sustainability

MDPI

Article

An Empirical Study on Visualizing the Intellectual Structure and Hotspots of Big Data Research from a Sustainable Perspective

Feng Hu [1,2,3], Wei Liu [4,*], Sang-Bing Tsai [5,6,*], Junbin Gao [3], Ning Bin [7] and Quan Chen [5,*]

1 School of Management, Guangdong University of Technology, Guangzhou 510520, China; fenghu@gdut.edu.cn
2 Institute of Big Data Strategic Research, Guangdong University of Technology, Guangzhou 510006, China
3 Discipline of Business Analytics, The University of Sydney Business School, The University of Sydney, Camperdown, NSW 2006, Australia; junbin.gao@sydney.edu.au
4 School of Business, Wuyi University, Nanping 354300, China
5 Zhongshan Institute, University of Electronic Science and Technology of China, Guangzhou 528400, China
6 Economics and Management College, Civil Aviation University of China, Tianjin 300300, China
7 School of Management, Guangdong University of Technology, Guangzhou 510520, China; bbb8087@gdut.edu.cn
* Correspondence: wei.liu2@sydney.edu.au (W.L.); sangbing@hotmail.com (S.-B.T.); zschenquan@gmail.com (Q.C.)

Received: 6 February 2018; Accepted: 26 February 2018; Published: 1 March 2018

Abstract: Big data has been extensively applied to many fields and wanted for sustainable development. However, increasingly growing publications and the dynamic nature of research fronts pose challenges to understand the current research situation and sustainable development directions of big data. In this paper, we visually conducted a bibliometric study of big data literatures from the Web of Science (WoS) between 2002 and 2016, involving 4927 effective journal articles in 1729 journals contributed by 16,404 authors from 4137 institutions. The bibliometric results reveal the current annual publications distribution, journals distribution and co-citation network, institutions distribution and collaboration network, authors distribution, collaboration network and co-citation network, and research hotspots. The results can help researchers worldwide to understand the panorama of current big data research, to find the potential research gaps, and to focus on the future sustainable development directions.

Keywords: big data; visualizing; intellectual structure; big data environment; co-citation network; collaboration network; sustainability

1. Introduction

With the growing popularity of mobile terminals, Internet of Things (IoT), social networks, cloud computing and mobile commerce, myriad data are generated, and the era of big data is coming. The advent of big data has promoted the revolution of data-driven thinking and decision making. Governments, industry and academia have paid great attention to big data strategy, technologies and applications. More and more people worldwide have made tremendous efforts in large-scale heterogeneous data collection, organization, storage, analysis, mining and applications under the big data environment. Big data has become a hot topic of discussion. For example, Nature and Science published special issues "Big Data" in 2008 and "Dealing with Data" in 2011 respectively. In May 2011, the McKinsey global institute (MGI) released the research report "Big Data: The Next Frontier for Innovation, Competition, and Productivity" [1]. In March 2012, U.S. President Office of Science and Technology Policy declared in public that the United States government would invest

$200 million to launch "The Big Data Research and Development Initiative" [2]. At the same time, big data had been extensively applied into many fields, such as IoT, social networks, health care, intellisense, environment and sustainable development, and so on [3]. For example, according to the UN Sustainable Development Goals (SDG) [4], big data such as from satellite imagery and sensor networks make environment and development indicators increasingly measurable. Worldwide research institutions and scholars had devoted themselves to big data science research and wanted big data for sustainable development. However, more and more research outcomes have been emerging and growing rapidly [5–7]. Moreover, the dynamic nature of a research front poses challenges for scientists, research policy makers, and many others to keep up with the rapid advances of the state of the art in science [8]. It is still difficult for scholars to understand the current research situation and sustainable development trends of big data. Therefore, how to identify intellectual structure, to detect emerging trends and sudden changes of big data research is increasingly essential.

In recent years, with the rapidly increasing publications related to big data, some scholars have begun to aggregate relevant existing literatures, performed the bibliometric analysis, and visualized the intellectual structure, hotspots and evolution paths to provide knowledge support for other researchers in different fields based on bibliometrics [9–11]. Bibliometrics comprehensively utilizes multi-disciplinary knowledge and methods, such as mathematics, statistics, philology, etc., to analyse the distribution regularities, the developments and research trends of a certain scientific field, and finally visualizes the research results. However, so far few quantitative depictions have been given of the intellectual structure and hotspots of big data research. A few existing surveys mainly focus on specific big data subfields and themes, such as big data and IoT applications on circular economy [10], social networks, health care [11], and supply chain [12]. However, these surveys were absent in the panorama of the big data field and were not conducted on the sustainability of big data research. It is still difficult for readers to deeply understand the current intellectual structure and sustainable development directions of big data research.

In this paper, we performed a bibliometric analysis distinct from the above existing surveys in several aspects. Firstly, this study retrieved all journal articles of big data between 2002 and 2016 in the WoS database, which include Science Citation Index Expanded (SCI-EXPANDED) journals, Social Sciences Citation Index (SSCI) journals, and Arts & Humanities Citation Index (A&HCI) journals. Secondly, this study did not simply describe the traditional concentrated distribution regularities. More importantly, visualization techniques and co-word analysis were used to demonstrate visually the intellectual structures, collaboration networks, and research hotspots of big data between 2002 and 2016 from the following perspectives: publications distribution, core journals, core institutions and collaboration network, core authors and collaboration network, as well as high-frequency keywords network. It provided a vivid overall picture of big data research. This study will help would-be big data researchers know the current research situation, research gaps, what journals they should follow, what authors they should focus, how to seek co-researchers, and work out the details in big data research activities. Moreover, it will also be helpful to improve and upgrade the sustainable research and development, applications, and policy making of big data at different levels in the future. Thirdly, this study went beyond traditional citation counts. Journal co-citation analysis (JCA) and author co-citation analysis (ACA), provided by CiteSpace, were used to detect some special pioneers and journals in the big data field from the following perspectives: the most co-cited frequency, intellectual turning points, and highest citation bursts. These pioneers and journals had contributed to the sustainable development of big data research from different perspectives.

The paper is organized as follows: In Section 2, we describe the methodology, including original data sources and research methods. In Section 3, we demonstrate the bibliometric analysis results, and visualize the intellectual structure and hotspots of big data research. In particular, we detect the distribution characteristics, intellectual turning points, strongest citation bursts, and research hotspots. In Section 4, we finally present the discussion and conclusions.

2. Methodology

2.1. Data Source

The first step consisted in collecting bibliographic data from robust and reliable data sources. Previous bibliographic data were extracted from different data sources. Some collected data from a single journal [13] or multiple journals [14]. Others did not discriminate the journals sources but regarded the citation databases [9,10,15]. Commonly used citation databases include the Web of Science (WoS), Scopus, Google Scholar (GS), and PubMed [15]. Each database has its own advantages and drawbacks. A certain database could be stronger in types, quantities, and countries of publications, while the others focus more on literature evaluation methods and indicators. For example, compared to WoS, Scopus significantly alters the relative ranking of those scholars who appear in the middle of the rankings and GS stands out in its coverage of conference proceedings as well as international, non-English language journals [16]. In this paper, we select the WoS as our data source. The WoS is an ideal single research destination to explore the citation universe across subjects and around the world, and provides everyone access to the most reliable, integrated, multidisciplinary research connected through linked content citation metrics from multiple sources within a single interface. Furthermore, the WoS adheres to a strict evaluation process, and only the most influential, relevant, and credible information is included.

In addition, with the growing emergence of social media, there are a variety of important ways for scientists to spread their academic ideas, such as monographs, conference proceedings, and personal blogs or web pages. Nevertheless, compared with the books, reports, and other equal ways, academic journals tend to be more direct, consistent and important channels for scientists to publish, spread, accumulate, comment on and assume the lead in a specific scientific research fields [17,18]. Furthermore, most key studies are usually published in core international journals [10]. We therefore target the journal articles in the Web of Science.

On 20 May 2017, the WoS database was searched using the following basic terms: topic = "big data", literature type = "article", and publication years were restricted to "2002–2016". We eventually obtained 4927 effective journal articles. The bibliographic records, including titles, authors, institutions, keywords, references, etc., were downloaded. These journal articles were distributed across 1729 journals, and contributed by 16,404 authors from 4137 institutions.

2.2. Research Methods

The methods of bibliometrics have been widely applied in quantitative analyses in many knowledge fields [8–10,12–15,17,19]. It comprehensively uses the professional knowledge and methods of mathematics, statistics, information science, philology, and other disciplines to analyze the distribution regularities, intellectual base, research front, and evolution paths. Commonly used bibliometrics methods include co-word analysis [20], document co-citation analysis (DCA) [11,21], author co-citation analysis (ACA) [22,23] and many other variations [15]. In addition, information visualization, raised by Robertson in 1989, focuses on interactive visual representations of abstract data to reinforce human cognition. Visualization techniques include visualizations of hierarchies or trees [24], graph or network structures [25], temporal structures [26], geospatial visualizations, and coordinated views of multiple types of visualizations [15].

