CrossMark

Discovering the interdisciplinary nature of Big Data research through social network analysis and visualization

Jiming Hu[1] · Yin Zhang[2]

Abstract Big Data is a research field involving a large number of collaborating disciplines. Based on bibliometric data downloaded from the Web of Science, this study applies various social network analysis and visualization tools to examine the structure and patterns of interdisciplinary collaborations, as well as the recently evolving overall pattern. This study presents the descriptive statistics of disciplines involved in publishing Big Data research; and network indicators of the interdisciplinary collaborations among disciplines, interdisciplinary communities, interdisciplinary networks, and changes in discipline communities over time. The findings indicate that the scope of disciplines involved in Big Data research is broad, but that the disciplinary distribution is unbalanced. The overall collaboration among disciplines tends to be concentrated in several key fields. According to the network indicators, Computer Science, Engineering, and Business and Economics are the most important contributors to Big Data research, given their position and role in the research collaboration network. Centering around a few important disciplines, all fields related to Big Data research are aggregated into communities, suggesting some related research areas, and directions for Big Data research. An ever-changing roster of related disciplines provides support, as illustrated by the evolving graph of communities.

Keywords Big Data research · Interdisciplinary collaboration · Network structure and patterns · Visualization

✉ Jiming Hu
   hujiming@whu.edu.cn

   Yin Zhang
   yzhang4@kent.edu

[1]   School of Information Management, Wuhan University, Wuhan, China

[2]   School of Library and Information Science, Kent State University, Kent, OH, USA

⚙ Springer

Introduction

In recent years, researchers from many countries and governments have paid close attention to Big Data research (Chen and Zhang 2014). While the excessive data resulting from rapid information growth is bringing great benefits to many fields such as business, science, and public administration (Savitz 2012a), it is proving difficult in its retrieval, collection, storage, filtering/classification, analysis, sharing/providing, and security (Khan et al. 2014).

As researchers studying the theories and technologies of Big Data make attempts to apply them to various fields (Chen et al. 2014), researchers, institutions, and even national governments agree that significant numbers of multidisciplinary collaborations are needed to promote understanding and development of Big Data (Emani et al. 2015). Additionally, the body of academic literature provides evidence that Big Data applications lie in many scientific disciplines, and require experts from different fields to collaborate on many complex theories, approaches, and techniques (Clarke 2016).

Disciplinarity refers to the degree of mastery over methodology and the capacity to obtain, analyze, and employ specialized knowledge (Whitley 2000). But the exchange of ideas across disciplines promotes the progress of science, and interdisciplinarity and cross-disciplinarity have been buzzwords for the last few years, which are used to describe contributions from and collaborations among several or more disciplines (Klein 1990). The prevalent tendency is for disciplinarity to be substituted by interdisciplinarity (Klein 2000; Jacobs and Frickel 2009). Interdisciplinary research is often perceived as a mark of innovation, potentially more successful at making breakthroughs and generating outcomes (Rafols and Meyer 2007). Big Data research, for example, is a kind of typical interdisciplinary research field involved in many disciplines (Emani et al. 2015).

Although Big Data research has involved a large number of disciplines, how disciplines collaborate to promote the development is still not well understood, including the structure and patterns of collaboration among disciplines. This paper aims to address the paucity of studies examining the interdisciplinary nature of Big Data research, whilst building on some previous studies that mapped and visualized the interdisciplinary collaboration of other fields, for example, Demography (Liu and Wang 2005), Cognitive Science (Leydesdorff and Goldstone 2014), and overall collaborations based on journal–journal citations (Leydesdorff et al. 2015). Additionally, this study aims to utilize the co-occurrence data between Subject Categories (SCs) related to Big Data research to discover the structure and pattern of the interdisciplinary network; its distribution and evolution over time; and the structural communities of interdisciplinary collaboration. It will then visualize these interdisciplinary networks. The results will help us explicitly understand the multidisciplinary and interdisciplinary status and development of Big Data research.

Literature review

Background of Big Data research and development

Big Data is an emerging field of practice and it is a challenge to comprehensively define Big Data in a way that is agreeable to all disciplines and in all contexts. Drawing on an extensive review of literature, Kitchin (2014) summarizes the major characteristics of Big Data as huge in volume, high in velocity, diverse in variety, exhaustive in scope, fine-

grained in resolution and uniquely indexical in identification, relational in nature, flexible in holding the traits of extensionality, and scalability. Big Data requires innovative techniques and technologies to perform its capture, curation, analysis, visualization and application (Casado and Younas 2015). Big Data research is considered to be of the top ten critical technology trends for the next five years (Savitz 2012a), as well as being a top ten strategic technology topic in 2013 (Savitz 2012b), and is considered one of the current and future frontiers of research (Agrawal and Chawla 2015). The evolving nature and status of Big Data research has been gauged via quantitative analysis of the proliferation of journal articles about Big Data and increasing industries and research approaches involved (Wamba et al. 2015), and joint efforts from academics, industries, and governments (Chen et al. 2014; Khan et al. 2014).

More and more fields are involved in addressing Big Data problems, such as scientific computation (e.g., Szalay 2011; Li et al. 2015), commerce and business (e.g., Olsson and Bull-Berg 2015; Erevelles et al. 2016), and other research fields (e.g., Goes 2014; Offroy and Duponchel 2016). Meanwhile, many industries and fields, with the development of techniques and technologies for utilizing Big Data, have already become highly data-driven, thus gaining many advantages and increasing operational efficiency (Khan et al. 2014). The innovative techniques of Big Data are also continually applied in various disciplines and fields to obtain useful outcomes (Chen et al. 2012, 2014; Singh et al. 2015).

Previous efforts in revealing and understanding the status of Big Data research

With publications on Big Data research proliferating in recent years, there have been efforts to elucidate the status of research in the field. A majority of such efforts have been devoted to qualitative reviews on Big Data research and development, including understanding fundamental concepts in Big Data research (e.g., De Mauro et al. 2014; Emani et al. 2015), exploring the background and development trends (e.g., Chen et al. 2014; Yacioob, et al. 2016), identifying related opportunities and challenges (e.g., Khan et al. 2014; Ekbia et al. 2015; Hilbert 2016), and recognizing applications and techniques (e.g., Al-Jarrah et al. 2015; Gil and Song 2016).

The status of Big Data research has been summarized on several fronts. First, Big Data has gained broad ground with the exponential, explosive increase of data in various fields through the aid of Information and Communication Technology (ICT), as well as valuable innovation and development opportunities (Chen et al. 2014; Khan et al. 2014). Second, Big Data research has been focused on theories and techniques which are considered its main directions, such as cloud computing, storage systems, and tools for Big Data mining and analysis (e.g., Chen and Zhang 2014; Emani et al. 2015). Third, the effective use of Big Data not only brings opportunities but also challenges to research, enterprises, and governments (Hilbert 2016). Big Data benefits intelligent decision-making and powerfully enhances competitive abilities (e.g., Chen and Zhang 2014). On the other hand, existing techniques or theories fail to process and analyze such immense data (Al-Jarrah et al. 2015; Wu et al. 2015), and even present vulnerabilities in areas such as privacy, security, and law (e.g., Bardi et al. 2014; Ekbia et al. 2015). The transformation and innovation of traditional theories, techniques, and approaches are key topics in the development of Big Data (e.g., Kambatla et al. 2014; Fang et al. 2015).

