A comparative study on big data research in China and the USA

Lv, X.; Zhou, P.

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

A comparative study on big data research in China and the USA [1]

Xiaozan Lv*, and Ping Zhou*

*lvxz1991@zju.edu.cn; pingzhou@zju.edu.cn
Department of Information Resources Management, Zhejiang University, No.866 Yuhangtang Road, Hangzhou, 310058 (China)

Abstracts
Based on publications indexed in Web of Science (WoS) of Clarivate from 2009-2015, this paper presents a comparative analysis of big-data related research produced by the USA and China in terms of publication productivity, interdisciplinarity, funding support and research topics from bibliometric perspectives. With exponential growth, both China and the USA have kept the leading positions in publication productivity. The year 2012 can be considered as a turning point of the USA when big data R&D became a national strategy. Although staring later, China develops faster and eventually caught up with the USA in 2015. The US researchers outperform their Chinese colleagues in high-quality papers as measured by the proportions of top-1% and top-10% most-frequently cited publications, and, indeed, publish in broader research areas. With a substantially higher percentage of funded publications, China has a higher degree of government involvement in big data research. Major funding agencies, the National Natural Science Foundation of China (NSFC) and the National Science Foundation (NSF) of the USA, for instance, have been able to provide specific support to big data research in multiple areas. Despite some differences related to big-data applications and certain social problems, China and the USA share similar research topics in core methods and technologies.

Keywords: Big data; Bibliometrics; the USA; China

Introduction

Emerged in the 1990s and gained momentum in the early 2000s, the term "big data" is originally used to describe datasets whose size is beyond the ability of traditional means to handle. However, the scope of the term has significantly expanded over the years, not only refers to the data itself, but also a set of different entities including - but not limited to - social phenomenon, analytical or storage technologies, processes and infrastructures (Mauro et al., 2015).

Nowadays, big data has played a major role in various domains such as science, research, engineering, medicine, healthcare, finance, business, and ultimately society itself (O'reilly Media, 2012). As "the new oil", it is an endless source of data for the economic and social world (Hasnat, 2018). In 2017, the annual global revenue for big data market has reached $33.5 billion and is expected to be doubled in the next four years[2]. More leading technology

[1] This study receives funding from National Natural Science Foundation of China (grant number: 71473219).
[2] Data source: https://www.statista.com/topics/1464/big-data/. Retrieved on 21st Jan, 2018.

companies, such as IBM, SAP, Oracle, Hewlett Packard and Accenture, have taken the initiative to apply big data applications and offer related services.

The potential value of big data has also attracted strong support from governments across the world. The United States first announced the *Big Data Research and Development Initiative*[3] in 2012, committing more than $200 million to big data research projects. Many other countries or regions jointed the campaign in the coming years. For example, the *Digital Agenda for Europe and Challenges for 2012*[4], the *UK data capability strategy: seizing the data opportunity*[5], the *Australian Public Service Big Data Strategy*[6], and the *Planning for the Development of Big Data Industry (2016-2020)*[7] of China.

In academic community, big data related publications from different areas has a dramatic growth in recent years (Porter et al., 2015). Originated from computer science, big data has its legacy in information technology developments, with research topics various from theoretical models and algorithms to core technologies and processing. However, its capacity and analytic capabilities promise to make an essential contribution in areas such as traffic management, logistics, health care, and education (e.g., Hazen et al., 2014; Mauro et al., 2015; Matthew et al., 2012; Smith et al., 2012). 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).

In the context of big data campaign, it is worthwhile to map the overall research status of the world, especially activities of the leading nations like China and the USA, so as to develop more efficient and more targeted research plans. The current paper will contribute in this regard by providing the overall publication production, interdisciplinarity, funding support and research topics of China and the USA in the period of 2009 - 2015.

Data and method

The data used in this paper are harvested from the Web of Science (WoS) of Clarivate, including the Science Citation Index Expanded (SCIE), Social Science Citation Index (SSCI), Arts & Humanities Citation Index (A&HCI), Conference Proceedings Citation Index—Science (CPCI-S), and Conference Proceedings Citation Index - Social Science & Humanities (CPCI-SSH).