With the development of information technologies, many representative software tools were exploited to facilitate the information visualization and science mapping of knowledge domains. Frequently used information visualization and science mapping software tools include some nonspecific science mapping software (e.g., Pajek, Gephi, or UCINET) and specific science mapping software tools, such as IN-SPIRE [27], VantagePoint [28], CiteSpace II [11,29,30], CoPalRed [31,32], Leydesdorff's Software [33], Bibexcel [34], Sci2 Tool [35] (Sci2 Team, 2009), VOSViewer [36], Network Workbench Tool [37], SciMAT [38], and so on. Each one presents different features, advantages, and drawbacks due to its own different analysis techniques and algorithms. As a result, there was no single

software tool effective and flexible enough to fulfill overall science mapping analysis [39]. Therefore, in this paper, we adopt using more than one software tool to perform deep science mapping analyses. For example, we use an ad hoc software tool SATI3.2 [40] to clean the data in the preprocessing stage, and apply UCINET6 and CiteSpace V to build networks and visualize scientific mapping. SATI3.2, developed by Qiyuan Liu at Zhejiang University (China), is also applied to field data extraction, item frequency statistics, co-occurrence matrix construction, and visual analysis based on NetDraw. It is freely downloadable at http://sati.liuqiyuan.com. UCINET6 for Windows, developed by Lin Freeman, Martin Everett and Steve Borgatti, is a software package for the analysis of social network data. It comes with the NetDraw network visualization tool and can be downloaded at https://sites.google.com/site/ucinetsoftware/home. CiteSpace V, developed by professor Chaomei Chen at Drexel University (USA), is used to focus on visual analysis and scientific mapping. It is a Java-based information visualization and scientific mapping software package and can be freely available at http://cluster.cis.drexel.edu/~cchen/citespace/. The main functions include co-word networks analysis and co-citation networks analyses of authors, documents, institutions and journals. More importantly, CiteSpace V facilitates the identification of the chronologic patterns of a specific knowledge domain, including research hotspots, intellectual turning points, and citation burst.

The aim of this article is to demonstrate visually the intellectual structure and hotspots in big data research from 2002 to 2016. Particularly, the distribution characteristics, intellectual turning points, and emerging trends are examined from the following perspectives: publications, journals, institutions and authors, as well as keywords analysis.

3. Results

3.1. Publications Distribution

To evaluate the outcomes of big data research between 2002 and 2016, we collected 4927 journal articles from WoS databases and tracked the annual publications distribution of big data research (shown in Figure 1). There were few journal articles on big data research before 2009. However, a growth spurt was generated from 2010 to 2016, when dozens, and eventually thousands, of journal articles emerged.

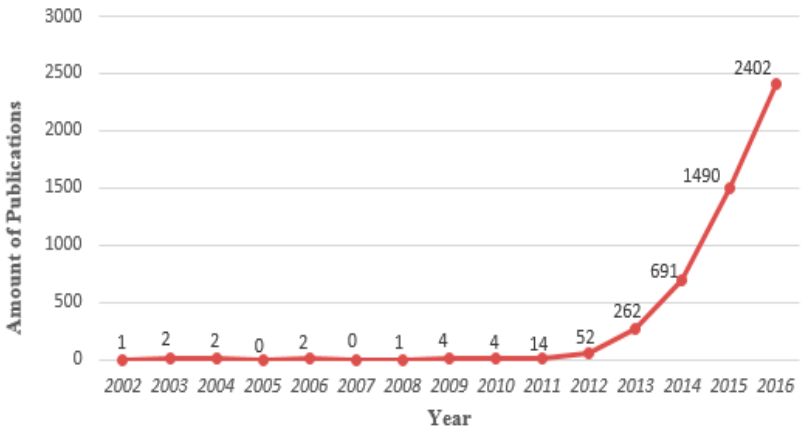

Figure 1. Annual publications distribution of big data research.

As shown in Figure 1, we roughly divided the development of big data research into two stages. Stage I (2002–2009) is an embryonic stage with few annual articles, which indicate that big data exploration just starts. The topics of big data research mostly are the introductions of theories, techniques, and methods related to big data, such as data-mining application architecture [41], SINFONI [42], MapReduce [43], Hive [44], the pathologies of big data [45], large-scale electrophysiology of big data [46], and so on. Stage II (2010–2016) has a rapid growth spurt in annual research outcomes.

In this stage, there were four articles in 2010; by 2016, the number of annual articles sharply increased to 2402, which represented that the number of annual articles had increased 600 times over the past six years. Such a significant change is attributable, to a great extent, to the growing research enthusiasm of governments, scholars and enterprises, such as the research report "Big Data: The Next Frontier for Innovation, Competition, and Productivity" [1], the declaration "The Big Data Research and Development Initiative" [2], the book "Big Data: A Revolution That Will Transform How We Live, Work, and Think" [47], and the worldwide opening of the big data subject. All of these promoted effectively the rapid development of scientific research works related to big data. The studies of big data gradually matured.

To further verify the rapid growth trend of research literatures related to big data in Stage II, we develop a curve-fitting, and find that the curve conforms to the exponential distribution: $y = 5.076e^{1.175t}$, where y is the amount of annual publications, and t is a time sequence between 2010 and 2016. Moreover, according to goodness of fit test, the closer R^2 (R Square, coefficient of determination) is to 1, the better fitting degree of the regression line. The quantitative result shows that $R^2 = 0.974$; R^2 is very close to 1. This result indicates the fitting regression curve has a good reliability of forecast and goodness of fit. Therefore, the annual publications of big data research between 2010 and 2016 grow exponentially and big data has become a hot topic. It is worth worldwide scholars to pay more attention.

Figure 2 shows the annual number of authors who published articles from 2002 to 2016. The line in Figure 2 is similar to the annual publications distribution in Figure 1. There were four authors in 2002, and 14 authors in 2010. However, the number of authors sharply increased to 9558 in 2016, which shows that the number of annual authors has increased hundreds of times over the past several years.

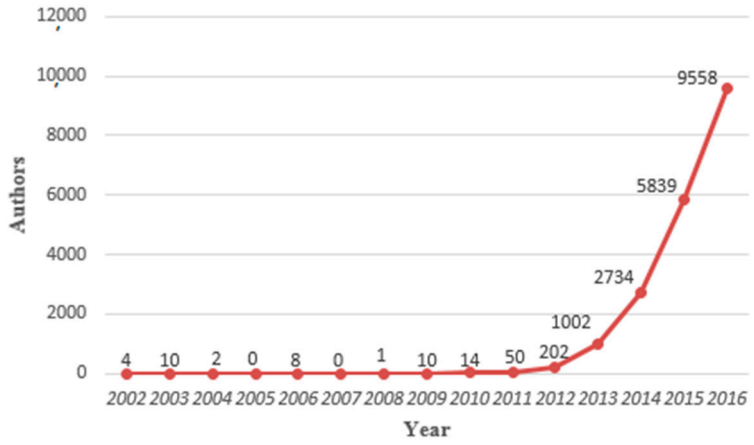

Figure 2. Annual authors distribution.

To further evaluate the annual collaboration ratio of researchers in the big data research field, we depicted average participants per article from 2002 to 2016 (shown in Figure 3). However, we excluded 2005 and 2007 because of mathematics. Figure 3 reveals a trend of collaboration among authors in the big data research field. In 2003, the average number of participants per article reached a maximum of five. However, the value hits rock bottom twice at 2004 and 2008 because of having an independent author in each article. After 2008, this number continued to rise, and reached 3.98 in 2016. Moreover, there was only a slight fluctuation from 2012 to 2016, which indicated that the average number of participants per article in the big data field were between three and four authors. The research collaboration, to some extent, ensured the quality of the publications.

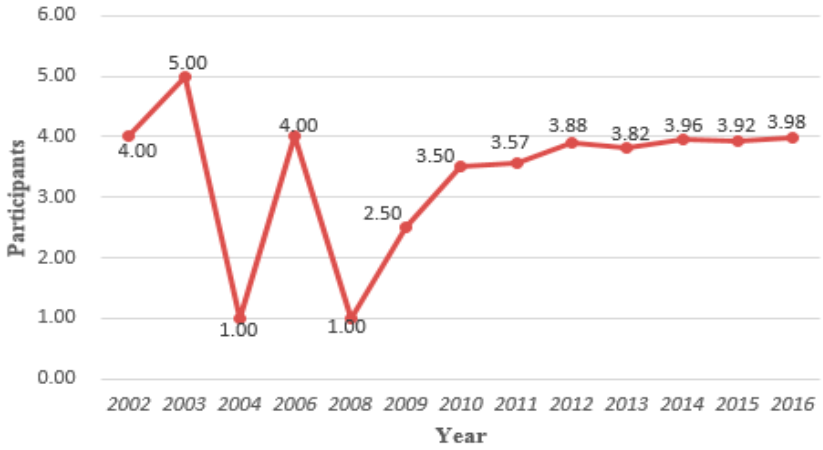

Figure 3. Average participants distribution per article.

