Structure and patterns of cross-national Big Data research collaborations

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

Received 1 December 2016
Revised 17 June 2017
Accepted 26 June 2017

Abstract

Purpose – The purpose of this paper is to reveal the structure and patterns of cross-national collaborations in Big Data research through application of various social network analysis and geographical visualization methods.
Design/methodology/approach – The sample includes articles containing Big Data research, covering all years, in the Web of Science Core Collection as of December 2015. First, co-occurrence data representing collaborations among nations were extracted from author affiliations. Second, the descriptive statistics, network indicators of collaborations, and research communities were calculated. Third, topological network maps, geographical maps integrated with topological network projections, and proportional maps were produced for visualization.
Findings – The results show that the scope of international collaborations in Big Data research is broad, but the distribution among nations is unbalanced and fragmented. The USA, China, and the UK were identified as the major contributors to this research area. Five research communities are identified, led by the USA, China, Italy, South Korea, and Brazil. Collaborations within each community vary, reflecting different levels of research development. The visualizations show that nations advance in Big Data research are centralized in North America, Europe, and Asia-Pacific.
Originality/value – This study applied various informetric methods and tools to reveal the collaboration structure and patterns among nations in Big Data research. Visualized maps help shed new light on global research efforts.

Keywords Research networks, Maps, Big Data research, Geographical visualization, International collaboration, Network structure and patterns

Paper type Research paper

1. Introduction

Big Data has become a critical research frontier, with the potential to revolutionize many fields, including business, science, and public administration (Savitz, 2012). Nature Publishing Group (2008) and Science/AAAS (2011), two premier scientific journals, produced special issues to analyze the significance, challenges, and impacts of Big Data. It is evident that Big Data has drawn huge attention from many nations in recent years because of its potential for increasing business productivity and breakthroughs in scientific research (Chen and Zhang, 2014).

Big Data is still a fast evolving field of both research and practice. It poses a challenge to comprehensively define Big Data, as the term has been used interchangeably to refer to many different and yet often intertwined aspects of studies, such as the characteristics of the data source, a class of analytic approaches – Big Data methods, and/or an overall approach to problem solving (Tonidandel *et al.*, 2016). 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).

Journal of Documentation
Vol. 73 No. 6, 2017
pp. 1119-1136

0022-0418
DOI 10.1108/JD-12-2016-0146

Since the beginning of this new century, there have been a large number of international theoretical and technical studies on Big Data involving many nations, regions, institutions, and researchers (Chen *et al.*, 2014; Emani *et al.*, 2015; Fang *et al.*, 2015). Many nations and governmental bodies play a vital role in promoting Big Data research and applications. In recent decades, the USA, European Union, and China have advanced in development planning of Big Data (Emani *et al.*, 2015). For example, in March 2012, the US Government announced the investment of 200 million dollars to launch the "Big Data Initiative" with a focus on foundational technology and public sector applications, aimed at promoting theoretical and technological research and strengthening scientific research, education, and national security (Kalil, 2015). Similarly, in 2013, the UK proposed a series of initiatives supporting the development of Big Data, primarily for the high-technology field, and government and public sector applications. In addition, China, Japan, Australia, Singapore, and several other nations issued related development strategies and accelerated research and application.

In recent years, many governments and institutions have promoted international collaborations in Big Data (Fang *et al.*, 2015). Previous research has shown that as researchers are encouraged to work and study collaboratively, and as international co-authorship has become mainstream in Big Data research (George *et al.*, 2014), there has been an increase in papers involving these collaborations published in journals (Michael and Miller, 2013). As a result, there is a great need to reveal the structure and patterns of international collaboration at the national level.

This study expands on previous research exploring patterns and characteristics of international collaborations in Big Data research. Different from the traditional co-authorship research, this study is the first step toward a greater understanding of the collaborations among nations as extracted from author affiliation addresses. The publication records retrieved from the Web of Science (WoS) database allow us to examine the extent of cross-national collaborations geographically. This research aims to reveal the collaboration patterns and network structures among nations, detect the structural communities of nations in terms of cross-national collaborations, and then visualize these collaborations using social network and geographical mapping. Furthermore, the characteristics of collaboration networks are examined in order to reveal the general status and position of each nation in Big Data research.

This research fills the gap with a detailed and systematic visual mapping of the collaborations in Big Data research, and examination of the characteristics of the overall international collaboration network and its sub-networks. The results of this study will contribute to a greater understanding of international collaboration in Big Data research and shed light on where each nation is positioned in this relatively new and evolving global research effort. The results may suggest potential opportunities for international community development in this research frontier that benefit all involved.

2. Literature review

2.1 The landscape of Big Data development

The past decade has seen an explosive, global increase in what is described as "Big Data," the large and complex data sets that cannot be processed by traditional means (Kalil, 2015). While Big Data creates business and research value, it also generates significant challenges (e.g. Marx, 2013; Chen *et al.*, 2012) in terms of networking, storage, management, analytics, and even ethics (Fang *et al.*, 2015). Therefore, researchers stress the urgent need to develop and innovate new techniques and technologies to process this data and benefit various specified purposes (Chen and Zhang, 2014). Big Data research has covered many different but yet often overlapped areas such as data sources, methods, and approaches to problem solving (Tonidandel *et al.*, 2016). It is covered as an abstract

concept, and the definition and features, as well as the value, of Big Data have been identified from various perspectives (e.g. Manyika *et al.*, 2011; Hitzler and Janowicz, 2013). Further, an abundance of Big Data research involves techniques and technologies for its capture, curation, analysis, and visualization (e.g. Nandi *et al.*, 2012; Wu *et al.*, 2014). Additionally, applications of theories and techniques of Big Data are mostly studied in the context of various disciplines and fields, such as commerce and business, management, culture and art, engineering and technology, and scientific research (Murdoch and Detsky, 2013; Hampton *et al.*, 2013).

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

Due to the evolving nature of Big Data research, reviews and empirical research have sought to reveal and understand its research status. Big Data research has been gauged by the quantitative proliferation of journal articles about Big Data and increasing industries and research approaches involved (Ekbia *et al.*, 2015; Wamba *et al.*, 2015), and joint efforts from academics, industries, and governments (Chen *et al.*, 2014; Khan *et al.*, 2014).

Most previous work used a qualitative analysis approach to shed light on the status of related research development. In a large-scale literature survey of Big Data applications, challenges, techniques, and technologies, Chen and Zhang (2014) concluded that Big Data will deliver a tremendous value to both governments and enterprises with increases in returns of their scientific investments. Similarly, Chen *et al.* (2014) presented both background and contemporary developments of Big Data, and suggested that the era of Big Data has arrived. After discussing related theories, technologies, and applications of Big Data, the authors speculated on possible future developments, concluding that there will be great breakthroughs in Big Data research areas.