We retrieved publications (i.e. “Article”, “Review” and “Proceeding papers”) related to big data in either China or the USA with the query: PY=2009-2015 and CU=(“Peoples R China” or USA) and TS=(“big data” or “bigdata” or “huge data” or “large scale data” or “large-scale data” or “massive data”). The period between 2009 and 2015 was selected due to the fact that research on big data began very recently and has gained momentum only during the last few years. Ultimately, this approach resulted in 6,502 records[8], of which 3,247 containing China (Chinese papers) and 3,619 containing the USA (the US papers).

Relevant questions will be investigated from perspectives including publication productivity, interdisciplinarity, funding and research topics. In terms of publication productivity, both

[3] https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal
[4] http://europa.eu/rapid/press-release_MEMO-10-200_en.htm
[5] https://www.gov.uk/government/publications/uk-data-capability-strategy
[6] https://www.finance.gov.au/files/2013/06/Draft-Big-Data-Strategy.pdf
[7] http://www.miit.gov.cn/n1146295/n1652858/n1653018/c5465700/content.html
[8] Retrieved on January 13, 2017.

absolute and relative measurements will be used. As big data involves multiple disciplines, we will investigate whether the number of research areas changes with time and whether differences exist between the two countries. Research topics are based on the keywords of each paper.

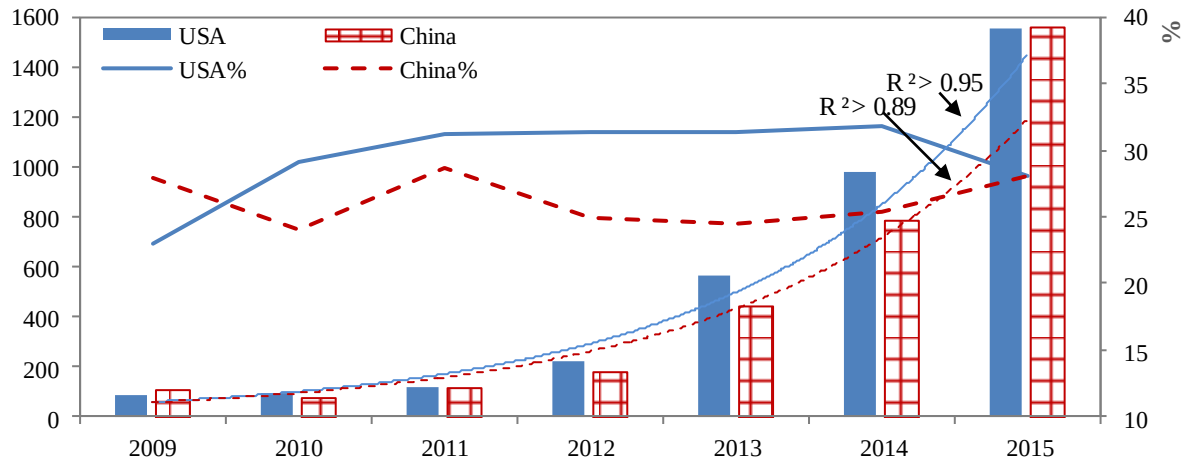
Results

Comparison between China and the USA will be focused on publication productivity including overall situation, highly-cited papers, interdisciplinarity, funding and research topics.