3.2. Journals Distribution and Co-Citation Network

3.2.1. Core Journals Identification

In this section, we examined 1729 different academic journals. According to Price law, core journals must be the journals which published more than N (note: N = 0.749 × square (69) ≈ 7) articles. According to the statistical analysis, there are 154 core journals. Table 1 lists the top 10 academic journals in descending order of publications. The core academic journal with the most publications in big data research is PLoS One (69), followed by IEEE Access (63), and Big Data (52). There is a narrow gap of less than five publications among Cluster Computing the Journal of Networks, Software Tools and Applications (45), Neurocomputing (45), Journal of Supercomputing (43), and Concurrency and Computation: Practice and Experience (41). IEEE Network, Information Sciences, and International Journal of Distributed Sensor Networks have equal publications (31).

In addition, according to the Journal Citation Reports in the WoS, *IEEE Network* simultaneously has the highest impact factor (IF, 7.230) and immediacy index (1.638) in these top 10 most publications core academic journals of big data research. Moreover, the top 10 academic journals published 450 articles, which account for 9.1% of overall published articles from 2002 to 2016. Simply, it indicates that 0.6% of academic journals in the big data research field published 9.1% of overall articles from 2002 to 2016. It conforms to what is known as a "Matthew effect" in academic journals distribution.

Table 1. Top 10 most publications core academic journals.

Journal	The Number of Publications	Impact Factor in 2016	Five Year Impact Factor	Immediacy Index
PLoS One	69	2.806	3.394	0.396
IEEE Access	63	3.244	3.870	0.607
Big Data	52	1.239	2.292	0.286
Cluster Computing—The Journal of Networks, Software Tools and Applications	45	2.040	2.076	0.339
Neurocomputing	45	3.317	3.211	0.819
Journal of Supercomputing	43	1.326	1.349	0.282
Concurrency and Computation: Practice and Experience	40	1.133	1.219	1.065
IEEE Network	31	7.230	6.410	1.638
Information Sciences	31	4.832	4.732	1.041
International Journal of Distributed Sensor Networks	31	1.239	1.315	0.238

3.2.2. Journals Co-Citation Network

Journals co-citation analyses usually are employed to discover the journals that formed the intellectual base of a knowledge domain. Figure 4 shows the highly cited journals co-citation network from 2002 to 2016. This network is constructed by the top 50 most cited references in each given time slices based on 337 iterations. It contains 195 journals and 489 links among them. Table 2 lists the top 10 highest co-cited journals from 2002 to 2016. The journals with frequencies more than 1000 include Nature (1899), Science (1844), Lecture Notes in Computer Science (1436), PLoS One (1210), Communications of the ACM (1197), and Proceedings of the National Academy of Sciences of the United States of America (1128). These six journals are the primary publishing outlets and the dominant citing sources for big data scholars, and contribute to the sustainable intellectual base formation of big data.

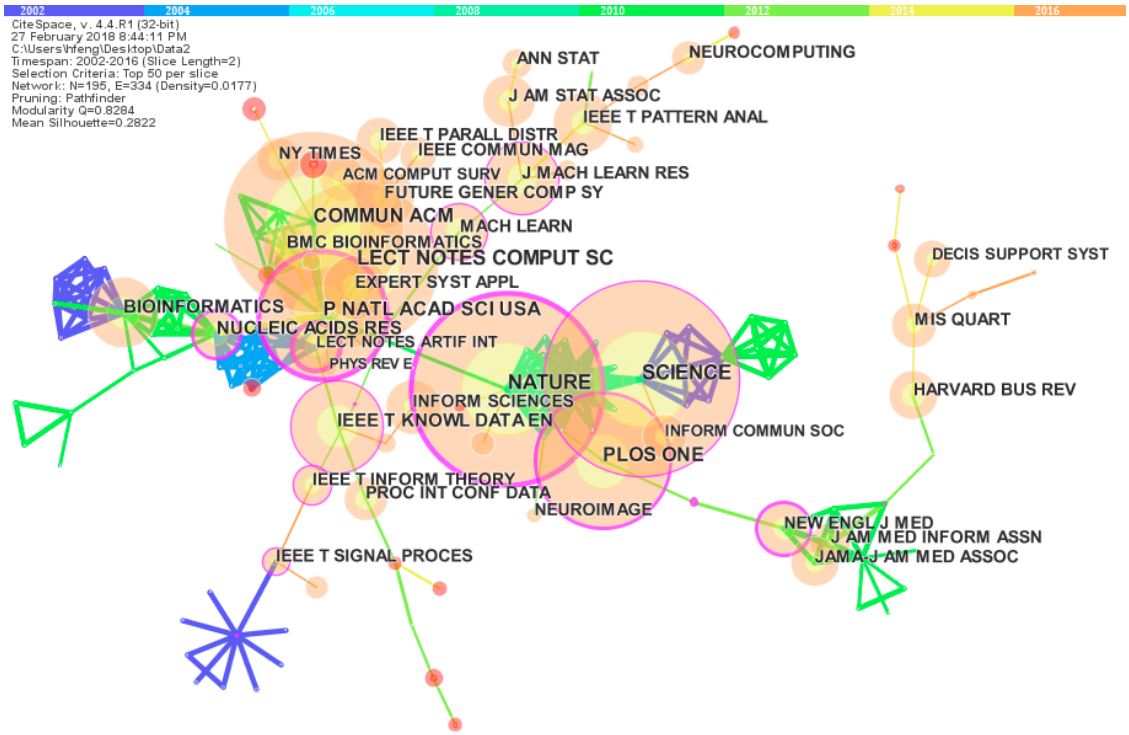

Figure 4. Journals co-citation network.

More interestingly, the nodes with purple tree rings around the outer rim indicate that some highly cited journals have high betweenness centrality (betweenness centrality ≥ 0.23), such as Nature (0.54), Proceedings of the National Academy of Sciences of the United States of America (0.56), Nucleic Acids Research (0.31), and PLoS One (0.23). These pivotal journals make connections to others in the journal co-citation network (see Figure 4). Some big nodes with thinner purple rings indicate that high co-citation scores do not necessarily have a high betweenness centrality. For example, Lecture Notes in Computer Science has a high co-citation frequency node (1436) and a lower betweenness centrality (0.04). Moreover, the journals in multidisciplinary sciences and Computer Science received more citations. It means that knowledge from multidisciplinary sciences and computer science is therefore a major intellectual resource for big data scholars. In addition, a significant co-citation burst journal is visualized by the node with red inner tree rings. The size of the red inner tree rings node represents the strength of its burst property. As shown in Figure 4, Big Data Revolution is a journal with red inner tree rings, suggesting that its citations have rapidly increased between 2014 and 2016.

Table 2. Frequency distribution and between centrality of the highest co-cited Journals.

Journal	Frequency	Centrality	IF	Categories
Nature	1899	0.54	40.137	Multidisciplinary Sciences
Science	1844	0.16	37.205	Multidisciplinary Sciences
Lecture Notes in Computer Science	1436	0.04	0.402	Computer Science (Theory and Methods)
PloS One	1210	0.23	2.806	Multidisciplinary Sciences
Communications of the ACM	1197	0.06	4.027	Computer Science (Hardware and Architecture; Software Engineering; Theory and Methods)
Proceedings of the National Academy of Sciences of the United States of America	1128	0.56	9.661	Multidisciplinary Sciences
Bioinformatics	908	0.1	7.307	Biochemical Research Methods; Biotechnology and Applied Microbiology; Mathematical and Computational Biology
Nucleic Acids Research	773	0.31	10.162	Biochemistry and Molecular Biology
IEEE Transactions on Knowledge and Data Engineering	720	0.11	3.438	Computer Science (Artificial Intelligence; Information Systems); Engineering, Electrical and Electronic
Journal of Machine Learning Research	579	0.12	5.000	Automation and Control Systems; Computer Science, Artificial Intelligence

Source: the Web of Science and Journal Citation Reports 2016; IF, impact factor in 2016.

