Measuring the interdisciplinarity of Big Data research: a longitudinal study

Jiming Hu
Department of Information Management, Wuhan University, Wuhan, China, and
Yin Zhang
Department of Library and Information Science, Kent State University, Kent, Ohio, USA

Received 29 December 2016
Revised 8 May 2017
30 September 2017
10 January 2018
Accepted 8 February 2018

Abstract

Purpose – The purpose of this paper is to measure the degree of interdisciplinary collaboration in Big Data research based on the co-occurrences of subject categories using Stirling's diversity index and specialization index.

Design/methodology/approach – Interdisciplinarity was measured utilizing the descriptive statistics of disciplines, network indicators showing relationships between disciplines and within individual disciplines, interdisciplinary communities, Stirling's diversity index and specialization index, and a strategic diagram revealing the development status and trends of discipline communities.

Findings – Comprehensively considering all results, the degree of interdisciplinarity of Big Data research is increasing over time, particularly, after 2013. There is a high level of interdisciplinarity in Big Data research involving a large number of disciplines, but it is unbalanced in distribution. The interdisciplinary collaborations are not intensive on the whole; most disciplines are aggregated into a few distinct communities with computer science, business and economics, mathematics, and biotechnology and applied microbiology as the core. Four major discipline communities in Big Data research represent different directions with different development statuses and trends. Community 1, with computer science as the core, is the most mature and central to the whole interdisciplinary network. Accounting for all network indicators, computer science, engineering, business and economics, social sciences, and mathematics are the most important disciplines in Big Data research.

Originality/value – This study deepens our understanding of the degree and trend of interdisciplinary collaboration in Big Data research through a longitudinal study and quantitative measures based on two indexes. It has practical implications to study and reveal the interdisciplinary phenomenon and characteristics of related developments of a specific research area, or to conduct comparative studies between different research areas.

Keywords Network analysis, Indicators, Big Data research, Measures, Interdisciplinarity

Paper type Research paper

Introduction

In the information age, Big Data holds great value to research and industry (Fang *et al.*, 2015), and also generates numerous opportunities and challenges (Marx, 2013). These challenges require a vast variety of theories, methods, techniques, and policy to address problems in capture, storage, curation, and analysis (Chen and Zhang, 2014). Big Data optimizes various applications in a large number of fields, including industry, agriculture, business and economics, traffic, transportation, medical care, families, law, public administration, etc. (Savitz, 2012; Bardi *et al.*, 2014; Ekbia *et al.*, 2015).

This study is supported by Ministry of Education of China (MOE) World-class Discipline "Library, Information, and Data Science," National Social Science Key Fund of China (NSSFC) (No. 15ZDC025), China Postdoctoral Science Foundation Special Funded Project (No. 2016T90736), National Natural Science Foundation of China Funded Project (No. 71303178) and Kent State University 2014 Postdoctoral Program for the Smart Big Data project.

Online Information Review

1468-4527
DOI 10.1108/OIR-12-2016-0361

Big Data, as an emerging field of research and practice, involves many disciplines and contexts. It is challenging to define Big Data given its evolving and interdisciplinary nature. Overall, major characteristics of Big Data include being huge in volume, high in velocity, diverse in variety, exhaustive in scope, fine-grained in resolution, uniquely indexical in identification, relational in nature, flexible in holding the traits of extensionality, and scalability (Kitchin, 2014). Furthermore, Big Data requires innovative techniques and technologies to perform.

Big Data research is a multidisciplinary and interdisciplinary research (IDR) field that requires collaborations from a great diversity of disciplines (Chen *et al.*, 2014; Hilbert, 2016). As reflected in retrieved articles related to Big Data research from the Web of Science (WoS) Core Collection, they span a large number of disciplines as represented by the subject categories (SCs) assigned by WoS; and the number of disciplines is increasing over time. More importantly, a majority of these papers have two or more SCs. It shows that Big Data research integrates different sciences and technologies into one field in a way that is characterized as interdisciplinary (Gandomi and Haider, 2015; Singh *et al.*, 2015). All of these observations provide evidence that Big Data research is regarded as an IDR (Wagner *et al.*, 2011) field. However, the degree of IDR in Big Data, named as interdisciplinarity (Tomov and Mutafov, 1996; Morillo *et al.*, 2003; Leydesdorff and Rafols, 2011) has not yet been studied or measured. It is important to study interdisciplinarity in Big Data research to understand the development of collaboration between disciplines.

This study aims to address the lack of study on revealing interdisciplinarity in Big Data research, utilizing SCs as the unit of analysis to achieve the following research purposes. First, using social network analysis, based on disciplines' co-occurrence data, this paper attempts to discover the structure and patterns of interdisciplinary collaboration, and to detect collaboration communities and trends over time. Second, relying on existing measurement indexes of interdisciplinarity (e.g. Stirling, 2007; Porter *et al.*, 2007; Rafols and Meyer, 2010, Leydesdorff and Goldstone, 2014), this paper utilizes Stirling's diversity index (Stirling, 2007) and specialization index (Porter *et al.*, 2007, 2008; Porter and Rafols, 2009) to ascertain the interdisciplinarity of Big Data research, revealing the degree of interdisciplinarity, supplemented by network indicators to describe status and trends.

Literature review

Background and status of Big Data research and development

Big Data is a research frontier (Manyika *et al.*, 2011) with significant growth in research articles in recent years (Akoka *et al.*, 2015). It is revolutionizing business, scientific research, and public administration (Chen and Zhang, 2014), with researchers from a variety of disciplines paying attention to the ability of Big Data to accelerate their research and applications (Wu *et al.*, 2015). The focus of most Big Data research is mainly on data generation, data acquisition, data storage, and data analysis (Chen *et al.*, 2014). The explosive growth rate of large data generates numerous challenges regarding technological innovation (Al-Jarrah *et al.*, 2015; Acharjya and Ahmed, 2016), security and management (Liang *et al.*, 2015), and even society and ethics (Smith *et al.*, 2015; Mittelstadt and Floridi, 2016). More importantly, Big Data also brings new opportunities (Michael and Miller, 2013) for discovering new values in all kinds of fields (Emani *et al.*, 2015) and making valuable and accurate predictions and decisions (Chen *et al.*, 2012; Suresh, 2016).