Disciplinary and interdisciplinary research

Interdisciplinary research (IDR), defined as the integration of disciplines within a research field (Qin et al. 1997), has become more common in scientific research (Birnbaum 1981). Although there are many terms, including inter-, multi-, trans- and cross-disciplinary; interdisciplinarity is the current, widely used term describing a property of collaborative research between, beyond, or across various disciplines, and for research spanning a variety of academic disciplines (Rafols and Meyer 2007).

Many previous studies of interdisciplinary collaboration have examined the structure and patterns of disciplines according to the research output of a field. First, researchers relied on statistical methods and social network analysis to visualize or map the relationship of co-occurrence, citations, and other bibliometric data, and then categorized the disciplines related to demography and revealed collaborative relationships (Liu and Wang 2005). Second, as noted in the aforementioned literature survey, the interdisciplinary structure of a research field has been mapped through journal citation analysis (e.g., Small 2010; Chi and Young 2013) to discover the collaboration of major disciplines and their relationships over time. Researchers studied the distribution and network of disciplines through the articles' references, to map and evaluate interdisciplinarity (e.g., Rafols and Meyer 2007, 2010). More importantly, Leydesdorff and his team also explored interdisciplinary research, as shown through cited journal maps (Leydesdorff et al. 2013, 2015). These studies provide relationship structures and patterns using a variety of interdisciplinary network maps, and even illustrate their dynamic evolution over time (Leydesdorff and Goldstone 2014).

Rationale for this study

The rapid growth in data size and scope, coupled with currently limited theories and techniques, has created a need for multidisciplinary collaboration among disciplines such as Mathematics, Computer Science, Engineering, and Social Sciences. Previous research and development has shown that Big Data research has become a multidisciplinary and interdisciplinary research field that involves a large number of disciplines. Additionally, an interdisciplinary approach has been noted and considered beneficial to advance Big Data research. For example, Computer Science, Engineering, and Statistics are three main disciplines contributing their respective approaches to Big Data models and algorithms (Fang et al. 2015).

Although previous efforts offer great insights into the status and development of Big Data research, a study revealing the specific structure and patterns of interdisciplinary collaboration in Big Data research is still lacking. To fill this gap in the literature, this study aims to map the interdisciplinary collaboration network of disciplines related to Big Data research, which could help researchers grasp the status and development of research collaboration among disciplines. Specially, this study addresses the following three research questions:

(1) What is the overall distribution and collaboration structure of related disciplines in Big Data research?
(2) What are the research communities formed from the interdisciplinary collaborations?
(3) What is the evolving tendency of interdisciplinary collaboration in Big Data research over time?

Methodology

Data collection and sample

The Web of Science (WoS) Core collection database was chosen as the source for literature related to Big Data research. During exploratory topic searches for related literature, it was observed that WoS did not include "Big Data" as its controlled vocabulary term in the Keywords Plus field, and that many records included "Big Data" in the abstract and/or author-provided keywords (DE field) simply as a general research background while the research itself was not about Big Data. Therefore, our sample searches revealed that combining title and author keywords turned out to be the most relevant indicator in identifying related research on Big Data.

The sample used in this study was retrieved from the WoS core collection and filtered using "Big Data" in both the title and author-provided keywords (DE) fields; for maximum recall, the timeframe covered the years from 1950 to 2015; and it included the document types of Article, Review, and Proceedings. The initial sample contained 1935 papers. After review, four records without SC data were excluded, leaving a final sample of 1931 papers, with the earliest published in 2004.

Methods and tools

For the purpose of the present study, we adopted the approach of using the subject categories (SCs) of publications in WoS as the basis for constructing the co-occurrence networks of disciplines, and to analyze and visualize the interdisciplinary collaboration networks (Rafols and Meyer 2010; Bjurström and Polk 2011). Specifically, the SC, used as disciplinary categories, is an accurate and simple unit of analysis to describe the disciplines involved in a research field (Rafols and Meyer 2010; Taskin and Aydinoglu 2015).

The co-occurrence methodology (Small and Griffith 1974; Coulter et al. 1998; Ding et al. 2001) serves as the basis of this study to associate different disciplines, i.e., SCs. Its effectiveness in identifying and revealing the underlying collaborative structure and patterns of terms (keywords, journals, authors, etc.) has been proven by many previous studies in other fields (e.g., Grauwin and Jensen 2011; Hu et al. 2011; Catala-Lopez et al. 2012).

First, in order to obtain the co-occurrence data representing the collaborative relationship between disciplines (co-discipline data), bibliographic data was downloaded from WoS and imported into SCI2 (Boerner 2011), which is an effective and widely used software application for network analysis and the visualization of scholarly datasets. Then a new file (.nwb) reflecting the co-discipline network was generated. In the data, nodes represent disciplines, with a corresponding number of occurrences, while relations between those disciplines are presented as links, with their number of co-occurrences. Note that a link equals two different disciplines co-occurring on at least one paper. This co-discipline data was then exported in formats to be read by Pajek software, which can better calculate network indicators and generate initial network maps.

Second, after generating the interdisciplinary network between disciplines, we used SCI2 to exclude isolated nodes, as unconnected, nonrelated SCs cannot reflect interdisciplinary research (Leydesdorff et al. 2013, 2015). The largest component of the interdisciplinary network was then extracted for further analysis, using Pajek for the calculation of network indicators (such as degree, density, centrality, etc.) (Doreian et al. 2013). Network indicators are useful for identifying the overall structure and pattern of the

research collaboration network (Yan et al. 2010), and for understanding a discipline's attributes, such as their power, stratification, ranking, and inequality in the network (Wasserman and Faust 1994).

Third, community structure, reflecting the clustering of disciplines, was detected using the Louvain algorithm (Blondel et al. 2008) in Pajek. An overall network graph of the interdisciplinary community, representing every year studied, was exported from Pajek to VOSviewer (van Eck and Waltman 2010) to conduct visualization. Additionally, network indicators of the largest network component reflect the overall status of interdisciplinary research, as well as of individual disciplines, in Big Data research. This approach also explicitly illustrates the interdisciplinary network and relations between involved disciplines. Cortext was used to visualize the evolution of individual disciplines and interdisciplinary communities, allowing for a layout of the dynamics as depicted by tubes in an alluvial model (cf. Rosvall and Bergstrom 2010). This method displayed developments within networks (Leydesdorff and Goldstone 2014).

Results

Disciplines involved in Big Data research

In this study, 109 disciplines are identified, and their statistical data is listed in Table 1. In Big Data research, the number of papers, disciplines and their co-occurrences are increasing over time; but the average number of disciplines involved with each paper only slightly varies, and generally only involves one or two disciplines.