Other reviews assess the challenges, identify underlying issues, and suggest possible solutions related to Big Data research. In a state-of-the-art overview of Big Data initiatives, technologies, and research in industries and academia, Fang *et al.* (2015) indicated that multidisciplinary collaborations and joint efforts from industries, academics, and governments are significant in advancing research. In a multidisciplinary review of Big Data within the sciences, humanities, policy, and trade literature, Ekbia *et al.* (2015) provided a synthesis of Big Data in an effort to identify some common underlying issues (e.g. scientific methodology, epistemology, aesthetics, ethics, and political economy) and provide a broader understanding of Big Data drivers, barriers, and challenges. Based on their review, the authors also suggested possible future directions (e.g. digital technologies, privacy protection, and multidisciplinary collaboration) for Big Data research and developments for related stakeholders.

Singh *et al.* (2015) took a quantitative approach in their scientometric analysis of research structure and patterns of Big Data research during 2010-2014. The study accomplished a more systematic review of Big Data through analyzing the research output in that timeframe as extracted from WoS and Scopus, revealing authorship relations and nation-level collaborations. Using scientometric indicators, it also elaborated on the major contributors, top publication sources, thematic trends, and emerging themes in this field. For nation-level collaborations, the study computed research output by nation and the characteristics of international collaboration papers (ICP). The study found that in terms of output, the USA was the top contributor to Big Data research, followed by China, the UK, and Germany. The study illustrated a snapshot of the cross-national collaboration network and observed that the USA and China collaborated most for ICP, followed by the collaborations between the USA and the UK. Overall, the USA had the highest ICP instances involving different countries. However, the study did not elaborate on the characteristics of the overall international collaboration network and its sub-networks, nor the roles that the nations play in the cross-national collaboration network.

2.3 Mapping methods in previous research

Previous research has not performed detailed and systematic mapping of the national-level collaborations in Big Data research. Some studies using the approach on other research fields demonstrate its effectiveness in revealing research patterns and providing a better understanding of a research area. These previous studies also illustrate techniques and tools for such analysis that are helpful in formulating the research plan of this study.

Several previous studies used mapping methods, based on the co-occurrence of nations, regions, or cities of co-authors in the published articles, to generate the international collaboration network. Grauwin and Jensen (2011) developed a set of routines to show various research maps using co-occurrence (e.g. institutions) and bibliographic coupling. Specifically, they proposed different ways to map relations among scientific institutions based on published articles using authors' addresses, and revealed the international collaborations of nations with a networking map. Catala-Lopez *et al.* (2012) conducted a systematic network analysis on co-authorship and institutional collaborations in the area of cost-effectiveness. They identified six major research groups in terms of the co-authorship relations, and drew the full network of co-authorship and institutional collaborations using Pajek. Based on co-authorship networking indicators, the study concluded that collaboration within this field was weak and fragmented.

Some previous studies have also drawn the collaboration network using geographical maps such as Google Maps (www.google.com/maps) and GPS Visualizer (www.gpsvisualizer.com), and plotted distribution patterns and networks of relations among regions. Bornmann and Leydesdorff (2012) produced regional maps showing where excellent papers have emerged and where these papers frequently occur in Information Science. On a global map, they located, by city, authors in Europe and the USA who have published highly cited papers. Circles of different sizes and colors represent cities according to productivity and normalized impact, based upon I3-metrics, as introduced by Leydesdorff and Bornmann (2011). Bornmann and Ozimek (2012) developed a toolbox for the statistical software package Stata for importing bibliometric data and processing author address information. After geocoding the author addresses using Yahoo!, geographical visualization of cities was conducted using distinct circles to show the spatial distribution of authors. Kristensen (2015) studied the geography of these international relations by analyzing bibliometric data, and visualized the global distribution of research articles and the co-authorship at the city and institutional levels, regardless of the number of co-authors. Clear stratification structures were found within USA disciplines by mapping the city and institutional outputs and co-authorships, which demonstrates the value of mapping visualizations to reveal hidden patterns involving geographical locations.

It is crucial to refer to the work accomplished by Leydesdorff and his colleagues, who provide a series of approaches, methods, and tools to address the co-occurrence structures from bibliometric and geographical perspectives. In performing the visualization from the geographic level, Leydesdorff and Rafols (2011) used the publication records in Science Citation Index (SCI) of two emerging technologies to illustrate the geographical network of cities with overlays to Google Maps. It found that the structure and patterns of collaboration network, with nodes and links, showed both preferential attachment and small-world characteristics. Other fields have conducted similar visual mapping. For example, researchers achieved the mapping of patent data, using overlays to Google Maps at the city level, with the aid of their own developed techniques (Leydesdorff and Bornmann, 2012). In their study, a geographical distribution of nanotechnology patents across cities was generated using Google Maps. More importantly, systematic and comprehensive guidelines have been provided to conduct the international co-authorship analysis in geographical view, involving the extraction of address information in bylines and then the abstraction of corresponding nations (Leydesdorff *et al.*, 2014).

2.4 Rationale for this study

A review of previous research shows most reviews of Big Data research follow a qualitative approach, which underscores a need for more quantitative research particularly in exploring the patterns and characteristics of international collaborations. Given the emphasis of Big Data research and development as a commonly recognized national strategy, as well as the importance of collaborations, it is timely and necessary to provide a state-of-the-art analysis to measure and map the Big Data research structure from the perspective of collaboration amongst nations, including geographical distribution and connections, in order to better understand the global research effort. Networking indicators, including centrality, density, and a nation's relation to the research community to which it belongs, can be calculated and employed to explain collaborative patterns and statuses.

In addition, this study projects the topological graph of international collaboration among nations on to the geographic map and illustrates the differences among national nodes and their links, visually positioning each nation's scholarly contributions and collaboration ties in the larger picture of the global research effort.

3. Methodology

3.1 Data collection and processing

A sample of Big Data publications in the WoS Core Collection database as of December 1, 2015 was selected through multiple steps for this study:

(1) Search strategy for publications on Big Data.

As mentioned in the literature review, the growth of Big Data research has been gauged by the quantitative proliferation of related literature, and the research status analyzed based on the publications (Ekbia *et al.*, 2015; Wamba *et al.*, 2015). To locate materials about Big Data research, researchers have specifically employed a strategy using "Big Data" in the title, or as a keyword in major academic databases (Ekbia *et al.*, 2015).

For this study, initial trial searches were conducted using "big data" as a search phrase in fields such as Title, Abstract, Author Keywords, and Keywords Plus. The retrieved publication records from the trial searches were reviewed and compared.