3.3. Institutions Distribution and Collaboration

3.3.1. Core Institutions Identification

It is significant to study the institutions distribution in a research field. Commonly the number of publications is an important index to measure academic level, scientific research ability, and status of the authors and their institutions in a specific field. Core institutions are important leaders in a research field. However, the names of academic institutions might change over time. Therefore, to avoid inconsistent signatures, we firstly need to standardize the names of academic institutions. In this section, we reserved the top level names, and constructed uniform names of academic institutions. Eventually we achieved 4137 different institutions.

According to Price law, core institutions must be the institutions who published more than N (note: N = 0.749 $\times$ square (153) $\approx$ 10) articles. According to the statistical analysis, there are 265 core institutions in development history of big data research from 2002 to 2016. Table 3 lists the top 10 most prolific academic institutions in descending order of publications. The top 10 most prolific academic institutions are almost all colleges and universities from USA and China. Among them, Chinese Academy of Sciences is the most prolific institution (153), followed by Tsinghua University (126) and University of California, Los Angeles (91). Stanford University and MIT just have a narrow gap less than two articles. The top 10 academic institutions published 895 articles, which account for 18.2% of overall published articles from 2002 to 2016. Simply, it reveals that 0.2% of institutions in the big data research field published 18.2% of overall articles published from 2002 to 2016. It conforms to what is known as a "Matthew effect" in institutions distribution.

Table 3. Top 10 most prolific academic institutions.

Institution	Country	The Number of Publications	Centrality	Year
Chinese Academy of Sciences	China	153	0.15	2012
Tsinghua University	China	126	0.04	2013
University of California, Los Angeles	USA	91	0.33	2009
Stanford University	USA	84	0.29	2013
MIT	USA	82	0.06	2011
University of Washington	USA	75	0.08	2012
University of Michigan	USA	73	0.06	2013
Harvard University	USA	72	0.17	2012
University of California, San Diego	USA	70	0.18	2010
University of Minnesota	USA	67	0.16	2011

3.3.2. Institutions Collaboration Network

To enhance overall research strength in a scientific field, scientific research collaboration usually is an important means, which allows researchers to play their own academic advantages and share information [48]. Moreover, the level of scientific research collaboration is one of important indexes to evaluate the academic level, scientific research ability, and status of institutions in a specific field. To discuss the scientific research collaboration in the big data research field, we constructed a scientific research collaboration network (shown in Figure 5).

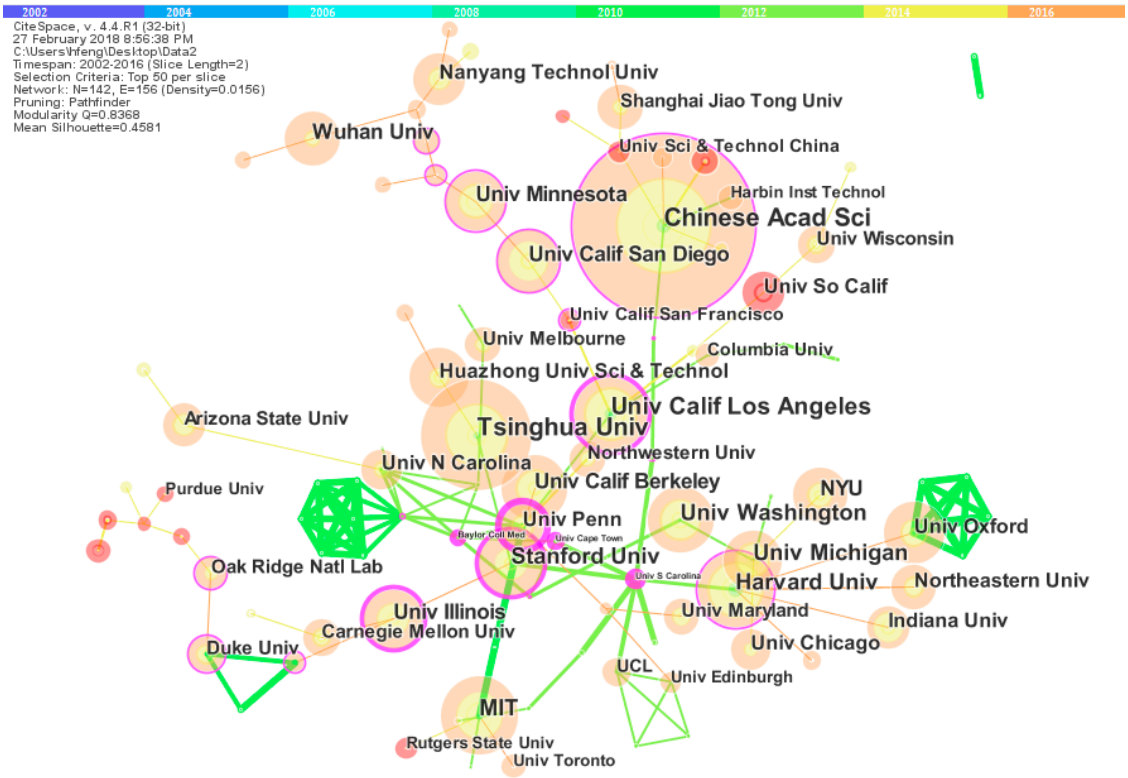

Figure 5. Institutions collaboration network.

This scientific research collaboration network consists of 142 nodes and 342 links. Each node represents an institution, and is depicted with a series of tree rings across multiple time slices. The size of each node is proportional to the total number of publications in each institution [8]. Each link between two nodes represents a scientific research collaboration relationship, and the thickness of a link represents the scientific research collaboration strength [49]. As shown in Figure 5, there are a wider scientific research collaboration among different institutions. For example, Chinese Academy of Sciences is a red tree ring node, which has the most publications 153 and cross-connects with University of Sydney, Harbin Institute of Technology, University of Science and Technology China, Peking University, Beijing Normal University, and Otto Von Guericke University. The gold-colored link between University of Sydney represents that the first scientific research collaboration year is between 2014 and 2015. However, the nodes with more publications do not certainly have stronger betweenness centrality scores. As listed in Table 3, compared with Stanford University (0.29), Chinese Academy of Sciences has a weaker betweenness centrality score (0.15). This means that Chinese Academy of Sciences plays a weaker intellectual pivotal role among the institutions collaboration network. Furthermore, University of South Carolina with the highest betweenness centrality score (0.63) has a very low co-occurrence frequency. These results reveal that the current research relationship is rather weak and diffuse. In addition, three thicker lines, which are linked with Otto Von Guericke University (link strength: 0.3), University of Sydney (link strength: 0.23), and Harbin Institute of Technology (link

strength: 0.21) respectively, indicate the stronger collaboration relationships. Moreover, two green lines, which are linked with Otto Von Guericke University and Harbin Institute of Technology, indicate that the first collaboration among them is in the 2012–2013 time slice.

3.4. Authors Distribution and Co-Citation Network

3.4.1. Core Authors Identification

It is interesting to study the core authors distribution in the big data research field. Usually the amount of publications is an important index to evaluate the academic level, advancement, and position of an author in a specific research field. In addition, core authors also are particular important leaders in a research field. However, the names of authors may be full and abbreviated names downloaded from the WoS. The same abbreviated name might stand for different full names. For example, Y ZHANG represent Yin ZHANG, Yi ZHANG, or Yong ZHANG. Similarly, Y WANG may represent Yige WANG, Yi WANG, or Yuhang WANG, et al. Moreover, a same full name may be different authors. For example, Yin ZHANG can be Yin ZHANG who comes from the School of Computer Science or Information Technology at Huazhong University of Science and Technology (HUST), or even Yin ZHANG who comes from the School of Economics and Law at Zhongnan University. They are different persons. To avoid inconsistent signatures, we therefore need to examine seriously the unique full names and affiliated institutions of the authors, count the amount of the articles, and order the different authors in descending articles. Eventually, we got 16,404 different authors who published 4927 articles from 2002 to 2016. It indicates that the average number of collaborator per article is between three and four in the big data research field. This result coincides with the publications distribution (see "publications distribution" section).

According to Price law, core authors must be the authors who published more than M (note: $M = 0.749 \times \text{square}(18) \approx 3$) articles. According to the statistical analysis, there are 229 core authors. Table 4 lists the top 10 most prolific authors by the amount of articles from 2002 to 2016. Among them, Ranjan, Rajiv ranks first with 18 articles, and Zomaya, Albert Y ranks second with 17 articles. If we do not exclude the collaborative articles, the top 10 authors published 138 articles, which account for 2.8% of overall articles published from 2002 to 2016. This means approximately 0.6% of overall authors published 2.8% of overall articles between 2002 and 2016. It conforms to what is known as a "Matthew effect" in core authors distribution. However, the number of all core authors is only 229, which accounts for 1.4% of overall authors. This means that 98.6% authors are not core authors. This result shows that research strengths are still comparatively weak and fragmented. Moreover, from the geographical perspective, the core authors from Australia account for 50% of the top 10 core authors, which means that Australia currently has a stronger research strength in the big data field compared with other countries.