Table 2 lists the top 39 disciplines involved in Big Data research, each with greater than ten occurrences. The leading disciplines are Computer Science, Engineering, Telecommunications, Business and Economics, Social Sciences, Information Science and Library Science, Education and Educational Research, Automation and Control Systems, Operations Research and Management Science, and Mathematics. The two largest disciplines, Computer Science and Engineering, account for 55.67% of the total occurrences of disciplines. The second-tier of disciplines, having published at least 60 papers related to Big Data research, are Telecommunications, Business and Economics, and Social Sciences. Information Science and Library Science, Education and Educational Research, Automation and Control Systems, Operations Research and Management Science, Mathematics, and Materials Science are also key disciplines. These top ten disciplines contribute 74.8% of all discipline occurrences, demonstrating an unbalanced distribution in Big Data research.

Table 1 The basic statistics of sample papers and SCs over time

Year	Number of papers	Number of discipline occurrences in papers	Number of unique disciplines	Mean disciplines in each paper
2004–2011	12	16	12	1.33
2012	75	114	20	1.52
2013	434	700	54	1.61
2014	642	977	76	1.52
2015	758	1218	91	1.59
Overall	1931	3023	109	1.57

Table 2 39 disciplines with greater than 10 occurrences involved in Big Data research

| Rank	Discipline	The number of occurrences	Rank	Discipline	The number of occurrences
1	Computer Science	1161	21	Robotics	25
2	Engineering	522	22	Energy and Fuels	20
3	Telecommunications	155	23	Remote Sensing	18
4	Business and Economics	93	24	Biotechnology and Applied Microbiology	17
5	Social Sciences	60	24-tie	Chemistry	17
6	Information Science and Library Science	49	26	Genetics and Heredity	15
7	Education and Educational Research	47	26-tie	Mathematical and Computational Biology	15
8	Automation and Control Systems	46	26-tie	Psychology	15
9	Operations Research and Management Science	43	29	Geography	14
10	Materials Science	42	29-tie	Geology	14
10-tie	Mathematics	42	29-tie	Imaging Science and Photographic Technology	14
12	Government and Law	39	32	Biochemistry and Molecular Biology	13
13	Science and Technology	36	32-tie	Neurosciences and Neurology	13
14	Health Care Sciences and Services	35	32-tie	Research and Experimental Medicine	13
15	Medical Informatics	35	35	Pharmacology and Pharmacy	12
16	Optics	34	35-tie	Sociology	12
17	Public Administration	31	35-tie	Transportation	12
18	Environmental Sciences and Ecology	28	38	Mechanics	10
19	Physics	27	38-tie	Physical Geography	10
20	Communication	25			

Network analysis of interdisciplinary collaborations

Descriptive statistics

Table 3 provides descriptive statistics about the interdisciplinary collaboration networks in Big Data research. First, the largest network components, overall and for individual years, are a high proportion of the entire interdisciplinary network; and the scale of interdisciplinary collaboration grows annually. However, the average degree and density changed little from 2013 to 2015. Second, Fig. 1 shows the evolution of network indicators. It is noted that density of interdisciplinary collaboration networks, measuring the closeness degree of connections, is very low, indicating weak collaboration between disciplines in Big Data research. The network betweenness centralization, measuring the degree of dependence on one or some nodes that could play a bridging role, is decreasing annually and the overall is the lowest. This

Table 3 Descriptive statistics of interdisciplinary collaboration networks

Year	Number of nodes	Number of lines	Average degree	Density
2004–2011	3	2	0.6667	0.0606
2012	8	8	1.3	0.0684
2013	36	54	2.2222	0.0419
2014	51	111	3.1316	0.0418
2015	61	142	3.2088	0.0357
Overall	86	231	4.3119	0.0399
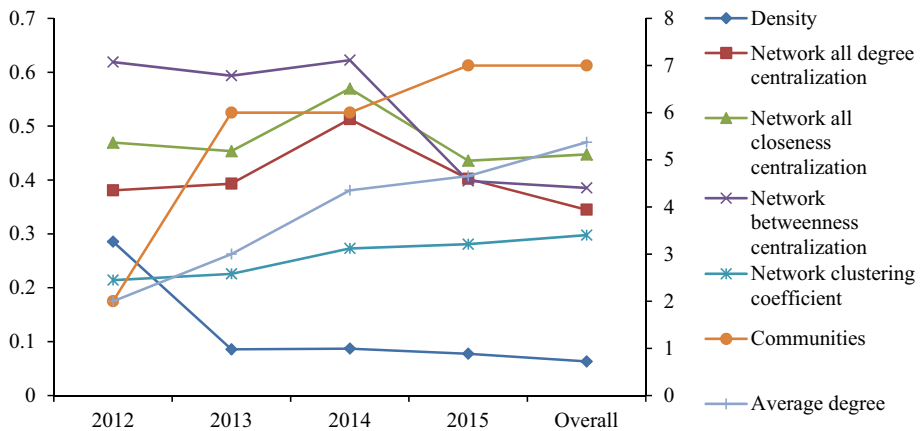
Fig. 1 The evolution of network indicators of the largest component of interdisciplinary in Big Data research networks during 2012–2015 and all years: network indicators (*left axis*) and the number of communities (*right axis*)

indicates that indirect links through a third discipline are fewer than direct ones between disciplines, and also proves weak interdisciplinary collaboration in Big Data. Therefore, betweenness centrality could be used as an indicator to measure the interdisciplinarity of an individual research field (Leydesdorff 2007). The high level of degree centralization and closeness centralization, respectively measuring the aggregation and independent degrees connecting other nodes (Wasserman and Faust 1994), is indicative of interdisciplinary research in Big Data tending to be centralized, and denoting significant differences between groups or individual disciplines (Khan et al. 2011). That is to say, the majority of disciplines are closely connected to a few powerful disciplines, and any two disciplines are relatively independent unless they are clustered into one group. Most of the interdisciplinary networks' clustering coefficients, measuring the possibility of being divided into groups, are at the 0.2–0.3 level, indicating two disciplines are more likely to collaborate and cluster into one group (Yan et al. 2010). The results of community detection echo those of the clustering coefficient, as about six or seven communities are detected in this collaboration network.

Network characteristics of individual disciplines

The network centrality of individual nodes represents the position and capacity that could dominate collaboration in the whole network (Rafols and Meyer 2007). Taking the number

of occurrences and sum of the indicators for each discipline, those important in interdisciplinary networks were selected. With high degree centrality, Computer Science, Engineering, Social Sciences, Business and Economics, and Automation and Control Systems represent the central disciplines directly connected with others. In the whole interdisciplinary network, they tend to have both a greater capacity and possibility to influence other disciplines. In a view of closeness centrality (Freeman 1979), the distance of Computer Science, Engineering, Automation and Control Systems, Business and Economics, and Mathematics to any other discipline in the interdisciplinary network is short; they are more powerful and lead distinct communities. Different from degree centrality and closeness centrality, the top ten disciplines with high betweenness centrality are more diverse but not as focused. These disciplines play the role of connecting different disciplines and communities, such as Computer Science, Engineering, Business and Economics, Information Science and Library Science, and Social Sciences. Although Big Data research is concentrated in Computer Science, Engineering, and Business and Economics, etc., these disciplines with high betweenness centrality connect more extensively with others, suggesting some shared interests with other disciplines.