It was found that the WoS Core Collection database did not have "big data" in its Keywords Plus field. There were records containing "big data" in Abstract or Author Keywords but not in Title. Some of the publications represented in such records did not appear to be about Big Data research itself. Rather, "big data" was merely used as a setting or context for other research. In addition, the set of publications based on the Title field search was identical to that based on both the Title and Author Keywords search, and a review of the records confirmed the relevancy of the publications. As a result, the search strategy was to use "big data" as a search phrase in the Title and Author Keywords fields for a relevant sample of publications on the topic.

(2) Sample and data collection.

Using this search strategy, the initial data set consisted of the bibliographic records of 1,426 publications, including journal articles, reviews, and proceedings papers. Through screening, 1,350 containing author affiliation address information were selected as the final data sample. Table I lists the number of publications in the sample by year. Big Data publications started proliferating in 2012, and the majority of the publications have been co-authored.

As a first step in data processing, "Author Address," in the "C1" field in the bibliographic data, is abstracted to the corresponding nation and its capital due to the following two reasons.

Table I.
The Big Data research papers sample

Year	Number of papers	Number of papers with author address information	Number of papers co-authored
2004	1	0	0 (0.00%)
2009	2	2	1 (50.00%)
2010	1	1	1 (100.00%)
2011	8	6	5 (62.50%)
2012	74	59	53 (71.62%)
2013	429	405	352 (82.05%)
2014	543	521	431 (79.37%)
2015	368	356	310 (84.24%)
Overall	1,426	1,350	1,153 (80.86%)

First, given the study's primary research purpose, the information regarding nations corresponding to authors is supportive of researching international collaborations in Big Data. Second, it is necessary to provide one specific location at the city level, using geographic coordinates, to represent the nation for geographical analysis. At this point, the relationship data of co-occurrence between nations was extracted according to the corresponding co-authorship. The updated data file, including the relationship of co-occurrence between nations (co-nation data), was imported into analytical tools for further analysis to obtain statistics of the sample data set and measures such as co-occurrences, and to coordinate data for visualizing international collaborations and calculating network indicators.

3.2 Methods and tools

This study is conducted based on the theory of co-occurrence (e.g. Small and Griffith, 1974; Coulter *et al.*, 1998; Ding *et al.*, 2001), and on related methods and approaches developed in prior studies which demonstrated co-occurrence to be an effective and powerful tool for identifying and revealing the research structure and patterns underlying research fields or disciplines (e.g. Leydesdorff and Welbers, 2011; Ronda-Pupo and Guerras-Martin, 2012). This study chose nations as the analysis object to achieve its goal of examining international collaborations in Big Data research. Furthermore, this study proposes two methodological approaches to map the international collaboration of Big Data research: social network and geographical mapping analysis.

3.2.1 Social network analysis. According to the methodology of co-occurrence, such as co-author or co-keyword analysis (Newman, 2004; Borner *et al.*, 2003; Hellsten *et al.*, 2007), the co-occurrence data between nations can be obtained using SCI2 (Boerner, 2011) in a new bibliometric file as co-nation data. In co-nation data, nodes represent nations, while a link represents the fact that two authors from different nations have collaborated on at least one paper. The co-nation data are then used to generate the topological collaboration network in SCI2 and calculate the network indicators in Pajek (Doreian *et al.*, 2013).

The community structure is generated using the Louvain algorithm (Blondel *et al.*, 2008) in Pajek. This overall network graph of community structure is exported from Pajek to VOSviewer (van Eck and Waltman, 2010), and each network graph, representing corresponding research communities, is shown using the GUESS visualization component (Adar *et al.*, 2007) in SCI2. This approach best visualizes the attributed results of repeated comparison among these visualized tools.

A particular nation is selected to show its individual ego network and international collaboration patterns. Note that network indicators of each nation node and research community are used to illuminate the visualization results. These community networks can be compared in terms of density, degree, clustering coefficient, and so on. In all visualization

graphs, the sizes of nodes representing nations and the widths of links representing their connections are, respectively, proportional to the number of papers that include those nations and to the co-occurrences between them.

3.2.2 Geographical mapping analysis. Supported by the social network analysis mentioned above, this study attempts to project the topological international collaboration network onto a worldwide geographical map. The geographical maps generated by CiteSpace and GPS Visualizer are combined and compared; they also supplement each other to explain the collaboration patterns.

The coordinate data of each nation and its capital are generated using Google Earth KML generator in CiteSpace (Chen, 2006) and GPS Visualizer's Address Locator (www.gpsvisualizer.com/geocoder). CiteSpace can help generate a KML file that includes both the co-nation relationships and nation coordinates. After scaling the coordinate nodes according to the frequency of nations, the KML file is imported into GPS Visualizer to generate the global collaboration network. GPS Visualizer can zoom in or entertain a global view.

An important process is to adjust the KML file, as the study found some coordinates of cities in the KML file generated by CiteSpace were not precise. GPS Visualizer's Address Locator helped correct them. After careful verification, the KML file was finalized after adding the properties of nodes and links representing nations and connections (such as the sizes and colors of nodes and links). Subsequently, the finalized KML was imported into GPS Visualizer to generate the geographical map. In addition, when combined with the "Geospatial" function in SCI2, the choropleth map and proportional symbol map showing the distribution of nations were generated. The depth of color and node sizes can differentiate nations according to their number of publications in Big Data research.

4. Results

4.1 Network analysis of international collaborations

4.1.1 Descriptive statistics. In this study, 66 nations are identified, as shown in Table II. The USA and China, having the largest number of papers (505 and 316), are the two most prominent nations in publishing Big Data research (50.31 percent of the total number). More specifically, there are 695 connections with the USA and 217 connections with China, approximately 72.9 percent of the total connections among nations. It could be said they lead the world in Big Data research. The UK (including England, Scotland, Wales, and Northern Ireland), Germany, India, Australia, South Korea, and Canada are the second-tier members, having published at least 50 papers related to Big Data. Japan, Italy, Spain, and France, with greater than 30 papers, also contribute to Big Data research and closely collaborate with other nations. The Netherlands and eight other nations contributed at least ten papers each in Big Data in these years. The remaining nations are either relatively new or collaborative contributors along with other leading nations in Big Data research.

4.1.2 Network characteristics. In total, 12 nations lack connections with others, but the remaining nations (54) comprise the largest connected component of international collaboration relations. According to the theory of social network analysis, there is a path between any two nations of this collaboration network in the largest component (Nooy *et al.*, 2005). Usually, the largest component is chosen to be the focus because it could indicate the extent of international collaboration in Big Data research (Kretschmer, 2004). This result suggests the scope of international collaboration (54 of 66 nations) in Big Data research is relatively broad at 81.82 percent.