Because Big Data research is interdisciplinary, researchers with different disciplinary backgrounds share their respective approaches to solve issues and promote applications (Fang *et al.*, 2015). Computer science, engineering, and mathematics provide theories, methods and tools to capture and analyze the large-scale data in almost all fields, and make valuable contributions in other disciplines or fields, such as social science (e.g. Olmedilla *et al.*, 2016), business and economics (e.g. Einav and Levin, 2014), and health care

(e.g. Taglang and Jackson, 2016). For example, a bibliometric analysis of health care Big Data showed that collaborations amongst authors with various disciplinary backgrounds are numerous (Gu *et al.*, 2017). In sum, collaboration amongst disciplines has been recognized as a main research pattern in Big Data research.

Interdisciplinary research and interdisciplinarity

The phenomenon that researchers collaborate with others from different disciplines and fields is described as IDR (Morillo *et al.*, 2003; Karlovcec and Mladenic, 2015). IDR is defined as the integration of information, techniques, concepts, and/or theories from two or more disciplines (Rafols and Meyer, 2010), and is increasingly recognized as the solution to today's challenging scientific and societal problems (Cassi *et al.*, 2014). Given that IDR is the key to accelerate scientific discoveries and solve significant issues (Wang *et al.*, 2015), it is important to empirically investigate the status and patterns of IDR, and to analyze its consequences and impacts.

IDR is often perceived as a mark of innovation; and a high degree of interdisciplinarity is seen as more successful in achieving breakthroughs and positive outcomes (Morillo *et al.*, 2003). Interdisciplinarity has been widely used to mean research spanning a variety of academic disciplines (Rafols and Meyer, 2010). Various indicators, such as Stirling's diversity index, integration and specialization, Brillouin's Index, and network centrality, are used to measure the interdisciplinary diversity involved in research at various levels (e.g. paper, author, journal, institution, and research field) (Stirling, 2007; Porter *et al.*, 2007; Leydesdorff and Rafols, 2011; Huang and Chang, 2011; Wagner *et al.*, 2011).

Bibliometric studies on interdisciplinarity

Bibliometric studies have endeavored to calculate and measure the degree of interdisciplinarity using different units of analysis. One such unit, citations, indicate the drawing and integrating of knowledge from other resources, often in other disciplines (Wang *et al.*, 2015). Previous studies have revealed that direct and indirect citations for IDR is the most common technique in bibliometric methods (e.g. Rafols and Meyer, 2010; Huang and Chang, 2011). Co-authorship is another indicator, revealing interdisciplinary collaboration from different disciplines (Porter and Rafols, 2009); measures of co-authorship could also examine the degree of interdisciplinarity for a researcher, a journal, and even a county, etc. (Schummer, 2004; Porter *et al.*, 2008; Luzar *et al.*, 2014; Karlovcec and Mladenic, 2015). Journals are another unit of analysis, categorized into disciplines in the journal citation report; the citation relationship among journals also represents collaboration among disciplines (Morillo *et al.*, 2003). The interdisciplinarity of one research field or a journal has been examined through the assignation of journals to more than one discipline on the basis of journal–journal citations (e.g. Leydesdorff, 2007; Leydesdorff and Rafols, 2011).

Other indicators or approaches measuring interdisciplinarity have been widely used (Wagner *et al.*, 2011); but there has been no agreement on a single best indicator (Stopar *et al.*, 2016). In general, there are two main approaches or indicators that are widely approved and used. First, Stirling's diversity index is a general quantitative non-parametric heuristic and allows for the systematic exploration of interdisciplinary diversity under different perspectives or units of analysis (Stirling, 2007). Unlike other indexes, it overcomes at least partially the issue of an arbitrary choice of a predefined categorization (Cassi *et al.*, 2014). Therefore, Stirling's diversity index is more systematic or robust for measuring interdisciplinarity than other indexes (Stirling, 2007; Leydesdorff and Rafols, 2011), such as Shannon's index (Rafols and Meyer, 2010; Karlovcec and Mladenic, 2015), or TF-IDF (Xu *et al.*, 2016). Second, researchers have used social network analysis to calculate measurements of interdisciplinarity, such as network coherence or a similar network of publications (Rafols and Meyer, 2010), degree centrality of a co-occurrence network among

disciplines, and betweenness centrality of the citation network among journals (Leydesdorff, 2007; Leydesdorff and Rafols, 2011). Network indicators demonstrate the collaboration among disciplines that could describe the interdisciplinarity to some extent, but they are usually supplemental to the interdisciplinary indicators mentioned above.

Another indicator specific to WoS, and a unit of analysis in this study, are the assignation of SCs, which define and delimit disciplines (Morillo *et al.*, 2003). Porter *et al.* (2008) developed and tested measures of the integration index and specialization index based on SCs, and achieved effective results for the measurement of interdisciplinarity. Furthermore, Porter and Rafols (2009) performed interdisciplinary computation on six research fields using SCs. Other previous studies measuring interdisciplinarity have been conducted on the basis of relationship among SCs (Karlovcec and Mladenic, 2015; Wang *et al.*, 2015). Additionally, SCs were directly used for the computation of interdisciplinarity, for example, references linked to SCs (Yegros-Yegros *et al.*, 2015) and co-occurrence of SCs in papers (Porter *et al.*, 2007).

Rationale for this study

Previous studies have shown Big Data research to be interdisciplinary in nature. However, there has been a lack of examination of the interdisciplinary measurement of Big Data research in general, and from bibliometric and quantitative perspectives in particular. Built upon bibliometric methods for examining interdisciplinarity, this study aims to examine quantitatively the interdisciplinary degree in Big Data research, to reveal its status and development trends over time, and to provide a panoramic view of collaboration relationships between disciplines. The findings will contribute to a more comprehensive understanding of the interdisciplinarity of Big Data research and related trends.