Publication Productivity

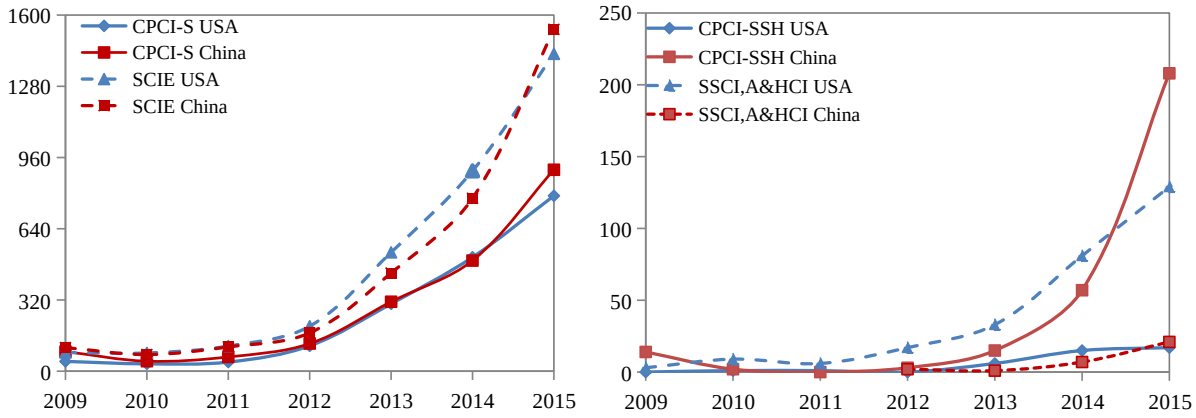
From 2009 to 2015, both China and the USA showed an exponential growth trend in the number of publications - this is especially true since 2012, as a notable rise can be observed. For the USA, the number of publications jumped from 87 in 2009 to 1,555 in 2015, surpassing that of China in most years. However, the output of China grew faster and finally caught up with the USA in 2015, reaching 1,557. As the main contributors to publication output, the USA and China no doubt occupied the most dominant positions in total big data publications in WoS, as the combined average share of the USA or/and China accounts for over 50% (29.4% and 26.2%, respectively) (Figure 1).

Figure 1: Annual publications (SCIE, CPCI-S, SSCI, CPCI-SSH, A&HCI)



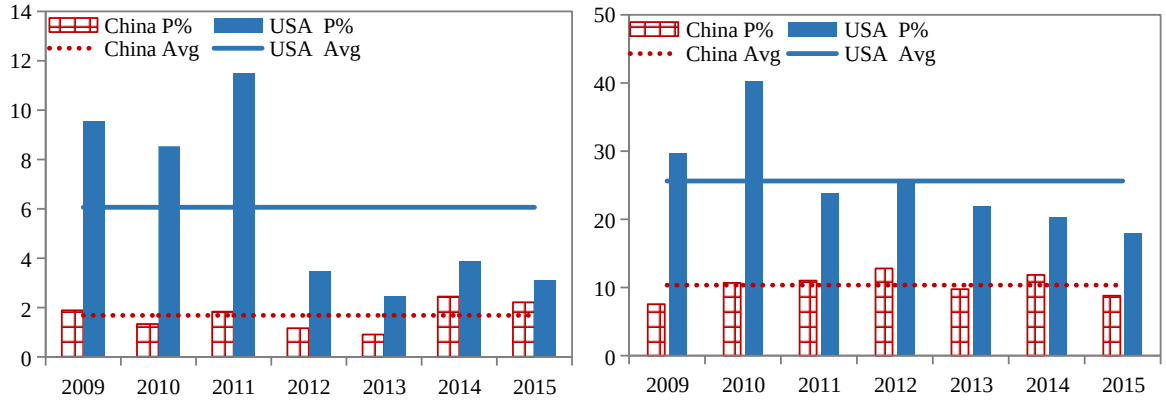
With regards to the publications indexed in different data sources, China and the USA shared similar situations for SCIE and CPCI-S until 2015, when China surpassed the USA. In the Social Sciences and Arts & Humanities (SSCI, A&HCI and CPCI-SSH), China falls far behind the USA in both its starting point and productivity of journal publications, and the gap continues to grow. However, with rapid growth since 2012, China published significantly more conference papers (Figure 2).

Figure 2: Publications indexed in different sources



As for highly-cited papers, we focus on SCIE and CPCI-S, since the two databases have covered over 95% of the publications in current research. With, respectively, 6.1% and 25.6% of top-1% and top-10% highly cited publications, the USA performs much better than China (1.7% and 10.3%). However, China has been able to improve its proportions, though slightly, in the top segments while the USA shows a pattern of decline (Figure 3).