Figures 2, 3, and 4 summarize the developments of the selected top five disciplines in terms of degree centrality, closeness centrality, and betweenness centrality. The development of degree centrality and closeness centrality is similar, while that of betweenness centrality has fluctuated widely. It indicates that the central disciplines in Big Data are relatively the same and are connected to many of the other disciplines. Disciplines playing the "bridge" role connecting any other two disciplines vary at different times. Taking various factors into consideration, Computer Science, Engineering, and Business and Economics prove to be the three most important disciplines in Big Data research. They are central in the interdisciplinary network and connect the entire collaboration network.

Interdisciplinary collaboration communities

Interdisciplinary collaboration communities detected in Big Data research are shown in Table 4. These results also prove the above conclusions that distinct interdisciplinary communities exist, led by a few central and important disciplines such as Computer Science, Engineering, Business and Economics, Social Sciences, Automation and Control Systems, Mathematics, Chemistry, Information Science and Library Science, and
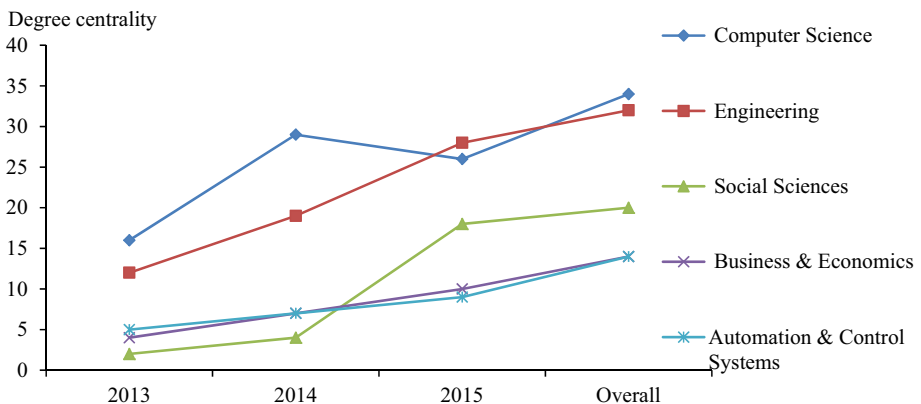

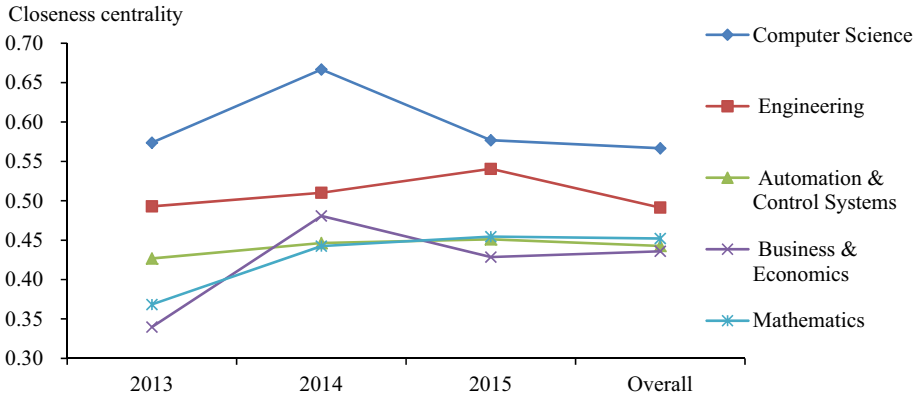**Fig. 2** Summary of the degree centrality of top five disciplines over the years

Closeness centrality

Fig. 3 Summary of the closeness centrality of top five disciplines over the years
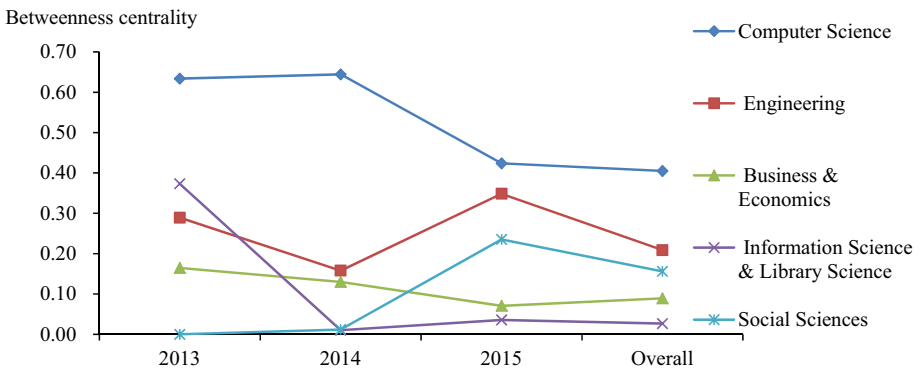
Betweenness centrality

Fig. 4 Summary of the betweenness centrality of top five disciplines over the years

Biochemistry and Molecular Biology. The number of interdisciplinary collaboration communities ranged between six and seven overall, indicating that interdisciplinary research in Big Data tends to be mature and stable.

Visualization of the interdisciplinary collaboration in Big Data research

The largest components of these networks, including distinct communities, have been visualized to display the interdisciplinary collaboration in Big Data research. These maps are displayed as Figs. 5, 6, 7, 8 and 9, and include the years from 2012 to 2015 as well as the overall interdisciplinary network. In these maps, disciplines and their relationships are shown clearly and sized proportionally, demonstrating that connections between disciplines within each community are closer than those between communities.

Several disciplines, each leading distinct community, are positioned centrally on the map. The remaining disciplines concentrate into one community due to their collaborations. In accordance with the results above, disciplines including Computer Science, Engineering, Business and Economics, and Mathematics prove central to the whole network, with a larger number of occurrences and co-occurrences. They respectively lead larger communities, with other disciplines connected to them.