Methodology

Data collection

Articles for the study were retrieved from WoS Core Collection using "big data" or similar phrases (e.g. large data, mass data) as the search term in the "Title" field and covering all years as of January 1, 2016. The search term was treated as a phrase to obtain more accurate and relevant search results. In formulating this search strategy, early test searches used "big data" as a subject term in other fields such as the "Topic" field a strategy used by other studies such as Huang *et al.* (2015). While such searches resulted in many more papers, after manually reviewing the search results, it was found that a large number of papers treated big data as a research background of another research topic rather than the foci of the research itself. To obtain more accurate and relevant search results, the final search strategy for the study was using "big data" as a phrase in "Title" field and applying filters to limit types of results to Article, Review, and Proceedings.

The search strategy resulted in 1,935 items. After removing six without SCs, 1,929 article records comprised the final sample of this study. The basic statistics of the sample are summarized in Table I. Data from 2004 to 2011 are merged because very few papers were related to Big Data in that period. In addition, some irregular SCs are substituted by its parent SCs, for example, "Social Sciences – Other Topics" is substituted by "Social Sciences."

Measures

As summarized in the literature review, SCs in WoS have been used, and proven reliable in previous research, to examine and serve as a data point in measuring and calculating the degree of interdisciplinarity of an area (Porter *et al.*, 2008; Karlovcec and Mladenic, 2015; Wang *et al.*, 2015). Following this approach, SCs associated with each paper, and their relationship of co-occurrence in papers, served as the basis of analysis for this study.

As previously mentioned, Stirling's diversity index and specialization index, as validated methods, are applied to measure the interdisciplinary degree of Big Data research in this study. Stirling's diversity index is a more scientific and effective way to measure the interdisciplinarity of Big Data research compared to other indexes. In addition, specialization index measures the concentralized degree of collaborations among disciplines, and complements Stirling's Diversity Index. Therefore, we employed both Stirling's diversity index and specialization index to measure the interdisciplinarity of Big Data research from the perspective of the diversity and specialization of involved disciplines. Stirling's diversity index incorporates various components that could examine "variety", "balance", and "disparity" together to measure the extent of diversity. Because it is a classic index for diversity that could account for the distances or similarities between SCs, it is more effective than others (Rafols and Meyer, 2010). On the other hand, specialization index, developed by Leahey (2006), is a trusted measure of the extent of focus on one research field. In this study, a high specialization index score indicates low collaboration among disciplines, and too little interdisciplinarity (Porter *et al.*, 2007). These two indexes helped us obtain an effective and accurate view of the interdisciplinarity of Big Data research.

Data processing and analysis

In order to utilize the measurement indexes mentioned above, we constructed the network data of SCs to represent the interdisciplinary collaboration relationship among disciplines. First, the bibliometric records of papers related to Big Data research were imported into SCI2 (Boerner, 2011) to extract SCs in each paper and generate the co-occurrence network among disciplines; this was named the network data file. In the network data file, nodes represent disciplines with a corresponding number of occurrences, and links represent relations among disciplines with their number of co-occurrences.

Second, we used SCI2 to exclude isolated nodes because isolated or nonrelated SCs cannot reflect interdisciplinary relations, and then generated the largest component of the interdisciplinary network. The largest component was used to calculate the network indicators by Pajek (Doreian *et al.*, 2013) such as degree, density, and clustering coefficient, and to detect the community concentrated with highly related disciplines by Louvain algorithm (Blondel *et al.*, 2008). Network measures such as centrality and density are two important indicators about characteristics of a network (Callon *et al.*, 1991). Centrality measures the degree of interaction of a node or a network with other nodes or networks. It equals to the importance of a discipline or a collaboration community in the entire research field. Density measures the internal strength of network, and indicates the degree of maturity and development of a collaboration community.

Third, the network data file was transformed into the co-occurrence matrix used to calculate the similarity between disciplines according to their co-occurrence relationship. Similarity or dissimilarity is an important part of the equation of Stirling's diversity index.

Year	Number of papers	Number of papers with two or more SCs (%)	Total number of SCs	Unique number of SCs	Average SCs in each paper
2004–2011	12	4 (33)	16	12	1.33
2012	75	37 (49)	114	20	1.52
2013	432	236 (55)	700	54	1.61
2014	643	254 (40)	977	76	1.52
2015	773	338 (44)	1,218	91	1.58
Total	1,935	869 (45)	3,023	109	1.56

Table I.
The basic statistics of papers and SCs related to Big Data research over time

We combined Stirling's diversity index and specialization index to measure the degree of interdisciplinarity in Big Data research. Stirling's diversity index, starting from a flexible general heuristic, is a more straightforward and relevant approach to measure the diversity of collaboration among disciplines (Stirling, 2007). We calculated Stirling's Diversity Index using the following equation:

$$D = \sum_{ij} p_i \cdot p_j \cdot d_{ij}, \tag{1}$$

$$d_{ij} = 1 - s_{ij}, \tag{2}$$

$$\cos(x, y) = \frac{\sum_i (x_i y_i)}{\sqrt{\left(\sum_i x_i^2\right)\left(\sum_i y_i^2\right)}}, \tag{3}$$

the values p_i and p_j are proportional representations of disciplines i and j in the interdisciplinary collaboration data; d_{ij} is the degree of difference (disparity) attributed to disciplines i and j, and derived from Salton's cosine similarity s_{ij} (Equations (2) and (3)) (Salton and McGill, 1983). It is considered one of the most common measures of similarity, and has been employed in previous studies (e.g. Rafols and Meyer, 2010). As mentioned above, in order to calculate Stirling's diversity index, a similarity matrix of SCs must be constructed according to the co-occurrence matrix; and then the values of cosine similarity are substituted into Equation (1) for final calculation.

As Porter *et al.* (2007) pointed out, specialization index (Equation (4) is not the antithesis of diversity, but a potentially orthogonal dimension. That is to say, specialization index represents the degree of concentration of disciplines in Big Data research:

$$S = \frac{\sum_i \left(f_i^2\right)}{\left(\sum_i f_i\right)^2}, \tag{4}$$

f_i is the number of occurrences of discipline i. According the study by Rafols and Meyer (2010), both of Stirling's diversity index and specialization index are concerned with measuring "diversity." But as calculated here, specialization index is the inverse of diversity (i.e. 1/diversity) (Porter *et al.*, 2008).