Table 4. Top 10 most prolific authors.

Author	Institution	Country	Publications
Rajiv Ranjan	Computational Informatics, CSIRO, Australian National University	Australia	18
Albert Y. Omiya	School of Information Technologies, University of Sydney	Australia	17
Lizhe Wang	Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences	China	16
Xuyun Zhang	Engineering and Information Technology, University of Technology Sydney	Australia	14
Jinjun Chen	Engineering and Information Technology, University of Technology Sydney	Australia	14
Laurence T Yang.	School of Computer Science and Technology, Huazhong University of Science and Technology	China	13
Chang Liu	Engineering and Information Technology, University of Technology Sydney	Australia	12
Keqin Li	Department of Computer Science State University of New York New Paltz	USA	12
Francisco Herrera	Dept. of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada	Spain	11
Samee U. Khan	electrical and computer engineering at North Dakota State University	USA	11

3.4.2. Core Authors Collaboration Network

To deeply understand the current research collaboration of core authors, we also developed the social network analysis based on UCINET (shown in Figure 6). Because the original network has lower density (0.0240), we deleted some isolates and pendants (nodes with degree one) to increase the identifiability of the network. Eventually, the core authors collaboration network consists of 44 nodes and four small networks. The two bigger networks have 25 nodes and 12 nodes separately. This means that more core authors nodes tend to be the isolates or the pendants. In general, the overall core authors collaboration network is relatively decentralized. This result reveals that the research collaboration among core authors is not enough close in the big data field.

As shown in Figure 6, the size of each node represents the between centrality score. According to the between centrality measure, Rajiv Ranjan is the central node with highest between centrality, as it form the densest bridges with other nodes. In addition, Laurence T. Yang, Albert Y. Zomaya, and Kim-Kwang Raymond Choo also have a higher between centrality. More interesting, nine authors (Rajiv Ranjan, Albert Y. Zomaya, Lizhe Wang, Xuyun Zhang, Jinjun Chen, Laurence T. Yang, Chang Liu, Keqin Li, and Samee U. Khan) listed in Table 4 have close bonds with each other in the biggest network.

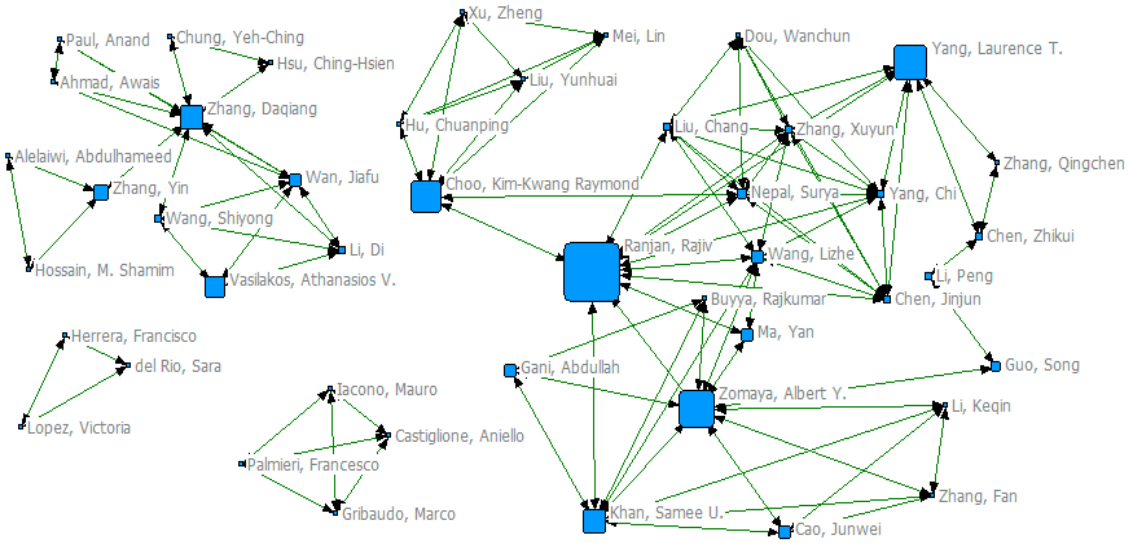

Figure 6. Core authors collaboration network.

3.4.3. Authors Co-Citation Network

Unlike the core authors collaboration analysis, authors co-citation analysis focuses on the co-cited authors who published the co-cited articles. Authors co-citation relationship is critical to understand the academic communication and knowledge base diffusion in a specific research field [11]. The more two authors are co-cited, the closer the intellectual relationship is. Figure 7 shows the overall landscape view of authors co-citation network in the big data research field. The top 50 most cited authors in each slice are used to construct the authors co-citation network based on 137,929 valid distinct references. This network consists of 262 nodes and 593 links. Moreover, this network has a very high modularity (0.9102), which can be considered that the specialties in science mapping are clearly defined in terms of co-citation clusters. The mean silhouette score (0.4179) is relatively lower mainly because of the numerous small clusters [15]. Therefore, we just need to focus on the major clusters.

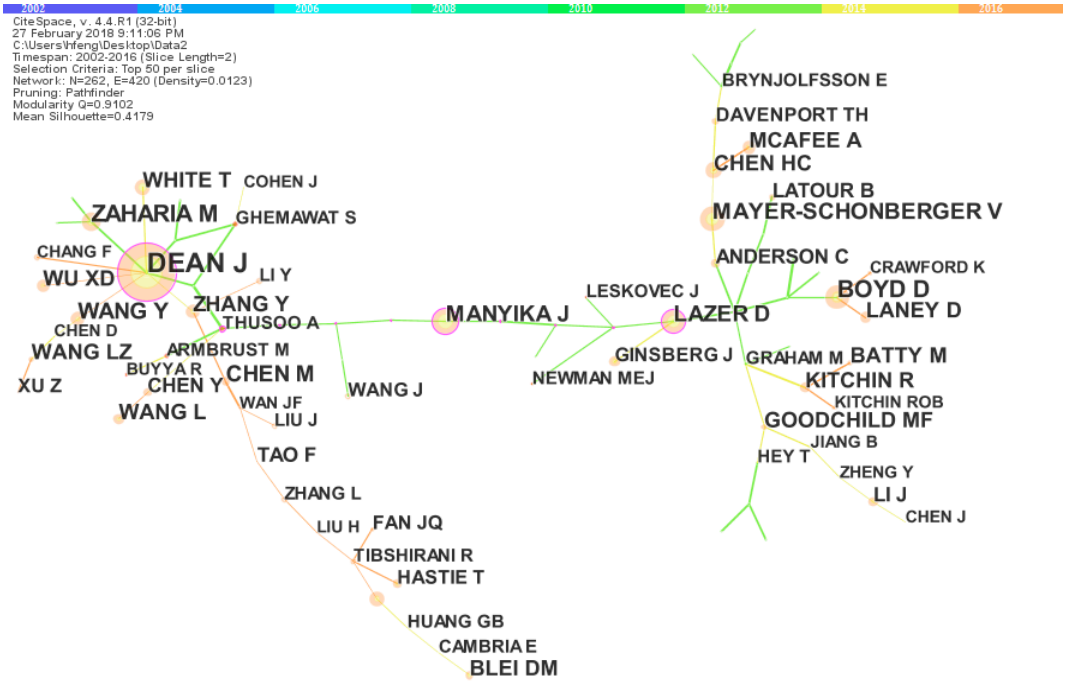

Figure 7. Authors co-citation network.

As shown in Figure 7, each node with a series of tree rings across multiple time slices represents an author. The size of each node is proportional to the total authors co-citation frequency. Each link between two nodes represents a co-citation relationship, and the thickness of a link shows the co-citation link strengths [49]. For example, Dean J is the biggest tree rings node, which has the most co-citation articles (493) and cross-connects with White T, Wu XD, Wang Y, Zaharia M, Isard M, and Condie T. The green-colored link with Zaharia M represents that the first co-citation year is 2012. In addition, three thicker lines, which are linked with Zaharia M (link strength: 0.57), Isard M (link strength: 0.53), and Condie T (link strength: 0.53) respectively, indicate some stronger co-citation relationships. Table 5 lists the most co-citation authors in the big data research field. Most of them come from USA. The highest co-citation author is Jeffrey Dean (493) at Google Inc., followed by Danah Boyd (224) at Microsoft Research, Matei Zaharia (212), James Manyika (202), Viktor Mayer-schönberger (192), Lazer David (192), Breiman Leo (170), LiZhe Wang (136), Hsinchun Chen (135), and Andrew Mcafee (128).

Table 5. Top 10 most co-citation authors.