Figure 3: Top-1% (left) and top-10% (right) papers (SCIE & CPCI-S)

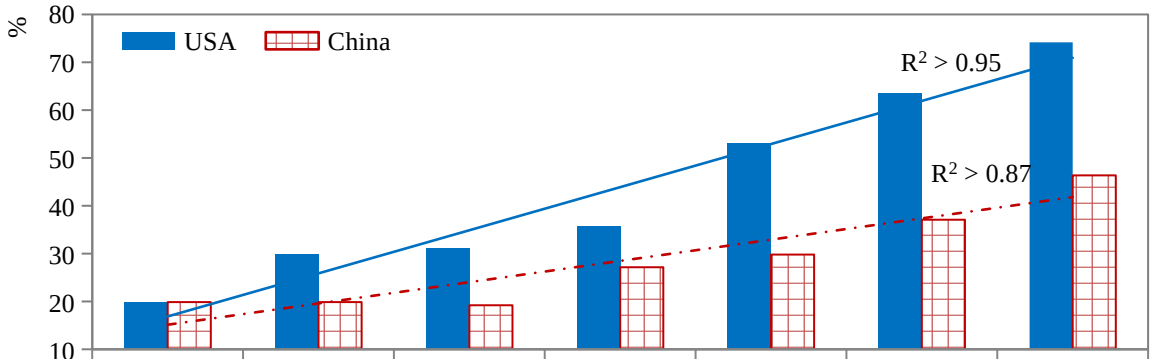


There are various reasons for why 2013 witnessed such a great leap forward in terms of big data research. As previous research (Huang and Zhang, 2016) points out, the important reason is likely to be the 2012 "*Big Data Research and Development Initiative*", with subsequent support from the National Science Foundation (NSF). Influenced by the USA to some extent, the Chinese government later proposed its own "*Big Data Research and Development Program*", with a total investment in the tens of billions of dollars for the construction of relevant projects. These actions led to a number of universities, companies, research institutes, and even multiple government agencies becoming encouraged to take part in big-data related research, contributing to the increase in paper output.

Interdisciplinarity

Research areas are identified based on the Subject Categories (SCs) of each paper. Generally, the annual number of SCs is increasing in both countries with the same starting point of 30 in 2009. By 2015, the USA had 1.6 times as many as China, reaching 112 in total. Additionally, we can see that the US shows rapid expansion over a wide range, and has covered over 70% of all the subject categories in WoS (151) up to 2015, while the proportion in China is only 46.4%. More concretely, growth rates in the proportions of SCs have a linear increase in both countries, with R^2>0.95 and 0.87 respectively, indicating that the US growth rate is faster (Figure 4).

Figure 4: Proportions of SCs in all research areas in WoS

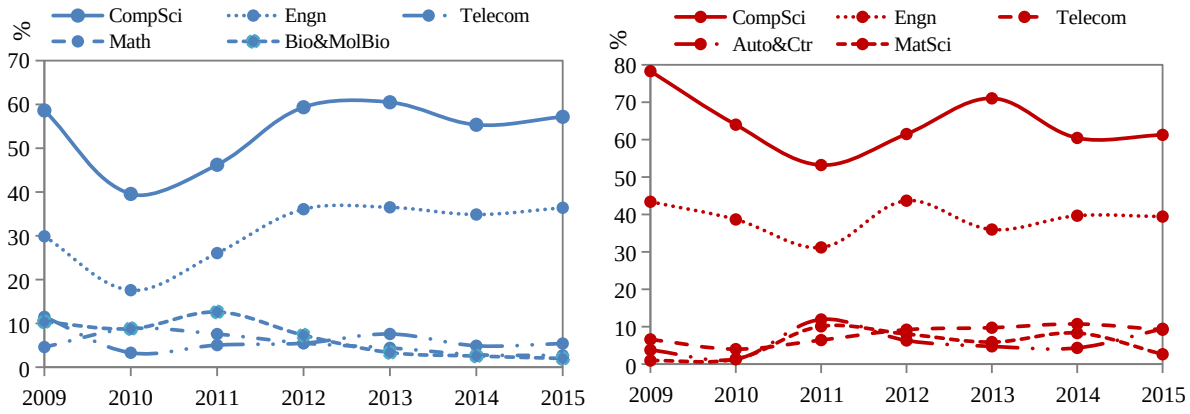


	2009	2010	2011	2012	2013	2014	2015
USA	19.9	29.8	31.1	35.8	53.0	63.6	74.2
China	19.9	19.9	19.2	27.2	29.8	37.1	46.4

The highest ratio of publications (72.8% in total) contributed to two research areas: Computer Science (CompSci) and Engineering (Engn), receiving more than 76% of all the US and 79% of all the Chinese papers. The second-tier of areas, which indicates at least 100 published papers, are Telecommunications (Telecom), Mathematics (Math), and Biochemistry & Molecular Biology (Bio&MolBio) in the USA and Telecommunications (Telecom), Automation & Control Systems (Auto&Ctr), and Materials Science (MatSci) in China. In addition, Optics, Physics, Education & Educational Research, Public Administration, and Information Science & Library Science are also key contributors.