Finally, network indicators of the co-occurrence network among disciplines are calculated to understand the development status and trends in Big Data research. Specifically, a quadrant diagram using Stirling's diversity index and specialization index was drawn to show the interdisciplinarity of each year in Big Data research in order to display its development trends. Also, the strategic diagram (Stegmann and Grohmann, 2003), a one quadrant diagram using centrality and density, was also drawn to show major discipline communities, and to reveal their interdisciplinary statuses and trends. That is, the higher the centrality is for a discipline community, the more central the community is in the whole network; and the higher the density is for a discipline community, the more maturity or potential the community has.

Results

Disciplines involved in Big Data research

In this study, 109 disciplines are identified, and Table II lists the top 39 disciplines occurring greater than ten times in Big Data research. In general, the number of disciplines involved in Big Data research is increasing over time, and the top ten disciplines, 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, as the most important ones, account for 73.37 percent of total occurrences. Computer science and engineering, the two largest disciplines in Big Data research, contribute 55.67 percent of the total occurrences, indicating the unbalanced distribution of disciplines in Big Data research.

Social network analysis of interdisciplinary collaborations among disciplines

Table III provides the descriptive statistics for the largest component, which shows that the scale of interdisciplinary collaboration in Big Data research is generally increasing over time. But the intensity of such collaboration is weak due to low density, indicating a weak collaboration between disciplines in Big Data. The values of degree centralization and closeness centralization are all relatively high, indicating that interdisciplinary collaboration among disciplines tends to be around a few important disciplines at the center. It is also consistent with the results of clustering coefficients, suggesting most disciplines are connected to a few core ones, and cluster into some distinct collaboration communities.

No.	Discipline	The number of occurrences	No.	Discipline	The number of occurrences
1	Computer science	1,161	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	25	Chemistry	17
6	Information science and library science	49	26	Genetics and heredity	15
7	Education and educational research	47	27	Mathematical and computational biology	15
8	Automation and control systems	46	28	Psychology	15
9	Operations research and management science	43	29	Geography	14
10	Materials science	42	30	Geology	14
11	Mathematics	42	31	Imaging science and photographic technology	14
12	Government and law	39	32	Biochemistry and molecular biology	13
13	Science and technology	36	33	Neurosciences and neurology	13
14	Health care sciences and services	35	34	Research and experimental medicine	13
15	Medical informatics	35	35	Pharmacology and pharmacy	12
16	Optics	34	36	Sociology	12
17	Public administration	31	37	Transportation	12
18	Environmental sciences and ecology	28	38	Mechanics	10
19	Physics	27	39	Physical geography	10
20	Communication	25			

Table II.
39 disciplines with an occurrence greater than ten involved in Big Data research

The network betweenness centralization, as an indicator of interdisciplinarity (Leydesdorff, 2007), changed much over time, with its lowest point at all years combined. This indicates that indirect links through a third discipline are fewer than the direct ones between disciplines. It also provides evidence that disciplines are more inclined to connect to others belonging to the same community.

Network indicators of individual disciplines represent their position and capacity to dominate collaboration in the interdisciplinary network (Freeman, 1979). The top ten disciplines with major network centrality indicators (degree, closeness, and betweenness) are listed in Table IV. First, disciplines with high degree centrality are the central, such as computer science, engineering, social sciences, mathematics, business and economics, and automation and control systems. These disciplines are directly and tightly connected with others, and might have a greater capacity and possibility to influence others. A high closeness centrality of one discipline represents the short distance between it and others, such as computer science, engineering, social sciences, automation and control systems, business and economics, and mathematics; they may be more integrated and would lead distinct communities. Values of betweenness centrality for the top 10 disciplines are all low except for computer science, indicating that the connectivity of this interdisciplinary network is

Table III. Descriptive statistics for the largest component of disciplines' co-occurrence network

Year	Number of nodes	Number of lines	Average degree	Network all centralization				Clustering coefficient
				Density	Degree	Closeness	Betweenness	
2004–2011	3	2	1.33	0.67	1.00	1.00	1.00	0.00
2012	8	8	2.00	0.29	0.38	0.47	0.62	0.21
2013	36	54	3.00	0.09	0.39	0.45	0.59	0.23
2014	51	111	4.35	0.09	0.51	0.57	0.62	0.27
2015	61	142	4.66	0.08	0.40	0.44	0.40	0.28
All years	86	231	5.37	0.06	0.34	0.45	0.39	0.30

Table IV. Top ten disciplines in terms of degree, betweenness, and closeness centrality (2004–2015)

Ranking	Discipline	Degree	Discipline	Closeness	Discipline	Betweenness
1	Computer science	34	Computer science	0.57	Computer science	0.40
2	Engineering	32	Engineering	0.49	Engineering	0.21
3	Social sciences	20	Social sciences	0.46	Social sciences	0.16
4	Mathematics	14	Mathematics	0.45	Biochemistry and molecular biology	0.14
5	Automation and control systems	14	Automation and control systems	0.44	Neurosciences and neurology	0.12
6	Business and economics	14	Business and economics	0.44	Biotechnology and applied microbiology	0.11
7	Chemistry	13	Chemistry	0.42	Business and economics	0.09
8	Information science and library science	12	Information science and library science	0.43	Cell Biology	0.07
9	Biochemistry and molecular biology	11	Biochemistry and molecular biology	0.44	Cardiovascular system and cardiology	0.07
10	Materials science	11	Materials science	0.42	Research and experimental medicine	0.07

relatively weak; and only computer science, engineering, business and economics, and social sciences play an effective role of "bridge" connecting the majority of disciplines.

With the aid of the Louvain algorithm embedded in Pajek, communities of collaboration among disciplines were detected to reveal the aggregating characteristic. Shown in Table V, there are seven interdisciplinary communities named C1–C7, including four major communities led by computer science, business and economics, mathematics, and biotechnology and applied microbiology, respectively . It indicates that the majority of disciplines are connected with these central and important disciplines and clustered into interdisciplinary communities.