Author	Country	First Co-Citation Year	Frequency	Centrality
Jeffrey Dean	USA	2012	493	0.1
Danah Boyd	USA	2012	224	0.01
Matei Zaharia	USA	2012	212	0.01
James Manyika	USA	2012	202	0.11
Viktor Mayer-schönberger	USA	2014	192	0.04
Lazer David	USA	2012	192	0.1
Breiman Leo	USA	2014	170	0.01
LiZhe Wang	China	2014	136	0
Hsinchun Chen	China	2014	135	0.03
Andrew Mcafee	USA	2014	128	0

The node with purple tree rings around the outer rim indicates this co-cited author has a high betweenness centrality, and this author tends to be a pivotal scholar whose work linked different

disciplines, research topics, or stages in the big data field. Table 6 lists all authors with high betweenness centrality (betweenness centrality $\geq$ 0.1). For example, Savage M (0.12) proposed "The Coming Crisis of Empirical Sociology" (2007) and "Contemporary Sociology and the Challenge of Descriptive Assemblage" (2009) to argue the challenges of "social" transactional data and descriptive assemblage. Savage M is a milestone author who argues how to develop sociology within the big data environment. Other authors with a strong betweenness centrality include Manyika J (0.11), Thusoo A (0.11), Schadt EE (0.11), Barabasi AL (0.11), and Chaudhuri S (0.11). Thusoo A (2009; 2010) presented the well-known Hive—a petabyte scale data warehouse using Hadoop. Schadt EE (2010) proposed the computational solutions to large-scale data management and analysis. Barabasi AL (2010) discussed the emergence of scaling in random networks and the development of large networks is governed by robust self-organizing phenomena that go beyond the particulars of the individual systems. However, it is not the case that a highly co-cited author positively has a high betweenness centrality. These authors are visualized by the small nodes with thicker purple tree rings, such as Savage M, Thusoo A, Schadt EE, Barabasi AL, and Chaudhuri S. Only a node simultaneously with a high co-citation frequency and a betweenness centrality is the milestone author. For example, as listed in Table 5, Manyika J (0.11) at McKinsey global institute (MGI, San Francisco, CA, USA) firstly released the research report "Big Data: The Next Frontier for Innovation, Competition, and Productivity" in May 2011. This report is a milestone publication, which triggered the research enthusiasm of scholars worldwide.

Table 6. High betweenness centrality authors.

Author	First Co-Citation Year	Frequency	Centrality
Savage M	2012	15	0.12
Manyika J	2012	202	0.11
Thusoo A	2012	57	0.11
Schadt EE	2012	11	0.11
Barabasi AL	2012	10	0.11
Chaudhuri S	2012	9	0.11
Dean J	2012	493	0.1
Lazer D	2012	192	0.1
Tien JM	2012	15	0.1
Stonebraker M	2012	11	0.1
Isard M	2012	9	0.1
Zhang D	2013	7	0.1

In addition, the node with red inner rings in Figure 7 means a significant co-citation burst. It reveals that the co-citation frequency of authors increased rapidly within a given time period. The size of the red inner tree rings node represents the strength of its burst property. As shown in Figure 7, there are 25 nodes with red inner tree rings. It means that there are 25 authors with co-citation bursts in big data research from 2002 to 2016. These authors may have profound impacts on the big data research, and their work should be paid more attention because they may impact the sustainable development directions of big data research. Table 7 lists the top 25 cited authors with strongest citation bursts. Among them, Ghemawat S with the strongest citation burst (11.6177) demonstrated the Google file system, a scalable distributed file system for large distributed data-intensive applications, which guided the big data storage research. Thusoo A, with the second strongest citation burst (9.8427), presented the well-known Hive based on Hadoop. In addition, Hey T, Armbrust M, Wang C, Cohen J, and Buyya R, etc. also made important contributions to the sustainable development of big data research from different perspectives.

Table 7. Top 25 Cited Authors with Strongest Citation Bursts.

Cited Authors	Year	Strength	Begin	End	2002–2016
Ghemawat S	2002	11.6177	2013	2016	
Thusoo A	2002	9.8427	2012	2016	
Hey T	2002	9.437	2012	2016	
Armbrust M	2002	9.0315	2012	2016	
Wang C	2002	8.9019	2014	2016	
Cohen J	2002	8.6382	2014	2016	
Buyya R	2002	8.3746	2014	2016	
Newman MEJ	2002	7.8172	2012	2016	
Chen J	2002	7.5844	2014	2016	
Leskovec J	2002	7.2113	2012	2016	
Schadt EE	2002	6.9707	2012	2013	
Savage M	2002	6.9707	2012	2013	
Brynjplfsson E	2002	6.4045	2012	2016	
Chen D	2002	5.2202	2014	2016	
Stonebraker M	2002	5.066	2012	2013	
Isard M	2002	5.066	2012	2013	
Barabasi AL	2002	4.4317	2012	2013	
Lotan G	2002	4.4317	2012	2013	
Yang H	2002	4.4317	2012	2013	
Christakis NA	2002	3.7977	2012	2013	
Chaudhuri S	2002	3.7977	2012	2013	
Callon M	2002	3.164	2012	2013	
Von Ahn L	2002	3.164	2012	2013	

3.5. Keywords Co-Word Network

Keywords usually provide the core content and principal research methods of each article. Keyword co-word analysis can be applied to identify research topics and monitor research frontiers of a knowledge domain [50]. To construct a reasonable keywords co-word network, SATI3.2 was used to extract the high frequency keywords and form keywords co-occurrence matrix. Moreover, commonly keywords must be integrated and unified because of synonymy and polysemy. We removed some broad words (such as algorithm, model, design, analysis, research, etc.), and eventually got the top 80 keywords. Table 8 lists the top 80 high frequency keywords.

Table 8. Top 80 high frequency keywords.

	Keywords	Frequency		Keywords	Frequency		Keywords	Frequency
1	Big Data	1834	28	bioinformatics	35	55	Reliability	19
2	cloud computing	213	29	technology	35	56	deep learning	19
3	machine learning	207	30	Database	35	57	Education	19
4	MapReduce	164	31	Security	33	58	parallel processing	19
5	data mining	144	32	data science	31	59	Sentiment analysis	19
6	big data analysis	128	33	text mining	30	60	computational social science	19
7	Hadoop	106	34	crowdsourcing	29	61	NoSQL	19
8	social media	101	35	Internet	29	62	Design	19
9	Internet of Things	80	36	ethics	29	63	innovation	18
10	privacy	77	37	scalability	29	64	knowledge	18
11	data analysis	71	38	Business Intelligence	28	65	informatics	18
12	Prediction	69	39	Data quality	26	66	software	17
13	computing	59	40	surveillance	26	67	epidemiology	17
14	Algorithm	58	41	open data	26	68	Spark	17
15	Twitter	54	42	Genomics	24	69	natural language processing	17
16	Classification	52	43	systems	24	70	precision medicine	17
17	networks	52	44	GIS	23	71	time series	17
18	optimization	51	45	Distributed computing	23	72	methodology	17
19	Cloud	49	46	Distributed	22	73	data management	16
20	model	49	47	Feature selection	21	74	Decision making	16
21	visualization	47	48	Measurement	20	75	sampling	16
22	Social network	46	49	statistics	20	76	Scheduling	16
23	performance	44	50	Healthcare	20	77	social	16

Table 8. *Cont.*

	Keywords	Frequency		Keywords	Frequency		Keywords	Frequency
24	clustering	44	51	Ontology	20	78	Parallel	16
25	smart city	39	52	Data protection	20	79	Storage	16
26	Information	37	53	Parallel computing	20	80	energy	15
27	management	35	54	energy efficiency	20			

As listed in Table 8, the highest frequency keyword absolutely is Big Data (1834), followed by cloud computing (213), machine learning (207), MapReduce (164), data mining (144), big data analysis (128), Hadoop (106), and social media (101). Other keywords with less than 100 frequency include Internet of Things (IoT, 80), privacy (77), data analysis (71), Prediction (69), Computing (59), Algorithm (58), Twitter (54), Classification (52), networks (52), and optimization (51). These keywords reflect the current research hotspots. Besides these high frequency keywords, the identification of some other relevant keywords indicate the current emerging research areas such as social network (46), smart city (39), bioinformatics (35), crowdsourcing (29), ethics (29), Genomics (24), GIS (23), Healthcare (20), Education (19), epidemiology (17), precision medicine (17), and energy (15).