Table III shows the network statistics and parameters of the entire co-nation network. In general, relatively higher indicators in Big Data research are equal to the higher level of international collaboration among nations. Density measures the potential degree of

No.	Nation	Number of papers	No.	Nation	Number of papers
1	USA	505	34	Norway	5
2	China	316	35	Croatia[a]	4
3	UK	78	36	Qatar	4
4	Germany	63	37	South Africa	4
5	Australia	60	38	Egypt[a]	3
6	India	60	39	Hungary	3
7	South Korea	58	40	Israel	3
8	Canada	52	41	Jordan[a]	3
9	Japan	42	42	Lebanon	3
10	Italy	35	43	Lithuania[a]	3
11	Spain	34	44	Mexico[a]	3
12	France	30	45	Pakistan	3
13	The Netherlands	22	46	Thailand	3
14	Taiwan (China)	19	47	Tunisia[a]	3
15	Singapore	18	48	Algeria[a]	2
16	Brazil	16	49	Estonia	2
17	Austria	15	50	Indonesia	2
18	Poland	15	51	Oman	2
19	Switzerland	15	52	Sri Lanka[a]	2
20	Romania	14	53	United Arab Emirates	2
21	Turkey	10	54	Uruguay	2
22	Ireland	9	55	Bangladesh[a]	1
23	Malaysia	9	56	Bolivia	1
24	Sweden	9	57	Cyprus	1
25	Czech Republic	7	58	Ecuador	1
26	Denmark	7	59	Iraq	1
27	Saudi Arabia	7	60	Kenya	1
28	Finland	6	61	Liechtenstein[a]	1
29	New Zealand	6	62	Luxembourg	1
30	Portugal	6	63	Mauritius[a]	1
31	Russia	6	64	Peru	1
32	Belgium	5	65	Philippines	1
33	Greece	5	66	Slovakia[a]	1

Note: [a]Nations without international collaborations

Table II.
The 66 nations in Big Data research

Indicators	Result
Number of nodes	66
Largest component	54
Number of lines	205
Density	0.1433
Average degree	7.5926 (Min: 1, Max: 40)
Network all degree centralization	0.6350
Network all closeness centralization	0.6548
Network betweenness centralization	0.4160
Network clustering coefficient	0.6375
Communities	5 (Modularity: 0.1447)

Table III.
The network indicators of international collaboration in Big Data research

international collaboration, and was only 0.14, a relatively low level indicating weak collaboration in Big Data research.

Other network indicators are relatively high, suggesting that in recent years, nations have been closely collaborating in Big Data research and a few are powerful players in this research area. High closeness centrality is indicated by the short distances between nodes

within the network. In this context, it could also reflect that nations may collaborate with others more directly than indirectly, and that a nation's controlling degree of influence on Big Data research is not as high as the whole. Therefore, collaborations between any two nations are relatively independent, except when influenced by some leading nations. A collaboration network with high closeness centrality would be clearly clustered into several independently collaborating groups. Betweenness centrality indicates that there are considerable indirect collaborations through several important or central nations. Several nations would play an important bridging role that connects other nations in research collaborations.

Furthermore, the clustering coefficient of this international collaboration network, at a relatively high level (Khan *et al.*, 2011), means that two nations are much more likely to have collaborated if they share a third collaborative nation. This finding is reinforced by the result of five clearly distinguished groups. Obtained by the community calculation, it also indicates the obvious propensity of nations to collaborate in Big Data research.

As we know, network centrality is equal to the nation's capacity to dominate the Big Data research in the international collaboration network (Ma *et al.*, 2015). As shown in Table IV, the top 10 nations by network indicators of degree, betweenness, and closeness centrality are selected to demonstrate their location and role in this international collaboration network. The USA, China, the UK, and other nations with higher degree centrality are more central to the network structure and tend to have a greater capacity and opportunity to influence other nations. Nations with a high betweenness centrality play the role of connecting different nations and groups, such as the USA and the UK. They are two important pivots in the international collaboration to promote Big Data research. Other nations in the top 10 list of betweenness, with the value less than 0.1, play the connecting role but are relatively weaker. In the view of closeness, as a more sophisticated centrality measure (Freeman, 1979), the distance of any nation to all others in the network is short; emphasized again is that the USA, China, and the UK hold the highest capacity to collaborate with a majority of the nations in this international collaboration network. Notably, the lists of degree and closeness centrality for these nations are almost the same, and nations with high degree centrality usually have high closeness centrality in this international collaboration network (Yin *et al.*, 2006).

Five research communities were identified (among the 54 nations) using the Louvain algorithm as shown in Table V: Community1 – the US group with a total of 20 nations, Community2 – the Italy group of seven nations, Community3 – the China group of 12 nations, Community4 – the South Korea group of six nations, and Community5 – the Brazil group of nine nations. Therefore, 54 nations are categorized into five communities according to their international collaboration characteristics in Big Data research; and each community means that there are high levels of collaboration within it. However, these five communities are different with frequency and network indicators. For example, from the

Ranking	Nation	Degree	Nation	Betweenness	Nation	Closeness
1	USA	40	USA	0.430	USA	0.803
2	UK	24	UK	0.104	China	0.624
3	China	21	Japan	0.077	UK	0.624
4	Australia	20	China	0.074	Australia	0.596
5	Canada	19	Spain	0.072	Canada	0.576
6	Germany	19	Australia	0.064	Germany	0.570
7	Spain	15	Saudi Arabia	0.051	Spain	0.558
8	France	15	France	0.049	France	0.552
9	Switzerland	13	India	0.041	Switzerland	0.541
10	Singapore	12	Germany	0.040	Brazil	0.535

Table IV.
Top 10 nations by network indicators: degree, betweenness, and closeness centrality

Community	Number	Nation	Frequency	Network characteristics
1	20	USA, UK, Germany, Canada, France, Switzerland, Austria, Turkey, Czech Republic, New Zealand, Russia, Greece, Belgium, South Africa, Qatar, Lebanon, Uruguay, United Arab Emirates, Luxembourg, Iraq	Total frequency: 814 Average frequency: 40.7	Edges: 55 Average degree: 5.5 (1-18) Density: 0.2895 Network clustering coefficient: 0.7348
2	7	Italy, Spain, Poland, Romania, Ireland, Israel, Ecuador	Total frequency: 111 Average frequency: 15.86	Edges: 9 Average degree: 2.5714 (1-6) Density: 0.4286 Network clustering coefficient: 0.4571
3	12	China, India, Australia, Japan, Taiwan, Singapore, Norway, Oman, Indonesia, Philippines, Cyprus, Kenya	Total frequency: 527 Average frequency: 43.92	Edges: 15 Average degree: 2.5 (1-7) Density: 0.2273 Network clustering coefficient:0.2202
4	6	South Korea, The Netherlands, Malaysia, Saudi Arabia, Pakistan, Estonia	Total frequency: 101 Average frequency: 16.83	Edges: 6 Average degree: 2.0 (1-3) Density: 0.4000 Network clustering coefficient:0.3889
5	9	Brazil, Sweden, Denmark, Finland, Portugal, Hungary, Thailand, Bolivia, Peru	Total frequency: 52 Average frequency: 5.78	Edges: 14 Average degree: 3.111 (1-5) Density: 0.3889 Network clustering coefficient: 0.4667