Assessment of interdisciplinarity based on Stirling's diversity index and specialization index

After performing social network analysis of co-occurrence among disciplines, the degree of interdisciplinarity could be calculated using Equation (1). The larger the value of Stirling's diversity index, the higher the degree of interdisciplinarity. According to Equation (4), the degree of specialization index was also calculated. The larger the value of the specialization index, the lower the degree of interdisciplinarity. The results are listed in Table VI and illustrated in Figure 1, which show the diversity index was increasing over time while the specialization index tended to decrease over time. Research in Big Data has shown a trend that integrates more diverse disciplines (high D) and becomes less specialized (low S) at the same time.

Community	Disciplines (top five)	Number of nodes	Number of edges
C1	Computer science; engineering; telecommunications; automation & control systems; operations research and management science	15	28
C2	Business and economics; social sciences; information science and library science; education and educational research; government and law	30	58
C3	Mathematics; science and technology; physics; chemistry; mathematical and computational biology	19	30
C4	Biotechnology and applied microbiology; genetics and heredity; research and experimental medicine; oncology; cardiovascular system and cardiology	10	12
C5	Materials science; environmental sciences and ecology; energy and fuels; geography; mechanics	6	7
C6	Medical informatics; health care sciences and services; general and internal medicine	3	2
C7	Remote sensing; imaging science and photographic technology; physical geography	3	3

Table V. The interdisciplinary communities in Big Data research (2004–2015)

Year	Stirling's diversity index	Specialization index
2004–2011	0.50	0.38
2012	0.50	0.40
2013	0.53	0.29
2014	0.57	0.20
2015	0.57	0.16
All years	0.72	0.19

Table VI. Stirling's diversity index and specialization index over time

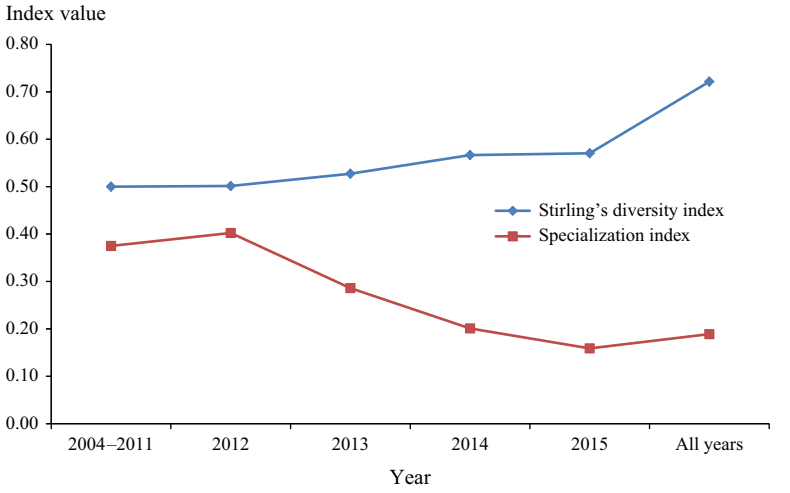

Figure 1.
Summary of Stirling's diversity index and specialization index over time

Figure 2 shows the relative position of interdisciplinarity in each year, and all years combined, in a two-dimensional chart using the values of Stirling's diversity index (*x*-axis) and specialization index (*y*-axis). The axes' origin (0.56, 0.27) is determined by the average of values. Year 2013 marked a milestone of interdisciplinarity of Big Data research. Research in Big Data before 2013 was more focused on a few disciplines; and after 2013 it turned to be more diverse and collaborative. These two indicators provide the quantitative evidence that, overall, interdisciplinary collaboration in Big Data is very extensive at a high level.

Trends of discipline communities in Big Data research

The strategic diagram of values of degree centrality and density provides indications of the development status and trends of the discipline communities. The diagram was obtained

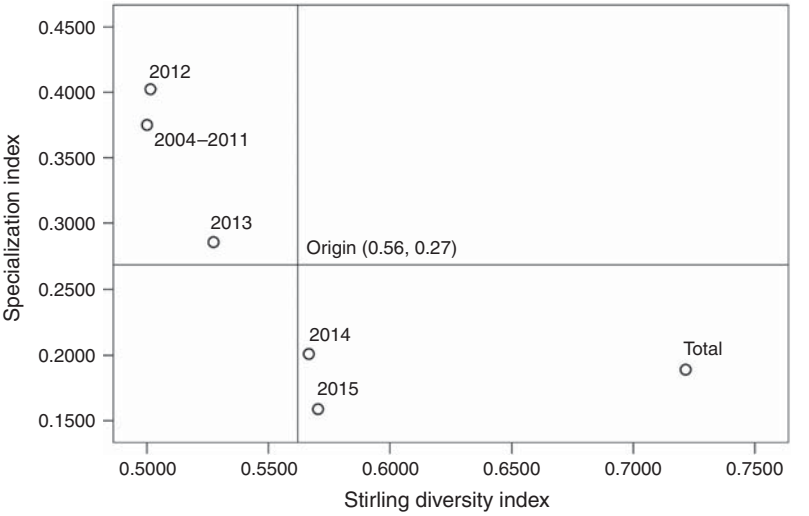

Figure 2.
The relative status of interdisciplinarity for each year and all years combined

through several steps. First, the value of degree centrality for every discipline was obtained, and summation of all disciplines in one community was then calculated. The centrality value for each community is equal to the average of summation. Second, the density value of each community was also calculated by Pajek. Table VII lists the values of centrality and density of communities. The strategic diagram is illustrated in Figure 3. The *x*-axis stands for degree centrality, while the *y*-axis stands for density, and the axes' origin (5.25, 0.21) is determined by the average of centrality and density.

C1, led by computer science, with high centrality and high density, are located in Quadrant I. Disciplines in C1, collaborating with each other intensively, are central to the whole interdisciplinary network and tend to be mature in Big Data research. It is the most important and central discipline community in Big Data research. C4 is located in Quadrant II with high density but low centrality, indicating that biotechnology and applied microbiology and other related disciplines in this community are connected intensively but not central to the whole interdisciplinary network. These disciplines have formed an independent and systematic subfield in Big Data research, and have the tendency to become the core area in the future (e.g. Belle *et al.*, 2015).