To understand the relationship among these keywords, we construct the keywords co-word network (shown in Figure 8). Each node represents a keyword. The size of each node is proportional to the betweenness centrality of keywords. It is not surprising that some well-known words, such as well-known topics including data mining, cloud computing, machine learning, MapReduce, Hadoop, social media, and visualization, have higher co-occurrence frequencies and betweenness centrality scores. Besides these topics, we find that data science, including data privacy, data management, data protection, and data quality, etc., also gradually enter the researchers' considerations. In addition, deep learning, algorithm, model, performance, optimization are some interesting findings in big data research. These keywords reveal the popular research hotspots and will have profound impacts on future sustainable development research directions of big data.

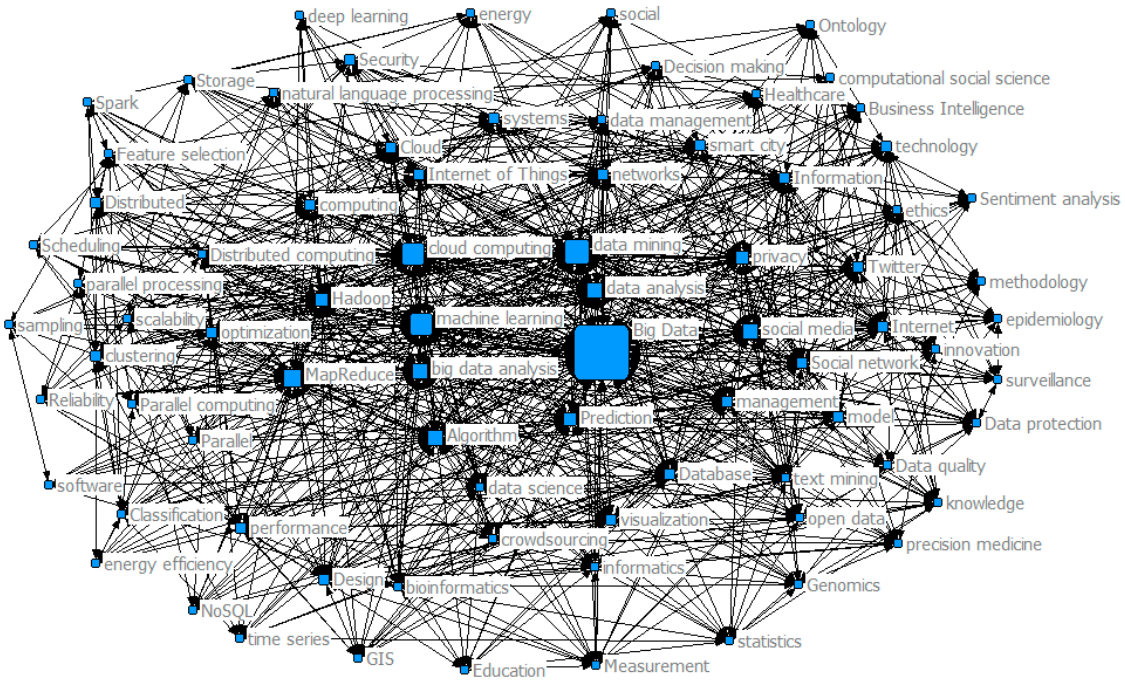

Figure 8. Keywords co-word network.

4. Discussion and Conclusions

In this study, we extracted the bibliometric data of 4927 effective journal articles listed in the WoS between 2002 and 2016, visualized the intellectual structure and hotspots of big data research from the bibliometric perspective, and presented the results in terms of publications distribution, journals distribution and co-citation network, institutions distribution and collaboration network, authors distribution, collaboration network and co-citation network, and keywords co-word network. The main findings of this study are as follows:

According to publications distribution, we found the annual growth trend of big data research outcomes and authors, as well as the changes of co-author numbers in each article. The research outcomes in the embryonic stage (2002–2009) were very few, but an exponential growth spurt was generated from 2010 to 2016. In addition, the growth trend of annual authors is similar to the annual publications distribution. Moreover, we found that the average number of participants per article in the big data field were between three and four authors.

The current core journal with the most publications was PLoS One, followed by IEEE Access and Big Data. However, the top five co-citation journals, which contributed to the sustainable intellectual base formation of big data, were Nature, Science, Lecture Notes in Computer Science, PLoS One, and Communications of the ACM. Among them, Nature had the highest betweenness centrality. Moreover, the most categories of top 10 co-citation journals were multidisciplinary sciences and computer science, which is closely related to the nature of big data science.

There was a wider scientific research collaboration among institutions in big data research. The top three core institutions in terms of publications were Chinese Academy of Science, Tsinghua University, and University of California, Los Angeles. However, the institutions with most publications had lower betweenness centrality scores, signifying that these institutions still were scattered and did not get general consent. Hence, the current research relationships among the institutions were rather weak and diffuse in the big data research field. With sustainable development and prosperity of big data research, the research collaboration relationships will be strengthened and increasingly firm.

According to the core authors identification, compared with USA and China, Australia had a current greater research strength in big data research. However, according to authors co-citation analysis, the top 10 most co-cited authors mainly came from USA and China. Moreover, some special authors with most co-citation frequency, high betweenness centrality and strong citation bursts were also identified, such as the most co-citation authors, pivotal scholars or intellectual turning pointers, and the strongest citation burst authors. These authors had contributed to the sustainable development of big data from different perspectives, and have a profound impact on the big data field. More attention should be paid to their work.

Keywords co-word analysis detected the current research hotspots and emerging topics, including not only the well-known research hotspots like data mining, cloud computing, machine learning, MapReduce, Hadoop, social media, and visualization, but also some emerging research topics, such as data science (data privacy, data management, data protection, and data quality, etc.), deep learning, and so on. Moreover, algorithm, model, performance and optimization are also gradually entering researchers' considerations. In addition, keywords co-word analysis also detected the current emerging and sustainable development applications areas of big data, such as social network, smart city, bioinformatics, crowdsourcing, ethics, Genomics, GIS, Healthcare, Education, epidemiology, precision medicine, and energy.

As an emerging hot topic, big data has changed the lives of human beings, and driven some changes in thinking, decision making, and research paradigms. Moreover, big data itself contains important strategic resources for social trends, market changes, scientific and technological development and national security. Many colleges and universities have opened big data disciplines and courses. However, as a new emerging cross-discipline, the sustainable development of big data still faces many very complicated and difficult challenges, such as the heterogeneity and incompleteness of data, the efficiency of big data processing, big data security and privacy protection, high energy

consumption, and so on. On the one hand, these challenges indicate some sustainable development directions of future big data research. On the other hand, these challenges are also unprecedented opportunities of big data sustainable development. With the increasing improvement of physical infrastructure constructions and policy making at national and institutional levels, and the further breakthroughs of information technologies (computer networks, distributed systems, cloud computing, data storage, machine learning, and so on), these above issues will be gradually solved. A bright future of big data science is coming.

Acknowledgments: This work was financially supported by Project 14YJC870008 (Humanity and Social Science Youth Foundation of Ministry of Education, China), Project GD13YTS01 (Philosophy and Social Science Foundation of Guangdong Province during the "12th Five-Year Plan"), Project 2016A070705052 (Public Welfare and Ability Construction Special Foundation of Science and Technology Plan, Guangdong Province), Project 2017GZYB98 (Philosophy and Social Science Foundation of Guangzhou City during the "13th Five-Year Plan"), Provincial Nature Science Foundation of Guangdong (2015A030310271 and 2015A030313679)and Project 2017B1015 (Zhongshan City Science and Technology Bureau).