Table V.
Communities in the largest component obtained by the Louvain algorithm in Pajek

perspective of frequency, the USA and China groups, with both total frequency and average frequency, are the two large groups that could be considered world leaders. These two groups occupy 83.55 percent of the total frequency and 68.75 percent of the total average frequency. However, from the perspective of network characteristics, higher density in the Italy and South Korea groups indicates relatively closer international collaboration in Big Data research within these groups, and the degree of close collaboration for each group is greater than 0.2. This relatively high level indicates that collaboration among nations within each group is closed and tends to be increasingly mature. In addition, the density of each research community is higher than that of the overall collaboration network, which demonstrates that collaborations within communities are closer than that between communities. Nations in Big Data research are relatively independent even as they form five separate research communities.

4.2 Visualization of the international collaboration in Big Data research

4.2.1 International collaboration patterns at the global level. Figure 1 portrays a full view of the international collaboration pattern graphically as overlays to geographical maps. The 66 nations and their connections are illustrated with the nation nodes scaled using their respective numbers of occurrence, and the width of the links proportionate to the number of co-occurrences. For example, the USA and China are the two largest nodes mapped, and the link representing the collaboration between them is shown using "Beijing, China – Washington, USA" and with "Width = 93." Therefore, nation nodes and collaboration links can be visualized at different scales in GPS Visualizer, which provides some interactive functions to view the basic network characteristics. GPS Visualizer also provides the

functionality to observe the partial map according to individual nations or links by zooming. For example, if you click the track of "Washington, USA – London, UK" in the "Track" window, the map zooms in on this link and its related nation nodes.

Figure 2 shows the distribution of nations in terms of their number of occurrences in the world map from the perspectives of a choropleth map and proportional symbol map. The figure distinguishes nations with a gradient of colors ranging from light to dark red, corresponding to the number of national occurrences. North America and Asia-Pacific (including the USA, Canada, China, India, Australia, South Korea, and Japan) could be seen as the main regions publishing a majority of the papers related to Big Data. The UK and Germany are two representative European nations, and are also categorized into the US group with more connections to other nations. Due to fewer publications and collaborations with other nations, a large proportion of nations in Africa and South America are not displayed here, indicating that Big Data research is undeveloped in these regions. In contrast, a majority of nations in Asia and Europe are colored, with several nations' color approaching dark red, indicating a high level of Big Data scholarship.

Nations with powerful research capacity are mostly concentrated north of Latitude 30, running through such places as Houston (a city in the USA), Israel, and Hubei (a province in China),

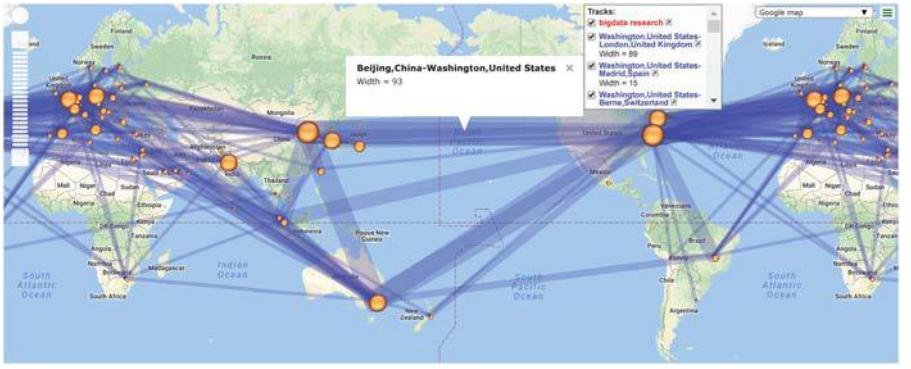

Figure 1.
The geographical international collaboration network of 66 nations in Big Data research

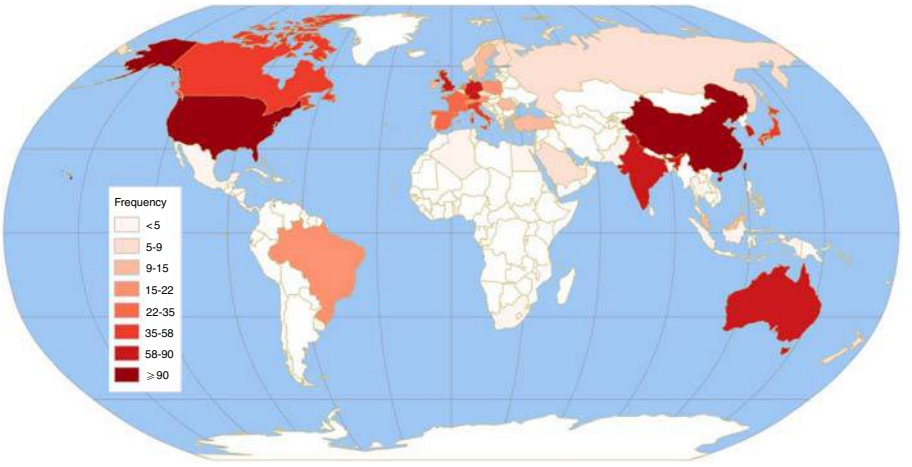

Figure 2.
The geographical choropleth map of 54 nations with number of occurrences greater than 2

as there are 38 nations with 86.5 percent of the number of occurrences. Nations south of Latitude 30 contribute only 13.5 percent of Big Data research, with only Australia (60) and India (60) considered major nations for Big Data research, followed by Taiwan (19), Singapore (18), and Brazil (16); the other nations all make fewer contributions to Big Data Research. In terms of contributions and developed degrees of Big Data research, Latitude 30 could be considered the geographic watershed.

4.2.2 Research community structure and patterns. In this study, five research communities are identified in the largest component (54 of the 66 nations) and distinguished by the different colors and sizes of their nodes and links. As shown in Figure 3, the five communities are the US group (red), China group (purple), South Korea group (blue), Italy group (green), and Brazil group (yellow). Figure 4 displays both the connections among research communities and the collaborations among nations within each community. The node size is proportionate to the number of related papers, and the width of links is proportionate to the number of co-occurrences for each pair of nations.

The USA, China, the UK, Australia, Canada, and Germany, with a higher degree of co-occurrence, are presented in Figure 3. These nations have larger nodes and more numerous links with others, indicating their importance and powerful capacity for Big Data research. The pattern of collaborations within each community being closer than that between communities is again identified.