C2 and C3 are located in Quadrant III with both low density and centrality, but very close to the *y*-axis. Disciplines in C2 and C3 are relatively peripheral and immature, suggesting that Big Data research conducted in business and economics and other related disciplines has not been well-developed, but has good prospects for development. Mathematics and other related disciplines, as fundamental disciplines in Big Data research, tend to be applied by researchers in other disciplines.

Community	Centrality	Density
C1 Computer science	8.00	0.27
C2 Business and economics	5.00	0.13
C3 Mathematics	5.11	0.18
C4 Biotechnology and applied microbiology	2.90	0.27

Table VII.
Centrality and density of four major discipline communities

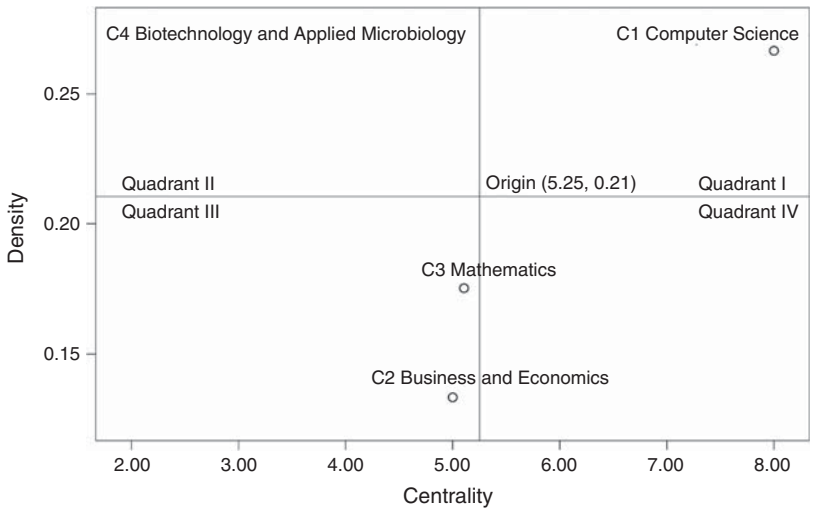

Figure 3.
The strategic diagram with major four discipline communities

Discussion

This study examined the interdisciplinarity of Big Data research with the aid of social network analysis and two index measures: Stirling's diversity index and specialization index. The results indicated that the degree of interdisciplinarity in Big Data research is increasing over time. This study found that Big Data research involves a large number of disciplines; but their distribution is very unbalanced in terms of number of published articles. Both the number of disciplines and scale of collaborations among them are increasing over time. The collaborations among disciplines are not concentrated but aggregated with a few powerful ones as the core. Disciplines tend to directly collaborate with others, rather than indirectly, to resolve the issues related to Big Data.

Particularly after 2013, disciplines in Big Data research are more diverse and collaborative. Big Data research has extended from hard core technological disciplines such as computer science, engineering, and telecommunications, to disciplines that produce, handle, and utilize data such as business and economics, social sciences, information science and library science, and scientific research in general. Such trend suggests the increasing importance of data in scholarly research and contributes to the understanding of the data scholarship phenomenon and complex relationships between data and scholarship (Borgman, 2015).

This study confirmed Big Data research being a multidisciplinary and IDR field that requires collaborations from a great diversity of disciplines (Chen *et al.*, 2014; Hilbert, 2016). It also found that the interdisciplinarity of most disciplines is very low, except for computer science and business and economics; and the difference among disciplines vary to a great extent. Several Big Data foundational disciplines (e.g. computer science) and applied disciplines (e.g. business and economics) as defined by Singh *et al.* (2015) have high interdisciplinarity, sharing both the most extensive and intensive collaborations with other disciplines.

Whilst interdisciplinarity has been growing, research collaborations among disciplines became more stable and mature after 2013, and formed seven distinct disciplinary communities. These disciplinary communities suggest divergent research directions in Big Data research with different statuses and development tendencies. Collaborations in some areas related to Big Data research techniques tend to be mature, maintaining a sound momentum in development. The collaborative applications in biotechnology and applied microbiology and other related disciplines have become more and more systematic, which led to a new independent subfield in Big Data research. Meanwhile, other communities or directions, especially the application research in disciplines such as social science, are in their infancy.

These findings illustrate disciplines are at various stages in the Big Data research endeavor and play different roles in the collaborative research effort. It adds empirical evidence to the suggestion that recognizing the disciplinary differences in research needs, practices, methods, and purposes would be helpful to understand data scholarship in various disciplinary contexts so that policies and support can be developed to accommodate the diversity of data scholarship across different disciplines (Borgman, 2015).

Conclusions

The findings of this study offer a better understanding of interdisciplinary collaboration in the Big Data research effort and shed new light on the interdisciplinary nature, status, and patterns of the research effort. The contribution of this study is to use multiple methods and tools, and a combined quantitative indicator to measure the interdisciplinary collaboration in a research field. This approach can serve as a general framework to study interdisciplinarity and is applicable in other research areas. Future studies may further validate and refine the approach through more empirical studies and with a comparative approach with other methods.

As this study had a non-exhaustive sample, the results should be interpreted with the limitations of the coverage of the literature database and the retrieval articles using our search strategy that emphasized more relevant articles than recall of all related articles. Future research could also build on this study by expanding the timeframe and scope of Big Data related research publications, in order to monitor and capture development trends. It could likewise identify future directions in this important research area that affects and involves fields that deal with large amounts of data. In addition, more in-depth research from other perspectives or sources, for example, revealing the interdisciplinary structure and patterns of the National Science Foundation's projects related to Big Data research, would also be valuable. Finally, it would be helpful to conduct some additional bibliometric analysis and in-depth content analysis of the papers in this area to know how these interdisciplinary collaborations are playing out, what the major benefits are for such collaborations, and why interdisciplinarity is common in some disciplines and rare in others.