In terms of the annual shares of the top-5 research areas, in the USA, the shares of CompSci and Engn are increasing while the others have decreased over time. However, China has the opposite case. Except for a slight decline in CompSci, the proportions of papers in the other four areas have shown a relatively stable, or even rising, trend (Figure 5).

Figure 5: Shares of top-5 SCs in the USA (left) and China (right)*



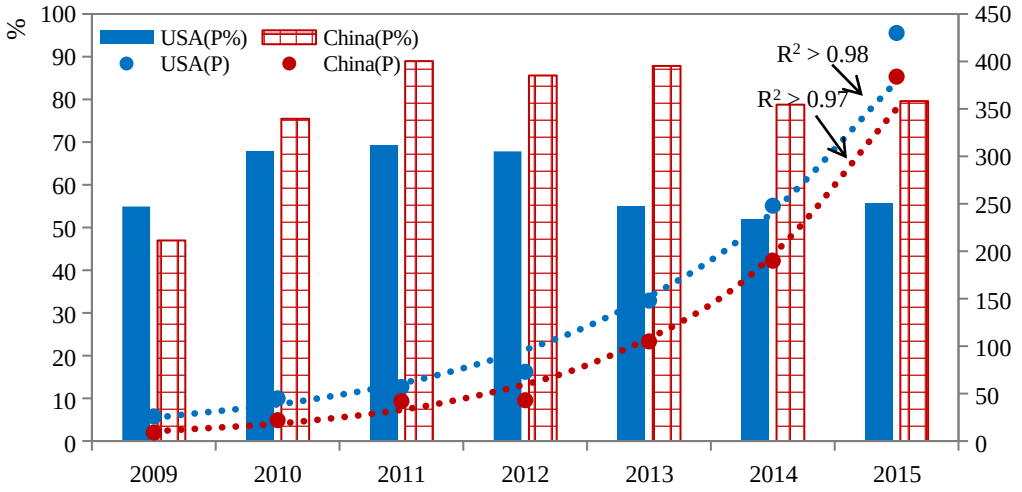
*A paper could have more than one research area, hence the total values are all greater than 100%.

Funding Support

Over the studied period (2009-2015), funded publications (SCI-E, SSCI, A&HCI) of China and the USA have shown exponential growth (R^2 >0.98 and 0.97, respectively), whereas China has grown faster. In addition, the average proportion of Chinese-funded papers has reached 80.8%, significantly higher than that of the US (56.7%). It can be seen that, despite

the rapid growth in absolute numbers of the two countries, the proportions has changed little, while the USA has experienced a downward trend since 2012 (Figure 6).

Figure 6: Numbers and shares of funded papers (SCIE, SSCI, A&HCI)



Among the funding sources, the National Natural Science Foundation of China (NSFC) takes the leading position in China, followed by the Ministry of Science and Technology (MOST) and the Ministry of Education (MOE). These three agencies (NSFC, MOST, MOE) have contributed to over 80% of funded publications in China. Moreover, the U.S. National Science Foundation (NSF) and Chinese Academy of Science (CAS) have also contributed significantly, together supporting about 9.3% of papers. As for the USA, the top-5 funding agencies are the NSF, the National Institutes of Health (NIH), NSFC, the U.S. Department of Energy (DOE), and MOST. The first two (NSF and NIH) hold the primary position, with a total 54.9% of funded papers receiving funds from them (28.4% and 28.0%, respectively), the remainder supported nearly 20% in all (Table 1).