The inner collaboration network of each research community is also shown in Figure 3, which more clearly distinguishes the groupings. The USA, China, Italy, South Korea, and Brazil are, respectively, central in each community network. All the other nations directly or indirectly connect to these nations. Note that the US group has the most intensive Big Data

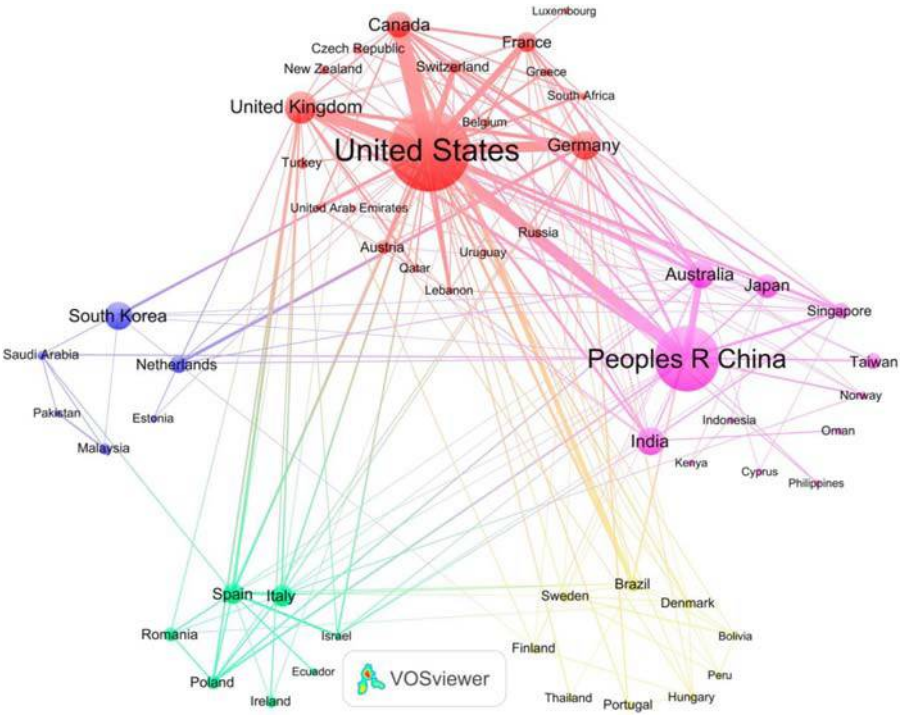

Figure 3.
Five communities of nations according to their collaborations

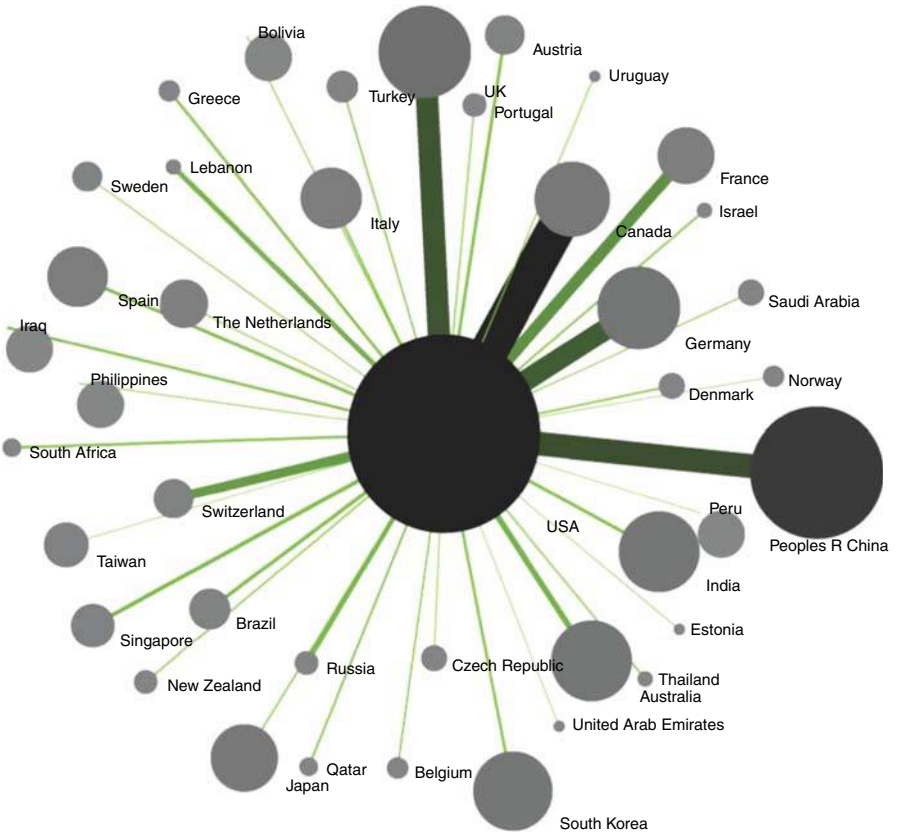

Figure 4.
The individual ego network of the USA collaborating with others

research, followed by the China group, while the Italy, South Korea, and Brazil groups seem sparser, with a smaller number of links. However, the high density in these three small research communities indicates intensive inner collaborations; conversely, the low density in the two large communities equals looser inner collaborations on the whole.

The international collaboration network among the 54 nations allows us to extract submaps reflecting the nearest neighbors of a single nation for the purpose of observing the neighboring nations with which it collaborates. Figure 4, for example, shows the individual ego network of the USA within the larger network. It can be found that Canada is their primary partner in Big Data research, followed by China, the UK, Germany, France, and other neighboring nations.

Figure 5 supplements this by showing the distribution of five research communities using a geographical choropleth map. The USA and China groups are relatively concentrated geographically, mainly located in North America, Europe, and Asia-Pacific. In contrast, the Brazil group is the most distributed, in South America, Europe, and Asia.

5. Discussion and conclusions

5.1 Findings and explanations

The analysis shows the international collaboration structure and patterns in Big Data research, including basic descriptive statistics of nations contributing to

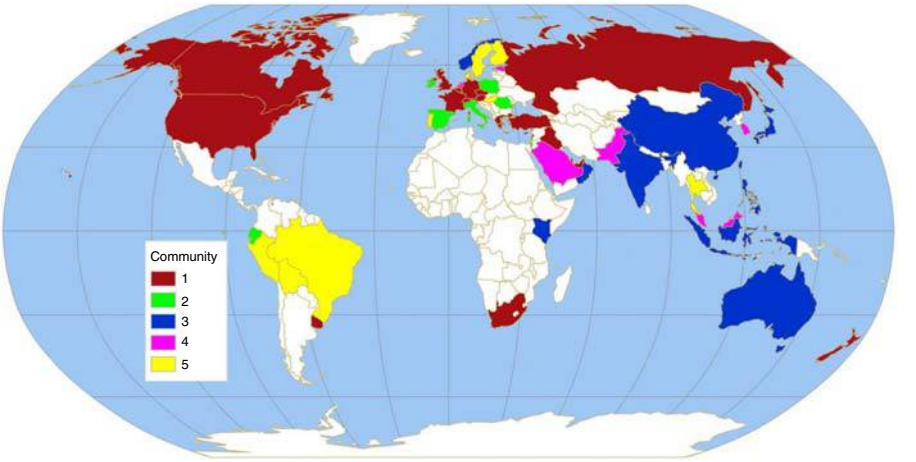

Figure 5.
The distribution of five research communities on a geographical map

Big Data research in recent years, network structure of collaborations among nations, patterns of collaboration communities, and visualized topological and geographical maps that vividly display the global collaboration structure and patterns. These new and comprehensive perspectives facilitate a greater understanding of research collaborations in Big Data as a global phenomenon.