Table 4 The interdisciplinary collaboration communities of Big Data research (2012–2015)

Year	Community	Number of disciplines	Representative disciplines
2012	C1-2012	6	Computer Science; Engineering; Physics; Operations Research and Management Science
	C2-2012	2	Information Science and Library Science; Business and Economics
2013	C1-2013	12	Computer Science; Engineering; Telecommunications; Optics; Operations Research and Management Science
	C2-2013	11	Business and Economics; Information Science and Library Science; Government and Law; Psychology; Education and Educational Research
	C3-2013	6	Geology; Remote Sensing; Automation and Control Systems; Imaging Science and Photographic Technology; Physical Geography
	C4-2013	3	Medical Informatics; Health Care Sciences and Services; General and Internal Medicine
	C5-2013	2	Materials Science; Mechanics
	C6-2013	2	Environmental Sciences and Ecology; Construction and Building Technology
2014	C1-2014	11	Computer Science; Engineering; Telecommunications; Automation and Control Systems; Physics
	C2-2014	17	Business and Economics; Health Care Sciences and Services; Education and Educational Research; Communication; Information Science and Library Science
	C3-2014	9	Public Administration; Government and Law; Remote Sensing; Agriculture; Sociology
	C4-2014	5	Materials Science; Energy and Fuels; Environmental Sciences and Ecology; Mechanics; Construction and Building Technology
	C5-2014	7	Science and Technology; Mathematics; Chemistry; Pharmacology and Pharmacy; Instruments and Instrumentation
	C6-2015	2	Optics; Astronomy and Astrophysics
2015	C1-2015	12	Computer Science; Engineering; Telecommunications; Operations Research and Management Science; Automation and Control Systems
	C2-2015	18	Social Sciences; Business and Economics; Education and Educational Research; Government and Law; Information Science and Library Science
	C3-2015	12	Mathematics; Medical Informatics; Biochemistry and Molecular Biology; Biotechnology and Applied Microbiology; Mathematical and Computational Biology
	C4-2015	7	Environmental Sciences and Ecology; Optics; Geography; Remote Sensing; Physical Geography
	C5-2015	5	Physics; Chemistry; Pharmacology and Pharmacy; Instruments and Instrumentation; Electrochemistry
	C6-2015	5	Cardiovascular System and Cardiology; Surgery; Neurosciences and Neurology; Transplantation; Respiratory System
	C7-2015	2	Psychology; Music

Table 4 continued

Year	Community	Number of disciplines	Representative disciplines
Overall	C1-Overall	15	Computer Science; Engineering; Telecommunications; Automation and Control Systems; Operations Research and Management Science
	C2-Overall	30	Business and Economics; Social Sciences; Information Science and Library Science; Education and Educational Research; Government and Law
	C3-Overall	19	Mathematics; Science and Technology; Physics; Chemistry; Mathematical and Computational Biology
	C4-Overall	6	Materials Science; Environmental Sciences and Ecology; Energy and Fuels; Geography; Mechanics
	C5-Overall	10	Biotechnology and Applied Microbiology; Genetics and Heredity; Research and Experimental Medicine; Oncology; Cardiovascular System and Cardiology
	C6-Overall	3	Medical Informatics; Health Care Sciences and Services; General and Internal Medicine
	C7-Overall	3	Remote Sensing; Imaging Science and Photographic Technology; Physical Geography
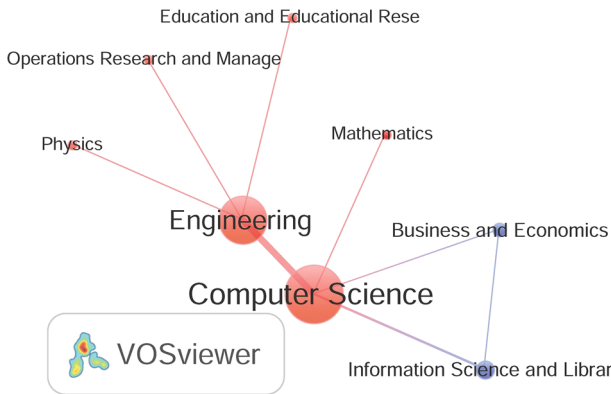
Fig. 5 Interdisciplinary collaboration communities (2012)

The scale of communities is unbalanced. Communities related to Computer Science and Business and Economics are the two largest, and could be viewed as leading fields performing Big Data research. The community related to Automation and Control Systems was independent in 2013, but aggregated into a larger community with Computer Science the following year. The community related to Medical Informatics was also independent in 2013, and then aggregated into larger communities related to Business and Economics in 2014 and Mathematics in 2015; though remaining a small independent field in Big Data research overall. The Mathematics community was also a small independent research field, only aggregating into a larger one in 2015. It is gradually growing into an important field for Big Data research. In addition, Medical Informatics, Remote Sensing, and other related disciplines are aggregated into small, independent research communities.
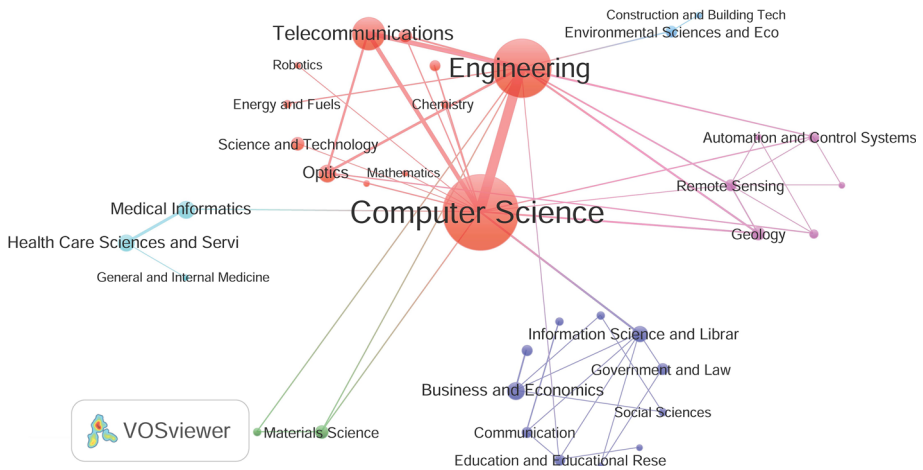
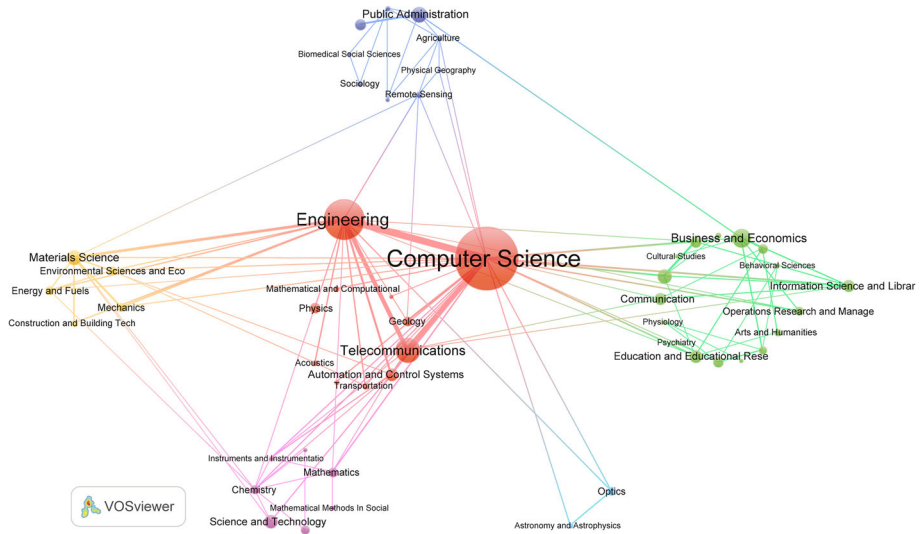**Fig. 6** Interdisciplinary collaboration communities (2013)

Fig. 7 Interdisciplinary collaboration communities (2014)

Computer Science is the fundamental discipline that could be widely applied in other disciplines to conduct large-scale computing and analyzing. In Big Data research, Computer Science stands as the main contributor, and collaborates with many other disciplines. Figure 10 shows the individual ego network of Computer Science and related collaborating disciplines. There are 34 disciplines collaborating with Computer Science. It occupies a large proportion (39.5% of 86 disciplines in the largest component) compared with others. Engineering is its primary partner, followed by Telecommunications and Automation and Control Systems.