Table 1. Top-5 funding sources in China and the USA

Country	Funders	Num	P%	P%	Num	Funders	Country
China	NSFC	544	68.4	28.4	292	NSF	USA
	MOST	255	32.1	28.0	288	NIH	
	MOE	197	24.8	9.8	101	NSFC	
	NSF	38	4.8	5.6	58	DOE	
	CAS	36	4.5	4.7	48	MOST	

We also find that funding is closely associated with international collaboration, as the NSF, NSFC and MOST are in each other's main funding sources list, and most of the NSF-funded Chinese papers and NSFC- or MOST-funded American papers are China-USA collaboration papers.

As a technology dominated field, it is reasonable that Computer Science and Engineering would account for the largest proportion of funding support in big data research. The phenomenon is more obvious in China, as more than 90% of NSFC-funding flows to these two fields. At the same time, there are distinctions in that the NSF-funding is more related to application areas, such as Biochemistry & Molecular Biology, Information Science & Library

Science, and Geography, while the NSFC shows more emphasis on technical and management domains.

Table 2. Research areas of funded papers in SCIE, SSCI, A&HCI (top-5)

SCIE				
	NSF-funded	**P%**	**P%**	**NSFC-funded**
1	Computer Science	48.6	64.5	Computer Science
2	Engineering	24.7	31.8	Engineering
3	Mathematics	11.1	16	Telecommunications
4	Biochemistry & Molecular Biology	7.3	6.6	Mathematics
5	Science & Technology - Other Topics	7.3	4.6	Science & Technology - Other Topics
SSCI, A&HCI				
	NSF-funded	**P%**	**P%**	**NSFC-funded**
1	Computer Science	34.2	36.4	Computer Science
2	Information Science & Library Science	21.1	13.6	Business & Economics
3	Science & Technology - Other Topics	15.8	13.6	Science & Technology - Other Topics
4	Geography	13.2	9.1	Mathematics
5	Engineering	13.2	9.1	Operations Research & Management Science

Research Topics

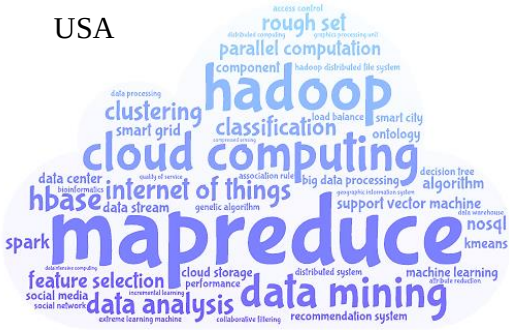
High-frequency keywords (top-50) are analyzed to determine the major research topics in big data research. Considering the differences among different research fields, we distinguish keywords based on the sources of data.

These results indicate that the USA and China have a considerable amount of shared research interests and focus on topics of core methods and technologies in SCIE and CPCI-S, such as “mapreduce”, “cloud computing”, “data mining”, “hadoop”, and “data analysis” (Figure 7).

In SSCI, A&HCI, and CPCI-SSH, the distinctions are obvious. The US concentrates more on social media as the main dimension affected by big data technologies and applications, such as “Internet” and “Twitter”. Additionally, data security, especially “privacy”, is a hot topic among both American and Chinese scholars, though “Internet finance” and “e-commerce” are more popular in China (Figure 8).

Figure 7: High-frequency keywords in SCIE, CPCI-S (top-50)

USA



China

Figure 8: High-frequency keywords in SSCI, A&HCI, CPCI-SSHS (top-50)

USA

China

Discussion and Conclusion

During the investigated time span of the present study, with exponential growth, both China and the USA have kept the leading positions in publication productivity. The year 2012 is a turning point of the USA when big data R&D became a national strategy. Although staring later, China develops faster and eventually caught up with the USA in 2015. As a technology-led field, the natural sciences (SCIE and CPCI-S), such as computer science and engineering, are reasonable sources for the main contributions. Although China surpassed the USA in paper output by 2015, the US researchers outperform their Chinese colleagues in high-quality papers as measured by the proportions of top-1% and top-10% most-frequently cited publications. In some ways, we may say that China is still in the stage of quantitative increment and is less capable of improving publication quality.