First, Big Data research involves a large number of nations with very broad collaborations. This distribution of nations related to Big Data is greatly unbalanced, as are the collaborations. Therefore, nations could be divided into different, greatly fragmented levels. The USA, China, and the UK contribute the most to this research, and comprise a large proportion of the collaborations. However, on the whole, this is not intensive collaboration; most nations are more likely to directly collaborate with several powerful nations in the frontiers of Computer Science or Information Science, which could support a variety of Big Data applications in many nations (Chen *et al.*, 2012). There are many frontiers where big data may be used to improve societies, and we may find ourselves able to make leaps-and-bounds advances that would not otherwise be possible.

Second, this study identifies five intensive research communities of nations. The USA, Italy, China, South Korea, and Brazil could be considered generally equivalent to leaders of regional collaboration communities related to Big Data research. Generally, while collaborations among nations in Big Data research are not intensive on the whole, those within communities are relatively intensive. These close connections among nations aggregated into one research community indicate some unity of their objectives in Big Data research and applications. In addition, the USA and China communities are the largest, with a high number of collaborations; they are the two most important research communities in the field of Big Data research.

Third, the geographical maps, including the collaboration network and choropleth maps, help us better understand these collaborations at the global level. In this study, nations related to Big Data research are clearly divided geographically. Nations with more contributions and collaborations are concentrated in North America, Asia-Pacific, and Western Europe. Larger communities tend to be concentrated geographically, while small ones are relatively scattered. Given the main contributors, Big Data research and collaborations are intimately related to a nation's economic development level, as well as its recognition of the importance of Big Data. As mentioned above, the USA, European Union, and China have invested great effort in their research and applications of Big Data

(Emani *et al.*, 2015). The findings of this study show and confirm that these countries have been key players in research in this area.

In addition, this study utilized geographical visualization and social network analysis to examine the collaborations in Big Data research at the national level, which is a different approach from previous studies' scientometric analysis of research work (e.g. Singh *et al.*, 2015), which was mainly based on authorships and country-level collaboration patterns and on major contributors (countries, institutions, and individuals) in the field. In this study, visual maps in topological and geographical views were generated to reveal the structure and patterns of international collaborations. Providing geographical network maps demonstrating the collaborative relationships among nations in Big Data research allowed us to directly and vividly illustrate the collaboration structure and patterns on the global level. The geographical choropleth and proportional symbol maps were provided to display the distribution of national contributions in Big Data research, which could also enable greater understanding of the status of a nation's research, as well as the collaboration communities across the geographical map. Moreover, based on geographical visualization methods and tools, the topological network map is successfully laid over the geographic map while retaining the size of the nodes and links representing nations and their collaborations, which are enhancements of previous studies (e.g. Leydesdorff *et al.*, 2014).

Finally, it should be noted that the sample articles used in this study were retrieved from the WoS Core Collection database, and the results of this study should be interpreted within the limitations of the database coverage. Future research could add samples of articles from other databases for complete coverage. Nevertheless, this study and its findings can serve as an example for analyzing international collaborations of one research field from topological and geographical perspectives, on the basis of tools and methods mentioned and illustrated in this paper.

5.2 Implications

This study facilitates the collaboration analysis of nations through geographical network and social network analyses, and explores the various structures and patterns of collaboration networks. The revelation and demonstration of collaboration patterns between countries could provide better understanding of Big Data related research efforts for researchers, institutions, and governments as many governments and institutions have promoted Big Data as a national strategy and emphasized international collaborations in the area. Examining and revealing the structure and patterns of international collaboration at the national-level contributes to the understanding of Big Data research as a national as well as a global phenomenon.

Future research can build on this study to shed light on the reasons, meanings, and practical implications of findings from this study beyond the scientometric context and examine them in social, political, and policy-related contexts. Continuous research in the international collaboration of Big Data research, from the perspectives of social informatics and human geography, would further illuminate the trends, facilitators, and barriers nations face when collaborating. In the meantime, it is also important to understand collaboration in Big Data research from other perspectives, such as by author, institution, discipline, topic, source of publication, or other bibliographic fields. Given the evolving nature of Big Data research, future studies should explore possible changes and evolutions in these structures based on the longitudinal data. Finally, to expand the understanding of Big Data research patterns and trends, government funding and policies should be included as part of the analysis. These findings may help governments and agencies understand the political and economic implications, evaluate their policies, and make informed decisions accordingly as they compare and relate Big Data research activities with those of peer nations.

References

Adar, E., Kim, M. and Soc, I.C. (2007), "SoftGUESS: visualization and exploration of code clones in context", paper presented at the 29th International Conference on Software Engineering (ICSE 2007), Minneapolis, MN, May 20-26, available at: http://skabber.nir.com.ua/files/icse2007-final.pdf (accessed May 30, 2016).

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, pp. P10008-P10008, doi: 10.1088/1742-5468/2008/10/P10008.

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

Borner, K., Chen, C.M. and Boyack, K.W. (2003), "Visualizing knowledge domains", *Annual Review of Information Science and Technology*, Vol. 37 No. 1, pp. 179-255.

Bornmann, L. and Leydesdorff, L. (2012), "Which are the best performing regions in information science in terms of highly cited papers? Some improvements of our previous mapping approaches", *Journal of Informetrics*, Vol. 6 No. 2, pp. 336-345.

Bornmann, L. and Ozimek, A. (2012), "Stata commands for importing bibliometric data and processing author address information", *Journal of Informetrics*, Vol. 6 No. 4, pp. 505-512.

Casado, R. and Younas, M. (2015), "Emerging trends and technologies in big data processing", *Concurrency and Computation-Practice & Experience*, Vol. 27 No. 8, pp. 2078-2091.