Figure 11 shows the overall evolution of discipline communities over time. The results are similar to the earlier analysis, due to the same community-detecting algorithm.
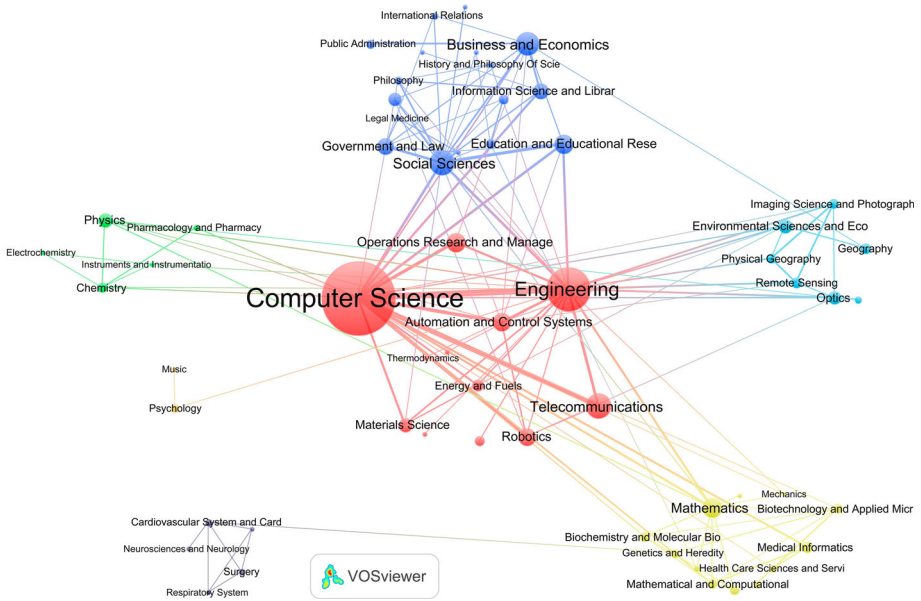
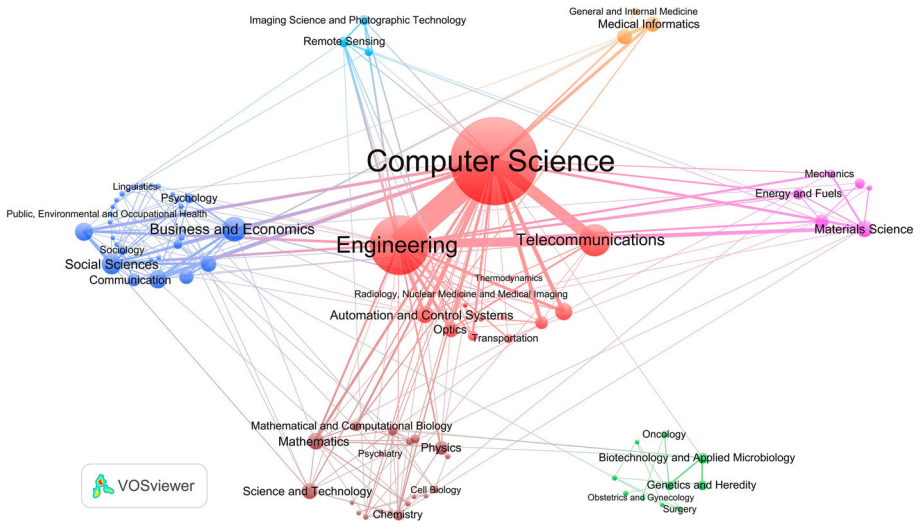**Fig. 8** Interdisciplinary collaboration communities (2015)

Fig. 9 Interdisciplinary collaboration communities (2004–2015)

Communities related to Computer Science and Engineering are shown with an increasing trend in the tubes. The community related to Automation and Control Systems was independent in 2013, and was substituted by other disciplines in 2015. Information Science and Library Science was an independent community in 2013 before aggregating into a larger community with Business and Economics. Mathematics and other related disciplines formed a relatively large community in 2015. Differing from previous results, Government
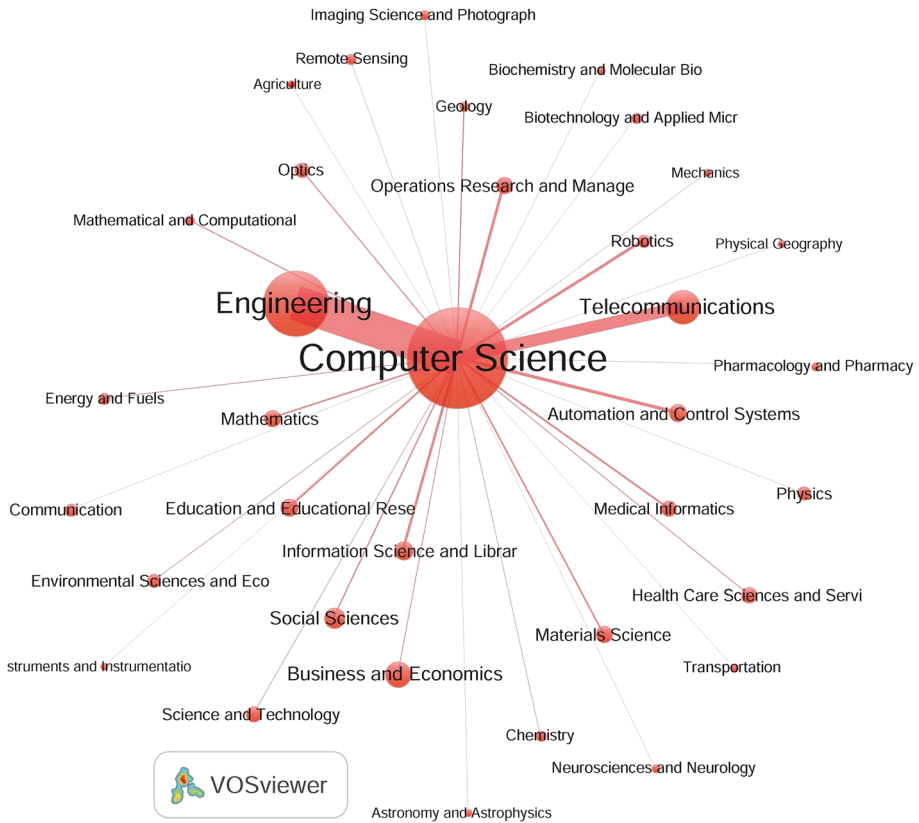
Fig. 10 The individual ego network of Computer Science and related collaborations

and Law separated from Business and Economics in 2015 to form an independent community with other related disciplines. In general, whereas the number of flows in Fig. 11 fluctuates, seven discipline communities are indicated, in agreement with the results above.