Strong support for big data research from governments can be reflected in levels of funding. Over 80% of all the Chinese and 56% of the US journal papers have received at least some funding. Currently, governmental agencies are the predominant funding agencies of big data research. Thus, it may indicate a higher degree of government involvement in big data research in China. Some major funding agencies, NSFC, MOST, MOE in China and NSF, NIH, DOE in the USA, for instance, have been able to provide specific support to big data research in multiple areas.

Big Data research involves an increasing number of research areas, of which the two largest are Computer Science and Engineering. As shown in previous studies, computer science is the fundamental discipline in big data research, providing methods and techniques for processing large-scale data sets (Chen et al. 2014), which are widely applied in certain fields as Engineering, Medicine, and Business (Khan et al. 2014). In contrast to the USA, China's big

data research area still has much empty and potentially developed space, for instance, General & Internal Medicine, Immunology, Nursing, and Psychiatry, which can be found in the US papers. To be specific, despite some common research interests in certain methods and techniques, e.g. “mapreduce”, “cloud computing”, “data mining”, “hadoop”, and “machine learning”, certain social problems have also been popular topics in related research. For example, “social media” attracted more focus from the US, while “e-commerce” attracted more from the Chinese, and “privacy” is a common one.

Above all, although both China and the USA are constantly advancing relevant research in big data, the USA, as the leading country in science and technology, has a head start. These differences are mainly reflected in the quality of publications, subsidized research areas, and research topics. Although the output of Chinese publications is quite large and has increased with time, basic technology and methods with limited applications are the main focus of the Chinese researchers. China perhaps should pay more attention to improving publication quality, increasing international cooperation with advanced countries, such as the USA, could be an effective breakthrough (Abramo et al., 2009). On the other hand, with the continuous development of big data technologies, questions of how to effectively use big data to develop its potential value and expand its application areas also deserve the attention of both the US and Chinese scholars.

References

Abramo G, D'Angelo C A, Di Costa F. (2009). Research collaboration and productivity: Is there correlation?[J]. *Higher Education, 57*(2), 155-171.

Chen H, Chiang R H L, Storey V C. (2012). Business intelligence and analytics: from big data to big impact[J]. *Mis Quarterly, 36*(4):1165-1188.

Chen C L P, Zhang C Y. (2014). Data-intensive applications, challenges, techniques and technologies: a survey on big data[J]. *Information Sciences, 275*(11), 314-347.

Chen M, Mao S, Liu Y. (2014). Big data: a survey[J]. *Mobile Networks & Applications, 19*(2), 171-209.

FANG H, ZHANG Z, WANG C J. (2015). A survey of big data research[J]. *IEEE Network, 29*(2011): 6–9.

HASNAT B. (2018). Big Data : An Institutional Perspective on Opportunities and Challenges[J]. *JOURNAL OF ECONOMIC ISSUES, LII*(2): 580–589.

Hazen B T, Boone C A, Ezell J D, et al. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications[J]. *International Journal of Production Economics, 154*(4), 72-80.

Huang M H, Huang M J. (2018). An analysis of global research funding from subject field and funding agencies perspectives in the G9 countries[J]. *Scientometrics, 115*(2), 1-15.

Khan N, Yaqoob I, Hashem I, et al. (2014). Big data: Survey, technologies, opportunities, and challenges[J]. *The Scientific World Journal*, 2014.

Matthew S, Christian S, Benjamin H, et al. (2012). Big Data privacy issues in public social media[C]. *IEEE International Conference on Digital Ecosystems Technologies (DEST)*, 1–6.

MARX V. (2013). The big challenges of big data[J]. *Nature, 498*: 255–260.

Mauro A D, Greco M, Grimaldi M. (2015). What is big data? a consensual definition and a review of key research topics[C]. *International Conference on Integrated Informationm, 1644*, 97-104.

O'Reily Media. Big Data Now: 2012 Edition[R]. *Sebastopol, CA: O'Reilly Media*, 2012.

Porter A L, Huang Y, Schuehle J, et al. (2015). Meta Data: Big Data Research Evolving across Disciplines, Players, and Topics[C]. *IEEE International Congress on Big Data*, 262-267.

Smith M, Szongott C, Henne B, Voigt G V. (2012). Big data privacy issues in public social media[C]. *IEEE International Conference on Digital Ecosystems Technologies*, 1-6.