References

Acharjya, D.P. and Ahmed, P.K. (2016), "A survey on Big Data analytics: challenges, open research issues and tools", *International Journal of Advanced Computer Science and Applications*, Vol. 7 No. 2, pp. 511-518.

Akoka, J., Comyn-Wattiau, I. and Laoufi, N. (2015), "Research on Big Data characterizing the field and its dimensions", in Jeusfeld, M.A. and Karlapalem, K. (Eds), *Advances in Conceptual Modeling, ER Workshops*, Springer, New York, NY, pp. 173-183.

Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K. and Taha, K. (2015), "Efficient machine learning for Big Data: a review", *Big Data Research*, Vol. 2 No. 3, pp. 87-93.

Bardi, M., Zhou, X., Li, S. and Lin, F. (2014), "Big Data security and privacy: a review", *China Communications*, Vol. 11 No. 2, pp. 135-145.

Belle, A., Thiagarajan, R., Soroushmehr, S.M.R., Navidi, F., Beard, D.A. and Najarian, K. (2015), "Big Data analytics in healthcare", *BioMed Research International*, 370194, doi: 10.1155/2015/370194.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. (2008), "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008 No. 10, P10008.

Boerner, K. (2011), "Plug-and-play macroscopes", *Communications of the ACM*, Vol. 54 No. 3, pp. 60-69.

Borgman, C.L. (2015), *Big Data, Little Data, No Data: Scholarship in the Networked World*, MIT Press, Cambridge, MA.

Callon, M., Courtial, J.P. and Laville, F. (1991), "Co-word analysis as a tool for describing the network of interactions between basic and technological research-the case of polymer chemistry", *Scientometrics*, Vol. 22 No. 1, pp. 155-205.

Cassi, L., Mescheba, W. and de Turckheim, E. (2014), "How to evaluate the degree of interdisciplinarity of an institution?", *Scientometrics*, Vol. 101 No. 3, pp. 1871-1895.

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

Chen, H., Chiang, R.H.L. and Storey, V.C. (2012), "Business intelligence and analytics: from Big Data to big impact", *MIS Quarterly*, Vol. 36 No. 4, pp. 1165-1188.

Chen, M., Mao, S. and Liu, Y. (2014), "Big Data: a survey", *Mobile Networks & Applications*, Vol. 19 No. 2, pp. 171-209.

Doreian, P., Lloyd, P. and Mrvar, A. (2013), "Partitioning large signed two-mode networks: problems and prospects", *Social Networks*, Vol. 35 No. 2, pp. 178-203.

Einav, L. and Levin, J. (2014), "Economics in the age of big data", *Science*, Vol. 346 No. 6210, p. 1243089.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V.R., Tsou, A., Weingart, S. and Sugimoto, C.R. (2015), "Big Data, bigger dilemmas: a critical review", *Journal of the Association for Information Science and Technology*, Vol. 66 No. 8, pp. 1523-1545.

Emani, C.K., Cullot, N. and Nicolle, C. (2015), "Understandable Big Data: a survey", *Computer Science Review*, Vol. 17, pp. 70-81.

Fang, H., Zhang, Z.Y., Wang, C.J., Daneshmand, M., Wang, C.G. and Wang, H.G. (2015), "A survey of Big Data research", *IEEE Network*, Vol. 29 No. 5, pp. 6-9.

Freeman, L.C. (1979), "Centrality in social networks conceptual clarification", *Social Networks*, Vol. 1 No. 3, pp. 215-239.

Gandomi, A. and Haider, M. (2015), "Beyond the hype: Big Data concepts, methods, and analytics", *International Journal of Information Management*, Vol. 35 No. 2, pp. 137-144.

Gu, D.X., Li, J.J., Li, X.G. and Liang, C.Y. (2017), "Visualizing the knowledge structure and evolution of Big Data research in healthcare informatics", *International Journal of Medical Informatics*, Vol. 98, pp. 22-32.

Hilbert, M. (2016), "Big Data for development: a review of promises and challenges", *Development Policy Review*, Vol. 34 No. 1, pp. 135-174.

Huang, M.-H. and Chang, Y.-W. (2011), "A study of interdisciplinarity in information science: using direct citation and co-authorship analysis", *Journal of Information Science*, Vol. 37 No. 4, pp. 369-378.

Huang, Y., Schuehle, J., Porter, A.L. and Youtie, J. (2015), "A systematic method to create search strategies for emerging technologies based on the Web of Science: illustrated for 'Big Data' ", *Scientometrics*, Vol. 105 No. 3, pp. 2005-2022.

Karlovcec, M. and Mladenic, D. (2015), "Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring", *Scientometrics*, Vol. 102 No. 1, pp. 433-454.

Kitchin, R. (2014), "Big Data, new epistemologies and paradigm shifts", *Big Data and Society*, Vol. 1 No. 1, pp. 1-12.

Leahey, E. (2006), "Gender differences in productivity – research specialization as a missing link", *Gender and Society*, Vol. 20 No. 6, pp. 754-780.

Leydesdorff, L. (2007), "Betweenness centrality as an indicator of the interdisciplinarity of scientific journals", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 9, pp. 1303-1319.

Leydesdorff, L. and 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*, Vol. 65 No. 1, pp. 164-177.

Leydesdorff, L. and Rafols, I. (2011), "Indicators of the interdisciplinarity of journals: diversity, centrality, and citations", *Journal of Informetrics*, Vol. 5 No. 1, pp. 87-100.

Liang, Q.L., Ren, J., Liang, J., Zhang, B.J., Pi, Y.M. and Zhao, C.L. (2015), "Security in Big Data", *Security and Communication Networks*, Vol. 8 No. 14, pp. 2383-2385.

Luzar, B., Levnajic, Z., Povh, J. and Perc, M. (2014), "Community structure and the evolution of interdisciplinarity in Slovenia's scientific collaboration network", *PLOS ONE*, Vol. 9 No. 4, doi: 10.1371/journal.pone.0094429.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A. (2011), "Big Data: the next frontier for innovation, competition, and productivity", available at: www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (accessed May 5, 2017).