Discussion

The development of interdisciplinary collaboration in Big Data research over recent years is illustrated in this study, including basic statistics for the related disciplines. The network structure of interdisciplinary collaboration and patterns of collaboration communities are supported by the visualized maps and evolution tube, illustrating the interdisciplinary structure and patterns. These results help offer a comprehensive understanding of the interdisciplinary nature of research in Big Data, and shed light on related efforts.

First, Big Data research involves an increasing number of disciplines and has generated several research areas. The distribution of disciplines related to Big Data is uneven, and multiple levels exist. Thirty-nine disciplines, each with a statistical frequency greater than ten, make the majority of contributions to Big Data research, the two largest being Computer Science and Engineering. As shown in previous studies, Computer Science provided techniques and methods for processing large scale data sets (Chen et al. 2014)
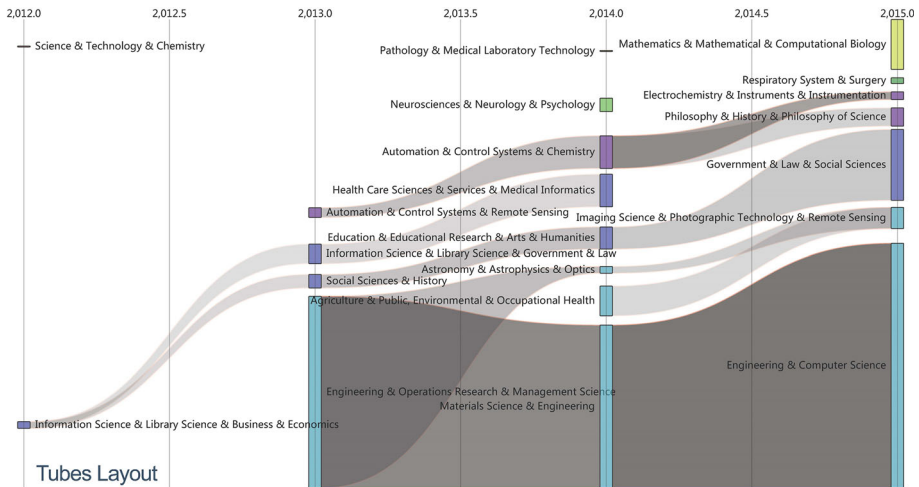
Fig. 11 Evolution of disciplinary communities over time

that are widely used in such fields as Engineering, Medicine, and Business (Khan et al. 2014). Therefore, Computer Science is the fundamental discipline in Big Data research, and supports and connects to other disciplines.

Second, the structure and patterns of collaborations among disciplines coalesced in 2013, rapidly maturing and stabilizing by 2014. In 2015, communities defined by collaborations clearly represent sub-fields or directions in Big Data research. The leading disciplines in each community are generally equivalent to the leading fields in Big Data research, such as Computer Science, Business and Economics, Mathematics, and Materials Science. Overall, the collaborations among individual disciplines across Big Data research are not as intensive as those within communities. The close connections among these disciplines aggregated into communities indicate how they support and supplement each other.

Third, we also find that the community related to Computer Science and Engineering is always central to the collaboration networks, connecting with and supporting other independent communities. For example, the capture and storage of large-scale data and analysis for decision support systems depends on the theories, methods, algorithms, and tools of Computer Science and Engineering. This community also plays an important bridging role connecting communities or research fields. Currently, along with the large proportion of Big Data research into fundamental theories, there is also widespread application research. In fact, 403 articles, 20.9% of the total sample, were results of 1342 projects as indicated in the publications. Projects were usually collaboratively conducted by researchers from multiple, related disciplines. This large number of interdisciplinary projects relating to Big Data research help facilitate interdisciplinary collaborations and generate intensive discipline communities.

Finally, it is also worth pointing out that the scope and nature of Big Data research contribute to a better understanding of Big Data practices and the definition of Big Data. Findings of this study may enhance the definition from more perspectives, especially taking into consideration of its interdisciplinary nature. For example, Big Data research

involves both Engineering Science and Social Sciences, and it should be defined considering the full range of various disciplines involved.

Conclusions and future research

The findings of this study provide a clear, comprehensive understanding of interdisciplinary collaboration in Big Data research. As the research develops, Big Data will continue to expand in scope and to address new problems, and the intellectual structure and collaboration patterns are expected to be accordingly transformed and constructed. The conceptual framework and methodology used in this study could be replicated for such future examination of trends in Big Data research, and any other research area, for revealing interdisciplinary collaborations.

The results of this study have shown the interdisciplinary and collaborative nature in Big Data research, which is propelled by wide applications in a broad range of disciplines (De Mauro et al. 2014). While we can use network indicators to understand the degree of interdisciplinary collaboration in Big Data research, measuring and calculating the degree of interdisciplinary collaboration could be the next step to further such understanding, built upon the current study. Our future research plan includes exploring approaches to measure the degree of interdisciplinary collaboration, and examining its effect on research productivity and scholarly impact.

Finally, future studies about Big Data research may be conducted in several areas. Comprehensive studies on the interdisciplinary collaboration of Big Data could be furthered from other perspectives such as social sciences, political sciences, and related policies. The research themes behind interdisciplinary collaborations should be analyzed, to increase understanding of cross-research among disciplines. Additionally, future studies may further trace the dynamic developments of interdisciplinary collaborations. The transformations of interdisciplinary collaborations also affect the frontiers and focus of the research. There is a place for qualitative approaches coupled with quantitative methods to understand the underlining issues and themes. Without qualitative approaches it is impossible to understand the impetus for collaborations, and is difficult to know the benefits of such collaborations.

References

Agrawal, D., & Chawla, S. (2015). The Big Data landscape: Hurdles and opportunities. In W. Chu, S. Kikuchi & S. Bhalla (Eds.), *Databases in networked information systems* (pp. 1–11).

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., & Taha, K. (2015). Efficient machine learning for Big Data: A review. *Big Data Research, 2*(3), 87–93.

Bardi, M., Zhou, X., Li, S., & Lin, F. (2014). Big Data security and privacy: A review. *China Communications, 11*(2), 135–145.

Birnbaum, P. H. (1981). Academic interdisciplinary research: Characteristics of successful projects. *Journal of the Society of Research Administrators, 13,* 5–16.

Bjurström, A., & Polk, M. (2011). Climate change and interdisciplinarity: A co-citation analysis of IPCC third assessment report. *Scientometrics, 87,* 525–550.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 8*(10), P10008.

Boerner, K. (2011). Plug-and-play macroscopes. *Communications of the ACM, 54*(3), 60–69.