Author Contributions: Writing: Feng Hu; Providing case and idea: Feng Hu, Wei Liu, Sang-Bing Tsai; Providing revised advice: Sang-Bing Tsai, Junbin Gao, Ning Bin, Quan Chen.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McKinsey Global Institute. Big Data: The Next Frontier for Innovation, Competition, and Productivity. 2011. Available online: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (accessed on 10 February 2017).
2. Office of Science and Technology Policy Executive Office of the President. Big Data Research and Development Initiative. 2012. Available online: https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf (accessed on 10 February 2017).
3. Feng, Z.Y.; Guo, X.H.; Zeng, D.J.; Chen, Y.B.; Chen, G.Q. On the research frontiers of business management in the context of Big Data. *J. Manag. Sci. China* **2013**, *16*, 1–9. (In Chinese)
4. Sustainable Development Goals. Available online: http://www.un.org/sustainabledevelopment/sustainable-development-goals/ (accessed on 30 January 2018).
5. McAfee, A.; Brynjolfsson, E. Big data: The management revolution. *Harv. Bus. Rev.* **2012**, *90*, 60–68. [PubMed]
6. Naimi, A.I.; Westreich, D.J. Big Data: A revolution that will transform how we live, work, and think. *Am. J. Epidemiol.* **2014**, *179*, 1143–1144. [CrossRef]
7. Jeon, S.; Hong, B. Monte Carlo simulation-based traffic speed forecasting using historical big data. *Future Gener. Comput. Syst.* **2016**, *65*, 182–195. [CrossRef]
8. Chen, C. CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 359–377. [CrossRef]
9. Zhao, R.; Wang, Q. The Mining and Analysis of Big Data Research Hotspots in the Field of Humanities and Social Science from Perspective of Information Measurement in China. *J. Intell.* **2016**, *35*, 93–98. (In Chinese)
10. Nobre, G.C.; Tavares, E. Scientific literature analysis on big data and internet of things applications on circular economy: A bibliometric study. *Scientometrics* **2017**, *111*, 463–492. [CrossRef]
11. Gu, D.; Li, J.; Li, X.; Liang, C. Visualizing the knowledge structure and evolution of big data research in healthcare informatics. *Int. J. Med. Inform.* **2017**, *98*, 22–32. [CrossRef] [PubMed]
12. Isasi, N.K.G.; Frazzon, E.M.; Uriona, M. Big Data and Business Analytics in the Supply China: A Review of the Literature. *IEEE Lat. Am. Trans.* **2015**, *13*, 3382–3391. [CrossRef]
13. Tsay, M.Y. Journal Bibliometric Analysis: A Case Study on the JASIST. *Malays. J. Libr. Inf. Sci.* **2008**, *13*, 121–139.
14. Chen, C.; Ibekwe-SanJuan, F.; Hou, J. The Structure and Dynamics of Co-Citation Clusters: A Multiple-Perspective Co-Citation Analysis. *J. Am. Soc. Inf. Sci. Technol.* **2010**, *61*, 1386–1409. [CrossRef]
15. Chen, C. Science Mapping: A Systematic Review of the Literature. *J. Data Inf. Sci.* **2017**, *2*, 1–40. [CrossRef]
16. Meho, L.I.; Yang, K. Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science versus Scopus and Google Scholar. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58*, 2105–2125. [CrossRef]

17. Pan, L.; Wang, S. A bibliometrics analysis on Chinese education research hotspots based on literature keywords co-occurrence knowledge map. *Educ. Res. Exp.* **2011**, *6*, 20–24.
18. No, H.J.; An, Y.; Park, Y. A structured approach to explore knowledge flows through technology-based business methods by integrating patent citation analysis and text mining. *Technol. Forecast. Soc. Chang.* **2015**, *97*, 181–192. [CrossRef]
19. Gautam, P. An overview of the Web of Science record of scientific publications (2004–2013) from Nepal: Focus on disciplinary diversity and international collaboration. *Scientometrics* **2017**, *113*, 1245–1267. [CrossRef]
20. Callon, M.; Courtial, J.P.; Turner, W.A.; Bauin, S. From translations to problematic networks—An introduction to co-word analysis. *Soc. Sci. Inf. Sci. Soc.* **1983**, *22*, 191–235. [CrossRef]
21. Small, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *J. Am. Soc. Inf. Sci.* **1973**, *24*, 265–269. [CrossRef]
22. Chen, C. Visualising semantic spaces and author co-citation networks in digital libraries. *Inf. Process. Manag.* **1999**, *35*, 401–420. [CrossRef]
23. White, H.D.; McCain, K.W. Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. *J. Am. Soc. Inf. Sci. Technol.* **1998**, *49*, 327–355.
24. Johnson, B.; Shneiderman, B. Tree-maps: A space filling approach to the visualization of hierarchical information structures. In Proceedings of the 2nd Conference on Visualization '91, San Diego, CA, USA, 22–25 October 1991; pp. 284–291.
25. Herman, I.; Melançon, G.; Marshall, M.S. Graph visualization and navigation in information visualization: A survey. *IEEE Trans. Vis. Comput. Graph.* **2000**, *6*, 24–44. [CrossRef]
26. Morris, S.A.; Yen, G.; Wu, Z.; Asnake, B. Timeline visualization of research fronts. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *55*, 413–422. [CrossRef]
27. Wise, J.A. The ecological approach to text visualization. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 1224–1233. [CrossRef]
28. Porter, A.L.; Cunningham, S.W. *Tech Mining: Exploiting New Technologies for Competitive Advantage*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2004; ISBN 9780471475675.
29. Chen, C. *Information Visualization: Beyond the Horizon*, 2nd ed.; Springer: New York, NY, USA, 2004.
30. Chen, C. Searching for intellectual turning points: Progressive knowledge domain visualization. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5303–5310. [CrossRef] [PubMed]
31. Bailón-Moreno, R.; Jurado-Alameda, E.; Ruíz-Baños, R. The scientific network of surfactants: Structural analysis. *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 949–960. [CrossRef]
32. Bailón-Moreno, R.; Jurado-Alameda, E.; Ruíz-Baños, R.; Courtial, J.P. Analysis of the scientific field of physical chemistry of surfactants with the unified scienctometric model. Fit of relational and activity indicators. *Scientometrics* **2005**, *63*, 259–276. [CrossRef]
33. Leydesdorff, L.; Schank, T. Dynamic animations of journal maps: Indicators of structural changes and interdisciplinary developments. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *59*, 1810–1818. [CrossRef]
34. Persson, O.; Danell, R.; Wiborg Schneider, J. How to use Bibexcel for various types of bibliometric analysis. In *Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at His 60th Birthday*; Åström, F., Danell, R., Larsen, B., Schneider, J.W., Eds.; International Society for Scientometrics and Informetrics: Leuven, Belgium, 2009; Volume 5, pp. 9–24.
35. Sci2 Team. Science of Science (Sci2) Tool. Indiana University and SciTech Strategies. 2009. Available online: https://sci2.cns.iu.edu (accessed on 10 August 2017).
36. Van Eck, N.J.; Waltman, L. Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523–538. [CrossRef] [PubMed]
37. Börner, K.; Huang, W.; Linnemeier, M.; Duhon, R.; Phillips, P.; Ma, N.; Price, M. Rete-netzwerk-red: Analyzing and visualizing scholarly networks using the network workbench tool. *Scientometrics* **2010**, *83*, 863–876. [CrossRef]
38. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. SciMAT: A new Science Mapping Analysis Software Tool. *J. Am. Soc. Inf. Sci. Technol.* **2012**, *63*, 1609–1630. [CrossRef]
39. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. Science Mapping Software Tools: Review, Analysis, and Cooperative Study among Tools. *J. Assoc. Inf. Sci. Technol.* **2011**, *62*, 1382–1402. [CrossRef]
40. Liu, Q.; Ye, Y. A Study on Mining Bibliographic Records by Designed Software SATI: Case Study on Library and Information Science. *J. Inf. Sources Manag.* **2012**, *1*, 50–58. (In Chinese)
41. Petersohn, H. Data-mining application architecture. *Wirtschaftsinformatik* **2004**, *46*, 15–21. [CrossRef]

42. Abuter, R.; Schreiber, J.; Eisenhauer, F.; Ott, T.; Horrobin, M.; Gillesen, S. SINFONI data reduction software. *New Astron. Rev.* **2006**, *50*, 398–400. [CrossRef]
43. Dean, J.; Ghemawat, S. MapReduce: Simplified data processing on large clusters. *Commun. ACM* **2008**, *51*, 107–113. [CrossRef]
44. Thusoo, A.; Sarma, J.S.; Jain, N.; Shao, Z.; Chakka, P.; Zhang, N.; Antony, S.; Liu, H.; Murthy, R. Hive—A Petabyte Scale Data Warehouse Using Hadoop. In Proceedings of the 26th International Conference on Data Engineering (ICDE 2010), Long Beach, CA, USA, 1–6 March 2010. [CrossRef]
45. Adam, J. The Pathologies of Big Data. *Commun. ACM* **2009**, *52*, 36–44.
46. Brinkmann, B.H.; Bower, M.R.; Stengel, K.A.; Worrell, G.A.; Stead, M. Large-scale Electrophysiology: Acquisition, Compression, Encryption, and Storage of Big Data. *J. Neurosci. Methods* **2009**, *180*, 185–192. [CrossRef] [PubMed]
47. Mayer-Schönberger, V.; Cukier, K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*; Houghton Mifflin Harcourt: Boston, MA, USA, 2013; ISBN-10: 0544227751; ISBN-13/EAN: 9780544227750.
48. Ebadi, A.; Schiffauerova, A. How to become an important player in scientific collaboration networks? *J. Informetr.* **2015**, *9*, 809–825. [CrossRef]
49. Navonil, M.; Nik, B.; Simon, J.E.T.; Stelios, S. Exploring the e-science knowledgebase through co-citation analysis. *Procedia Comput. Sci.* **2013**, *19*, 586–593.
50. Callon, M.; Courtial, J.; Laville, F. Co-word analysis as a tool for describing the network of interactions between basic and technological research—The case of polymer chemistry. *Scientometrics* **1991**, *22*, 155–205. [CrossRef]