Marx, V. (2013), "The big challenges of Big Data", *Nature*, Vol. 498 No. 7453, pp. 255-260.

Michael, K. and Miller, K.W. (2013), "Big Data: new opportunities and new challenges", *Computer*, Vol. 46 No. 6, pp. 22-24.

Mittelstadt, B.D. and Floridi, L. (2016), "The ethics of Big Data: current and foreseeable issues in biomedical contexts", *Science and Engineering Ethics*, Vol. 22 No. 2, pp. 303-341.

Morillo, F., Bordons, M. and Gomez, I. (2003), "Interdisciplinarity in science: a tentative typology of disciplines and research areas", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 13, pp. 1237-1249.

Olmedilla, M., Martinez-Torres, M.R. and Toral, S.L. (2016), "Harvesting Big Data in social science: a methodological approach for collecting online user-generated content", *Computer Standards and Interfaces*, Vol. 46, pp. 79-87.

Porter, A.L. and Rafols, I. (2009), "Is science becoming more interdisciplinary? Measuring and mapping six research fields over time", *Scientometrics*, Vol. 81 No. 3, pp. 719-745.

Porter, A.L., Roessner, J.D. and Heberger, A.E. (2008), "How interdisciplinary is a given body of research?", *Research Evaluation*, Vol. 17 No. 4, pp. 273-282.

Porter, A.L., Cohen, A.S., Roessner, J.D. and Perreault, M. (2007), "Measuring researcher interdisciplinarity", *Scientometrics*, Vol. 72 No. 1, pp. 117-147.

Rafols, I. and Meyer, M. (2010), "Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience", *Scientometrics*, Vol. 82 No. 2, pp. 263-287.

Salton, G. and McGill, M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY.

Savitz, E. (2012), "Gartner: 10 critical tech trends for the next five years", available at: www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years (accessed May 5, 2017).

Schummer, J. (2004), "Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology", *Scientometrics*, Vol. 59 No. 3, pp. 425-465.

Singh, V.K., Banshal, S.K., Singhal, K. and Uddin, A. (2015), "Scientometric mapping of research on 'Big Data'", *Scientometrics*, Vol. 105 No. 2, pp. 727-741.

Smith, A., Sparks, L. and Goulding, J. (2015), "Using commercial Big Data to inform social policy: possibilities, ethics, methods and obstacles", *Journal of Macromarketing*, Vol. 35 No. 1, pp. 141-141.

Stegmann, J. and Grohmann, G. (2003), "Hypothesis generation guided by co-word clustering", *Scientometrics*, Vol. 56 No. 1, pp. 111-135.

Stirling, A. (2007), "A general framework for analysing diversity in science, technology and society", *Journal of the Royal Society Interface*, Vol. 4 No. 15, pp. 707-719.

Stopar, K., Drobne, D., Eler, K. and Bartol, T. (2016), "Citation analysis and mapping of nanoscience and nanotechnology: identifying the scope and interdisciplinarity of research", *Scientometrics*, Vol. 106 No. 2, pp. 563-581.

Suresh, S. (2016), "Big Data and predictive analytics: applications in the care of children", *Pediatric Clinics of North America*, Vol. 63 No. 2, pp. 357-366.

Taglang, G. and Jackson, D.B. (2016), "Use of 'Big Data' in drug discovery and clinical trials", *Gynecologic Oncology*, Vol. 141 No. 1, pp. 17-23.

Tomov, D.T. and Mutafov, H.G. (1996), "Comparative indicators of interdisciplinarity in modern science", *Scientometrics*, Vol. 37 No. 2, pp. 267-278.

Wagner, C.S., Roessner, J.D., Bobb, K., Klein, J.T., Boyack, K.W., Keyton, J., Rafols, I. and Borner, K. (2011), "Approaches to understanding and measuring interdisciplinary scientific research (IDR): a review of the literature", *Journal of Informetrics*, Vol. 5 No. 1, pp. 14-26.

Wang, J., Thijs, B. and Glänzel, W. (2015), "Interdisciplinarity and impact: distinct effects of variety, balance, and disparity", *PLOS One*, Vol. 10 No. 5, p. e0127298, doi: 10.1371/journal.pone.0127298.

Wu, L., Yuan, L. and You, J. (2015), "Survey of large-scale data management systems for Big Data applications", *Journal of Computer Science and Technology*, Vol. 30 No. 1, pp. 163-183.

Xu, H., Guo, T., Yue, Z., Ru, L. and Fang, S. (2016), "Interdisciplinary topics of information science: a study based on the terms interdisciplinarity index series", *Scientometrics*, Vol. 106 No. 2, pp. 583-601.

Yegros-Yegros, A., Rafols, I. and D'Este, P. (2015), "Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity", *PLOS One*, Vol. 10 No. 8, doi: 10.1371/journal.pone.0135095.

Further reading

van Eck, N.J. and Waltman, L. (2010), "Software survey: VOSviewer, a computer program for bibliometric mapping", *Scientometrics*, Vol. 84 No. 2, pp. 523-538.

About the authors

Jiming Hu, PhD, is Associate Professor of School of Information Management at Wuhan University, China. His major research areas include information behavior, information visualization, and information service. He holds PhD, Master's, and Bachelor Degrees from the School of Information Management at Wuhan University in China. His scholarly publications consist of more than 20 papers and two books, as well as about ten national and international conference presentations, invited lectures and keynote speeches. He was the PI and Co-PI of five China National Science Foundations (NSF) and other projects. Jiming Hu is the corresponding author and can be contacted at: hujiming@whu.edu.cn

Yin Zhang, PhD, is Professor of Library and Information Science at Kent State University. Her major research areas include information uses and users, information-seeking behavior, information organization, and database systems and design. She holds a PhD Degree in Library and Information Science from the University of Illinois at Urbana-Champaign, MS and BS Degrees in Information Science from Wuhan University. She is Author of numerous refereed journal articles, book chapters and conference presentations, and Author or Editor of several books. She also has been PI or Co-PI on more than $1m in grants from the Institute of Museum and Library Services.