Casado, R., & Younas, M. (2015). Emerging trends and technologies in Big Data processing. *Concurrency and Computation-Practice and Experience, 27*(8), 2078–2091.

Catala-Lopez, F., Alonso-Arroyo, A., Aleixandre-Benavent, R., Ridao, M., Bolanos, M., Garcia-Altes, A., & Peiró, S. (2012). Coauthorship and institutional collaborations on cost-effectiveness analyses: A systematic network analysis. *Plos One, 7*(5), e38012. doi:10.1371/journal.pone.0038012.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From Big Data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

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

Chen, C. L. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275,* 314–347.

Chi, R. B., & Young, J. (2013). The interdisciplinary structure of research on intercultural relations: A co-citation network analysis study. *Scientometrics, 96*(1), 147–171.

Clarke, R. (2016). Big data, big risks. *Information Systems Journal, 26*(1), 77–90.

Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science, 49*(13), 1206–1223.

De Mauro, A., Greco, M., & Grimaldi, M. (2014). What is Big Data? A consensual definition and a review of key research topics. *AIP Conference Proceedings, 1644*(1), 97–104.

Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing and Management, 37*(6), 817–842.

Doreian, P., Lloyd, P., & Mrvar, A. (2013). Partitioning large signed two-mode networks: Problems and prospects. *Social Networks, 35*(2), 178–203.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., et al. (2015). Big Data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology, 66*(8), 1523–1545.

Emani, C. K., Cullot, N., & Nicolle, C. (2015). Understandable Big Data: A survey. *Computer Science Review, 17,* 70–81.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research, 69*(2), 897–904.

Fang, H., Zhang, Z., Wang, C. J., Daneshmand, M., Wang, C., & Wang, H. (2015). A survey of Big Data research. *IEEE Network, 29*(5), 6–9.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215–239.

Gil, D., & Song, I. (2016). Modeling and Management of Big Data: Challenges and opportunities. *Future Generation Computer Systems-The International Journal of escience, 63,* 96–99.

Goes, P. B. (2014). Big Data and IS research. *MIS Quarterly, 38*(3), III–VIII.

Grauwin, S., & Jensen, P. (2011). Mapping scientific institutions. *Scientometrics, 89*(3), 943–954.

Hilbert, M. (2016). Big Data for development: A review of promises and challenges. *Development Policy Review, 34*(1), 135–174.

Hu, C.-P., Hu, J.-M., Gao, Y., & Zhang, Y.-K. (2011). A journal co-citation analysis of library and information science in China. *Scientometrics, 86*(3), 657–670.

Jacobs, J. A., & Frickel, S. (2009). Interdisciplinarity: A critical assessment. *Annual Review of Sociology, 35,* 43–65.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in Big Data analytics. *Journal of Parallel and Distributed Computing, 74*(7), 2561–2573.

Khan, G. F., Moon, J., & Park, H. W. (2011). Network of the core: Mapping and visualizing the core of scientific domains. *Scientometrics, 89*(3), 759–779.

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, W. K. M., Alam, M., et al. (2014). Big data: Survey, technologies, opportunities, and challenges. *The Scientific World Journal.* doi:10.1155/2014/712826.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data and Society, 1*(1), 1–12. doi:10.1177/2053951714528481.

Klein, J. T. (1990). *Interdisciplinarity/history, theory, and practice.* Detroit: Wayne State University Press.

Klein, J. T. (2000). Interdisciplinarity and complexity: An evolving relationship. *Emergence: Complexity and Organization, 6*(1-2), 2–10.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology, 58*(9), 1303–1319.

Leydesdorff, L., de Moya-Anegon, F., & Guerrero-Bote, V. P. (2015). Journal maps, interactive overlays, and the measurement of interdisciplinarity on the basis of scopus data (1996–2012). *Journal of the Association for Information Science and Technology, 66*(5), 1001–1016.

Leydesdorff, L., & Goldstone, R. L. (2014). Interdisciplinarity at the journal and specialty level: The changing knowledge bases of the journal Cognitive Science. *Journal of the Association for Information Science and Technology, 65*(1), 164–177.

Leydesdorff, L., Rafols, I., & Chen, C. (2013). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal-journal citations. *Journal of the American Society for Information Science and Technology, 64*(12), 2573–2586.

Li, K., Wu, H., & Li, Z. (2015). Big data cloud and the frontier of computer science and technology. *Concurrency and Computation-Practice and Experience, 27*(18), 5719–5721.

Liu, Z., & Wang, C. Z. (2005). Mapping interdisciplinarity in demography: A journal network analysis. *Journal of Information Science, 31*(4), 308–316.

Offroy, M., & Duponchel, L. (2016). Topological data analysis: A promising Big Data exploration tool in biology, analytical chemistry and physical chemistry. *Analytica Chimica Acta, 910,* 1–11.

Olsson, N. O. E., & Bull-Berg, H. (2015). Use of Big Data in project evaluations. *International Journal of Managing Projects in Business, 8*(3), 491–512.

Qin, J., Lancaster, F. W., & Allen, B. (1997). Types and levels of collaboration in interdisciplinary research in the sciences. *Journal of the American Society for Information Science, 48*(10), 893–916.

Rafols, I., & Meyer, M. (2007). Diversity measures and network centralities as indicators of interdisciplinarity: Case studies in bionanoscience. In *Proceedings of ISSI 2007: 11th international conference of the international society for scientometrics and informetrics* (Vols. I and II, pp. 631–642).

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics, 82*(2), 263–287.

Rosvall, M., & Bergstrom, C. T. (2010). Mapping change in large networks. *PLoS ONE, 5*(1), e8694.

Savitz, E. (2012a, October 22). Gartner: 10 Critical tech trends for the next five years. Forbes. Retrieved from http://www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years.

Savitz, E. (2012b, October 23). Gartner: Top 10 strategic technology trends for 2013. Forbes. Retrieved from http://www.forbes.com/sites/ericsavitz/2012/10/23/gartner-top-10-strategic-technology-trends-for-2013.

Singh, V. K., Banshal, S. K., Singhal, K., & Uddin, A. (2015). Scientometric mapping of research on 'Big Data'. *Scientometrics, 105*(2), 727–741.

Small, H. (2010). Maps of science as interdisciplinary discourse: Co-citation contexts and the role of analogy. *Scientometrics, 83*(3), 835–849.

Small, H., & Griffith, B. C. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science Studies, 4*(1), 17–40.

Szalay, A. S. (2011). Extreme Data-Intensive Scientific Computing. *Computing in Science and Engineering, 13*(6), 34–41.

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

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'Big Data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics, 165,* 234–246.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Whitley, R. (2000). *The intellectual and social organization of the sciences* (2nd ed.). Oxford: Oxford University Press.

Wu, L., Yuan, L., & You, J. (2015). Survey of large-scale data management systems for Big Data applications. *Journal of Computer Science and Technology, 30*(1), 163–183.

Yacioob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., et al. (2016). Big data: From beginning to future. *International Journal of Information Management, 36*(6), 1231–1247.

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