



A decade of big data literature: analysis of trends in light of bibliometrics

Iftikhar Ahmad¹ · Gulzar Ahmed² · Syed Adeel Ali Shah¹ · Ejaz Ahmed³

Published online: 7 December 2018

© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

Bibliometrics is a quantitative tool for the analysis of literature published in a scientific field. Using Scopus as the data source, we perform a thorough analysis of scholarly works published in the field of big data from 2008 to 2017. The objective of the work is to find the most cited articles in the given time frame, the citation trends, the authorship trends as well as the trends of research work in the related area. The analysis shows that over 50% of publications do not receive any citations, and the average number of citations per publication is 3.17. It is also observed that single authorship of research publications has declined over the time. The analysis reveals the pioneering role played by the USA in advancing the research in big data, which has lately been taken over by China, and the large-scale usage of big data analytics in various domains of science.

Keywords Big data · Bibliometric analysis · Citation analysis

1 Introduction

In science, it is imperative to identify the importance of research publications as well as the trends of research [3,17,55]. Bibliometrics is one such tool that performs

✉ Iftikhar Ahmad
ia@uetpeshawar.edu.pk

Gulzar Ahmed
gulzarahmed@aup.edu.pk

Syed Adeel Ali Shah
adeel@nwfpuet.edu.pk

Ejaz Ahmed
ejazahmed@ieee.org

¹ Department of Computer Science and Information Technology, University of Engineering and Technology, Peshawar, Pakistan

² Institute of Business and Management Sciences, The University of Agriculture, Peshawar, Pakistan

³ Centre for Mobile Cloud Computing Research, University of Malaya, Kuala Lumpur, Malaysia

quantitative analysis of research publications [19,33,34,55]. Bibliometrics identifies the most influential work in a specific field mostly by utilizing citation counts as a metric. Citations in turn help build new works on top of the existing knowledge, and forms a connection between the novel approaches to those of its predecessors. Other purposes of citations include crediting the peers' work, providing background information, and contextualizing one's own work [53]. Citation count is also a good indicator of the influence and visibility of a scientific publication.

Bibliometrics and citation analysis is used in various fields to identify the most influential work and researchers, and in the analysis and evolution of a specific research theme [17,55]. It is applied in various domains of science such as medicine [19], physics [27], social sciences [34], and computer sciences [24].

Big data research has seen a greater interest in the last decade, and attracted researchers from transdisciplinary areas such as physical sciences, natural sciences, social sciences, and biomedical sciences [2,4,13,28,38,42,49,50]. The concept of big data originated from the information explosion that occurred because of widespread adaptation of information and communication technologies. This resulted in massive amount of data generation. For instance, the Australian Square Kilometre Array Pathfinder (ASKAP) acquires 7.5 terabytes/second of image data [47]. Likewise in genomics, it is estimated that the data size is doubling every 7 months [52]. Stephens et al. [52] compared the big data phenomenon in three fields namely genomics, astronomy, and online social platforms (YouTube and Twitter). The authors discussed data acquisition, data storage, data distribution, and data analysis aspects of the aforementioned fields. Major findings of the study were that genomics is one of the most demanding big data domains and requires technological development in many fields to meet the computational needs [52]. In the time frame of 2008–2017, Scopus has recorded over 35,000 publications in connection with big data. This motivated us to analyze the literature published in the field of "big data" via citation analysis.

The objective of this work is to investigate the evolution of big data literature using bibliometric analysis. The aim is to identify the most influential work, top venues for publications, citation trends, geographical and institutional trends, as well as authorship trends in literature published in the domain of big data. The study is based on the articles published in the period of 10 years (2008–2017) and covers over 35,000 records. The rest of the article is organized as follows: Section 2 presents a brief review of studies in the field of bibliometrics. Section 3 presents research questions, and the methodology for data extraction. Section 4 presents an in-depth analysis of the data. Section 5 discusses the limitation of the study, and finally, Sect. 6 concludes the work.

2 Literature review

There is a plethora of work dedicated to citation analysis in various fields [17,19,24,27,34]. We briefly discuss works in bibliometrics in relation to "big data" and then present major scholarly publications in citation analysis.

Nobre and Tavares [48] analyzed the literature related to the application of big data/IoT in the context of circular economy indexed in Scopus for the time frame 2006–2015. The study found that China and USA are the most active countries. Surprisingly, among countries producing large greenhouse gases, Brazil and Russia were not contributing much in terms of number of publications in big data. Kalantari [40] performed bibliometric analysis of 6572 papers indexed in Web of Science from 1980 to March 19, 2015. Using MS Excel, general concentration, dispersion, and movement of the data from the selected pool were analyzed. Liao et al. [43] performed bibliometric analysis of big data literature published in the field of medical big data. The authors used Science Citation Index Expanded and the Social Science Citation Index databases as data sources to extract 988 references. There were no restrictions on the time span. The novelty of the work is the application of multi-regression analysis considering the number of authors, number of pages, and number of references. It was observed that the medical big data literature has seen a rise after 2010. By analysis of the keywords, it was identified that the medical care is shifting its focus toward patient-centered model than disease-centered approach.