Catala-Lopez, F., Alonso-Arroyo, A., Aleixandre-Benavent, R., Ridao, M., Bolanos, M., Garcia-Altes, A., Sanfelix-Gimeno, G. and Peiro, S. (2012), "Coauthorship and institutional collaborations on cost-effectiveness analyses: a systematic network analysis", *Plos One*, Vol. 7 No. 5, p. e38012, doi: 10.1371/journal.pone.0038012.

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 No. 11, pp. 314-347.

Chen, C.M. (2006), "CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 3, pp. 359-377.

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 and Applications*, Vol. 19 No. 2, pp. 171-209.

Coulter, N., Monarch, I. and 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*, Vol. 49 No. 13, pp. 1206-1223.

Ding, Y., Chowdhury, G.G. and Foo, S. (2001), "Bibliometric cartography of information retrieval research by using co-word analysis", *Information Processing and Management*, Vol. 37 No. 6, pp. 817-842.

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.

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., Wang, C.J., Daneshmand, M., Wang, C. and Wang, H. (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.

George, G., Haas, M.R. and Pentland, A. (2014), "Big data and management", *Academy of Management Journal*, Vol. 57 No. 2, pp. 321-326.

Grauwin, S. and Jensen, P. (2011), "Mapping scientific institutions", *Scientometrics*, Vol. 89 No. 3, pp. 943-954.

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S. and Porter, J.H. (2013), "Big data and the future of ecology", *Frontiers in Ecology and the Environment*, Vol. 11 No. 3, pp. 156-162.

Hellsten, I., Lambiotte, R., Scharnhorst, A. and Ausloos, M. (2007), "Self-citations, co-authorships and keywords: a new approach to scientists' field mobility?", *Scientometrics*, Vol. 72 No. 3, pp. 469-486.

Hitzler, P. and Janowicz, K. (2013), "Linked data, Big Data, and the 4th paradigm", *Semantic Web*, Vol. 4 No. 3, pp. 233-235.

Kalil, T. (2015), "Big data is a big deal", March 29, available at: www.whitehouse.gov/blog/2012/03/29/big-data-big-deal (accessed May 30, 2016).

Khan, G.F., Moon, J. and Park, H.W. (2011), "Network of the core: mapping and visualizing the core of scientific domains", *Scientometrics*, Vol. 89 No. 3, pp. 759-779.

Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Ali, W.K.M., Alam, M., Shiraz, M. and Gani, A. (2014), "Big data: survey, technologies, opportunities, and challenges", *The Scientific World Journal*, Vol. 2014 No. 712826, doi: 10.1155/2014/712826.

Kitchin, R. (2014), "Big Data, new epistemologies and paradigm shifts", *Big Data & Society*, Vol. 1 No. 1, pp. 1-12, doi: 10.1177/2053951714528481.

Kretschmer, H. (2004), "Author productivity and geodesic distance in bibliographic co-authorship networks and visibility on the web", *Scientometrics*, Vol. 60 No. 3, pp. 409-420.

Kristensen, P.M. (2015), "Revisiting the 'American social science' – mapping the geography of international relations", *International Studies Perspectives*, Vol. 16 No. 3, pp. 246-269.

Leydesdorff, L. and Bornmann, L. (2011), "Integrated impact indicators compared with impact factors: an alternative research design with policy implications", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 11, pp. 2133-2146.

Leydesdorff, L. and Bornmann, L. (2012), "Mapping (USPTO) patent data using overlays to Google Maps", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 7, pp. 1442-1458.

Leydesdorff, L. and Rafols, I. (2011), "Local emergence and global diffusion of research technologies: an exploration of patterns of network formation", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 5, pp. 846-860.

Leydesdorff, L. and Welbers, K. (2011), "The semantic mapping of words and co-words in contexts", *Journal of Informetrics*, Vol. 5 No. 3, pp. 469-475.

Leydesdorff, L., Park, H.W. and Wagner, C. (2014), "International coauthorship relations in the social sciences citation index: is internationalization leading the network?", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 10, pp. 2111-2126.

Ma, H., Fang, C., Pang, B. and Wang, S. (2015), "Structure of Chinese city network as driven by technological knowledge flows", *Chinese Geographical Science*, Vol. 25 No. 4, pp. 498-510.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. (2011), "Big data: the next frontier for innovation, competition, and productivity", report, McKinsey Global Institute, available at: www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (accessed May 30, 2016).

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.

Murdoch, T.B. and Detsky, A.S. (2013), "The inevitable application of big data to health care", *Journal of the American Medical Association*, Vol. 309 No. 13, pp. 1351-1352.

Nandi, A., Yu, C., Bohannon, P. and Ramakrishnan, R. (2012), "Data cube materialization and mining over MapReduce", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24 No. 10, pp. 1747-1759.

Nature Publishing Group (2008), "Big Data", available at: www.nature.com/news/specials/bigdata/index.html (accessed May 30, 2016).

Newman, M.E.J. (2004), "Coauthorship networks and patterns of scientific collaboration", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101 No. S1, pp. 5200-5205.

Nooy, W., Mrvar, A. and Batagelj, V. (2005), *Exploratory Social Network Analysis with Pajek*, Cambridge University Press, New York, NY.

Ronda-Pupo, G.A. and Guerras-Martin, L.Á. (2012), "Dynamics of the evolution of the strategy concept 1962-2008: a co-word analysis", *Strategic Management Journal*, Vol. 33 No. 2, pp. 162-188.

Savitz, E. (2012), "Gartner: 10 critical tech trends for the next five years", October 22, available at: www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years (accessed May 30, 2016).

Science/AAAS (2011), "Special online collection: dealing with big data", available at: www.sciencemag.org/site/special/data (accessed May 30, 2016).

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.

Small, H. and Griffith, B.C. (1974), "The structure of scientific literatures I: identifying and graphing specialties", *Science Studies*, Vol. 4 No. 1, pp. 17-40.

Tonidandel, S., King, E.B. and Cortina, J.M. (2016), "Big Data methods: leveraging modern data analytic techniques to build organizational science", *Organizational Research Methods*, doi: 10.1177/1094428116677299.

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.

Wamba, S.F., Akter, S., Edwards, A., Chopin, G. and 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*, Vol. 165, pp. 234-246, doi: 10.1016/j.ijpe.2014.12.031.

Wu, X., Zhu, X., Wu, G.-Q. and Ding, W. (2014), "Data mining with Big Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26 No. 1, pp. 97-107.

Yin, L., Kretschmer, H., Hanneman, R.A. and Liu, Z. (2006), "Connection and stratification in research collaboration: an analysis of the COLLNET network", *Information Processing and Management*, Vol. 42 No. 6, pp. 1599-1613.

Further reading

Leydesdorff, L. and Persson, O. (2010), "Mapping the geography of science: distribution patterns and networks of relations among cities and institutes", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 8, pp. 1622-1634.

Corresponding author
Yin Zhang can be contacted at: yzhang4@kent.edu