, 35* (2), 137–144.

Ghani, N. A., Hamid, S., Hashem, I. A. T., & Ahmed, E. (2019). Social media big data analytics: A survey. *Computers in Human Behavior, 101*, 417–428.

Gupta, N. K., & Rohil, M. K. (2020). Big data security challenges and preventive solutions. In N. Sharma, A. Chakrabarti & V. E. Balas (Eds.), *Data Management, Analytics and Innovation* (pp. 285–299). Springer, Singapore.

Hu, J., & Zhang, Y. (2017). Discovering the interdisciplinary nature of Big Data research through social network analysis and visualization. *Scientometrics, 112* (1), 91–109.

Hu, J., & Zhang, Y. (2018). Measuring the interdisciplinarity of Big Data research: A longitudinal study. *Online Information Review, 42* (5) , 681–696. doi:10.1108/OIR–12–2016–0361

Hulsen, T., Jamuar, S. S., Moody, A. R., Karnes, J. H., Varga, O., Hedensted, S., ... & McKinney, E. F. (2019). From big data to precision medicine. *Frontiers in medicine, 6*, Article 34.

Iqbal, R., Doctor, F., More, B., Mahmud, S., & Yousuf, U. (2020). Big data analytics: Computational intelligence techniques and application areas. *Technological Forecasting and Social Change, 153*, 119253.

Jang, W., Kwon, H., Park, Y., & Lee, H. (2018). Predicting the degree of interdisciplinarity in academic fields: the case of nanotechnology. *Scientometrics, 116* (1), 231–254.

Jia, H., & Jia, C. (2019, June). Construction and application of data standard in Big Data environment. In *Proceedings of the 2019 International Conference on Big Data Engineering* (pp. 121–124).

Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University–Computer and Information Sciences, 30* (4), 431–448.

Kankanhalli, A., Hahn, J., Tan, S., & Gao, G. (2016). Big data and analytics in healthcare: Introduction to the special section. *Information Systems Frontiers, 18* (2), 233–235.

Liu, Z.(1993). Interdisciplinary research in the new era of interdisciplinary science. *Studies in Science of Science, 11* (2), 9–16. doi:10.16192/j.cnki.1003–2053.1993.02.003.

Lv, X., & Wang, H. (2019). A comparative study of the interdisplinarity of big data research in China and the USA. *Science Research Management, 40* (4), 1–13. doi:10.19571/j.cnki.1000–2995.2019.04.001.

Lv, X., Cai, X., & Zhou, P. (2021). Visual analysis on global Big Data collaboration network under the background of science and technology globalization. *Science and Technology Management Research, 41*(16), 26–36.

Qi, C. C. (2020). Big data management in the mining industry. *International Journal of Minerals, Metallurgy and Materials, 27* (2), 131–139.

Qiu, J., & Shen, C. (2021). Analysis of hot topics in domestic Big Data research based on LDA model. *Journal of Modern Information, 41* (9), 22–31.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies

in bionanoscience. *Scientometrics, 82* (2), 263–287.

Sheng, X., & Tang, J. (2022). Analysis of user interaction characteristics and rules in virtual academic community from the perspective of social network. *Journal of Modern Information, 42* (1), 64–71.

Stoianov, N., Urueña, M., Niemiec, M., Machnik, P., & Maestro, G. (2015). Integrated security infrastructures for law enforcement agencies. *Multimedia Tools and Applications, 74* (12), 4453–4468.

Sun, Z., & Wang, P. P. (2017). A mathematical foundation of big data. *New Mathematics and Natural Computation, 13* (2), 83–99.

Taleb, I., Serhani, M. A., & Dssouli, R. (2018, November). Big data quality assessment model for unstructured data. In *2018 International Conference on Innovations in Information Technology* (IIT) (pp. 69–74). IEEE.

Taşkln, Z., & Aydinoglu, A. U. (2015). Collaborative interdisciplinary astrobiology research: A bibliometric study of the NASA Astrobiology Institute. *Scientometrics, 103* (3), 1003–1022.

Wang, X. (2017). Comparative analysis of domestic and international Big Data research based on bibliometrics. *Library Work in Colleges and Universities, 37* (4),49–54.

Xiong, W., & Fu, H. (2021). Topics and its evolution of interdisciplinary research based on topic mining model. *Information Science, 39* (11), 117–126. doi:10.13833/j.issn.1007–7634.2021.11.016.

Yan, E., Ding, Y., & Zhu, Q. (2010). Mapping library and information science in China: A coauthorship network analysis. *Scientometrics, 83* (1), 115–131.

Zhang, J. (2016). The interdisciplinary research of big data and wireless channel: A cluster–nuclei based channel model. *China communications, 13* (2), 14–26.

Zhang, L., Sun, B., & Huang, Y. (2018). Interdisciplinarity measurement based on interdisciplinary collaborations: A case study on highly cited researchers of ESI social sciences. *Journal of the China Society for Scientific and Technical Information, 37* (3), 231–242.

Zhu, Q., & Li, L. (2008). Social network analysis and method & its application in information science. *Information studies: Theory & Application, 2008* (02), 179–183+174. doi:10.16353/j.cnki.1000–7490.2008.02.021