One of the earliest works in the field of citation analysis in computer science is that of Culnan [18]. Culnan [18] analyzed and compared the citation patterns of academics and practitioners who published in the proceedings of a national computer science conference. The study identified that both the groups under consideration (academicians and practitioners) cited the same core journals, as well as documents belonging to the same age group. Goodrum et al. [29] analyzed computer science literature present on the web in the form of PDF and postscripts using autonomous citation indexing (ACI). For ACI citeseer (now called citeseerX) was used, and for a comparative analysis, Institute for Scientific Information (ISI) SCISEARCH was used. Using the data, profiles of source documents and citation profiles of two sources are discussed. Wohlin [54] used ISI Web of Science as data source to identify the most influential journal articles in software engineering for the year 1999.

Hoonlor et al. [35] performed an in-depth analysis of citation data in computer science. They inferred that most publications mention the keyword “algorithm,” and most abstracts are related to databases, neural networks, and Internet. The study also identified web as an attractive source of data and application test beds, which resulted in more research in the areas of data mining, cloud computing, and information retrieval. The study also concluded that funding is essentially required to keep research momentum and progress in a specific field.

Chadegani et al. [12] compared Scopus and ISI Web of Science based on a set of research questions. They concluded that ISI Web of Science has strong coverage dating old publications, whereas Scopus covers high-quality journals and more recent articles. Both databases provided customized search ability and are equally favored by scientists and researchers.

Ioannidis et al. [37] surveyed the highly cited scientists in biomedical field to score their publications to answer the question “*Is your most cited work your best?*”. The scientists were asked to score their publications on six dimensions including publication difficulty, surprise, disruptive innovativeness, greater synthesis, broader interest, and continuous progress. On average, a low average score was observed for publication difficulty, surprise, and disruptive innovativeness. The authors concluded

that beside citations-based metrics, other measures must also be used for evaluation of scientific works. Ding et al. [20] criticized the traditional way of citation analysis and proposed to use content-based citation analysis (CCA) to use the value of a citation at syntactic and semantic levels.

Garousi and Mäntylä [24] performed a comprehensive bibliometric assessment by considering over 70,000 articles published in the field of software engineering from Scopus. The authors observed a considerable growth in the number of publications per year. However, approximately 45% of the papers are not cited. Using text mining techniques, web services, mobile and cloud computing, industrial (case) studies, source code and test generation were identified as the most hot research topics.

Garousi [25] reported the findings of a bibliometric study of Turkish software engineering community based on the research publication in software engineering outlets and indexed in Scopus until 2014. Author identified the top-ranked university, and scholar. The study also identified the contributions made by the Turkish software engineering community to be very low in comparison with rest of the world. Likewise, the study also identified the lack of diversity in topics covered.

Effendy and Yap [22] performed the trend analysis of research areas in computer science using Microsoft Academic Graph dataset. The authors proposed a new metric called FoS score to measure the level of interest in a specific research topic. Using the measure, they discussed citation trends, trends in conferences, evolution of research areas, and the relation between research areas.

3 Research questions and data set

3.1 Research questions

Following the pattern of Garousi and Fernandes [26], we formulate a set of research questions and base our analysis of the data on these key questions. The main objective of choosing the research questions is to identify the top cited papers, the contributions of various countries in advancing big data research, and identification of key research areas. The set of research questions are as following

- RQ 1. What are the top cited publications in the selected time frame?
- RQ 2. What are the top cited publications for each year in the selected time frame?
- RQ 3. What are the key topics/areas that are addressed in publications?
- RQ 4. What are the top venues for the most cited publications?
- RQ 5. What is the citation landscape of the publications? What is the average number of citations/publications?
- RQ 6. Which countries and institutions have contributed most in terms of publications count?
- RQ 7. What is the authorship trend? What is the average number of authors per publications?

Table 1 Query for Data Extraction

Query

```
( TITLE ( "Big Data" ) OR KEY ( "Big Data" ) ) AND ( PUBYEAR > 2007 AND
PUBYEAR < 2018 ) AND ( LIMIT-TO ( LANGUAGE,"English " ) )
```

3.2 Data set extraction

A key consideration in any citations-based study is the selection of data source. A number of online databases are available that provide access to the citation data. These sources include ISI Web of Knowledge, Scopus, Google Scholar and dblp. Out of these, ISI Web of Knowledge and Scopus are the two main sources used by the majority of researchers for citation data analysis [17,24,25,33,34]. Other citation databases are also used such as dblp by Hoonlor et al. [35]. Our choice for data collection is Scopus because of service availability, ease of use and authenticity of the data [26].

After selection of the source, the next step is the extraction of the required data from the data source. Scopus provides a flexible and customized way for data extraction from its database using various criteria such as search by author name, source name, affiliation, and keywords search. We used `title` and `keywords` as our main search criterion, i.e., we queried the database to return all documents where the word “Big Data” is located either in the title of the document or in the associated keywords. We also restricted the results to English language articles and articles published between 2008 and 2017 (both inclusive). The query is given in Table 1.

Scopus has indexed 13 documents prior to 2008 satisfying our search criterion. Out of these, only a single document¹ has received 19 citations. Likewise, no more than 3 papers/year are indexed by Scopus prior to 2008. Therefore, we chose 2008 as starting year for our data extraction. Thus, the time frame spans a decade of research in the field. Our search criteria have resulted in total of 34,655 documents. Using the graphical user interface provided by Scopus, we exclude some document types namely *conference review*, *editorial notes*, and *letters*. The main motivation behind the omission of these document type resides in the fact that they mostly provide information about the outlet such as scope of the conference/special issue, the number of submissions, and the acceptance rate, and normally have very low citation counts. The omission of these documents reduced the documents number to 33,623. As stated earlier that big data is a transdisciplinary field, we did not limit to literature from computer science only and allowed results from diverse fields such as business, medical science, and social sciences. Note that the data are downloaded on January 27, 2018, and the citation count might slightly differ at later dates. The dataset is stored in comma-separated values (CSV) format, and analysis is performed in the R statistical tool.

¹ “Big data” dynamic factor models for macroeconomic measurement and forecasting: A discussion of the papers by Lucrezia Reichlin and by Mark W. Watson (Book Chapter) Diebold, F.X. 2003 *Advances in Economics and Econometrics: Theory and Applications*, Eighth World Congress, Volume III pp. 115–122 19.

Table 2 Top 10 publications by citation count

Rank	Title	Year	Citation count
1	Business intelligence and analytics: From big data to big impact	2012	1098
2	Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon	2012	999
3	Data mining with big data	2014	681
4	Internet of Things: A survey on enabling technologies, protocols, and applications	2015	532
5	Experimental evidence of massive-scale emotional contagion through social networks	2014	526
6	The parable of google flu: Traps in big data analysis	2014	526
7	Big data: A survey	2014	525
8	Internet of things in industries: A survey	2014	520
9	Data-intensive applications, challenges, techniques and technologies: A survey on Big Data	2014	499
10	Private traits and attributes are predictable from digital records of human behavior	2013	457

4 Results

In this section, we present answers to the research questions designed in Sect. 3.1.

4.1 Identification of top 10 research publications

Table 2 presents the top 10 publications based on the absolute number of citations received. Rather unsurprisingly, most of the top 10 research publications include works that are survey in nature. One of the interesting exceptions is the work on Rank # 5 which reports findings of an experimental study of Facebook [1]. The authors investigated if emotional contagion occurs on Facebook by analyzing contents in the newsfeed of users. Reducing positive contents lead to the reduction of positive posts and an increase in negative posts. Same trends were observed when negative expressions were reduced in the news feed. The authors concluded that emotional states are transferable to other via emotional contagion.

It can be seen from Table 2 that all the publications (with the exception of one) are at least 4 years old, i.e., they are published in 2014 or earlier. These publications have accumulated more citations than others. Therefore, relying on absolute numbers of citations is age-dependent, and puts the recent publications at a disadvantage in comparison with older publications. To overcome this issue, we calculated the normalized score by dividing the absolute number of citations by the number of years since publications. We believe that the normalized score is a better indicator to gauge the quality of recent publications.

We observe that the top publication both in terms of the absolute number of citations and the normalized score is that of Chen et al. [13], which highlights the importance

Table 3 Top 10 publications by normalized score

Rank	Change to Table 2	Title	Year	Normalized score
1	–	Business intelligence and analytics: From big data to big impact	2012	183
2	+2	Internet of Things: A survey on enabling technologies, protocols, and applications	2015	177.34
3	–	Data mining with big data	2014	170.25
4	–2	Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon	2012	166.5
5	–	Experimental evidence of massive-scale emotional contagion through social networks	2014	131.5
6	–	The parable of google flu: Traps in big data analysis	2014	131.5
7	–	Big data: A survey	2014	131.25
8	–	Internet of things in industries: A survey	2014	130
9	–	Data-intensive applications, challenges, techniques and technologies: A survey on Big Data	2014	124.75
10	+3	The rise of “big data” on cloud computing: Review and open research issues	2015	121.34

of the work over the years. The work of Al-Fuqaha et al. [23] jumps from rank 4 to 2 in normalized score, highlighting the coverage of *IoT* in the recent literature. It can also be seen in Table 3 that there is no significant changes in ranks of publication in comparison with Table 2. The top 9 remains the same with only 2 and 4 swapping places. The only new entry in Table 3 is the work of Hashem et al. [31] which gained 3 ranks.

4.2 Most cited articles for each year

Table 4 identifies top cited papers for each year based on the absolute citation count. The normalized score is also provided. The table portrays the evolution of big data research. In the early years, research mostly involved the potential usage, structure, and applications of big data [6,36,39]. In the later years, the top cited papers covered different areas such as development of big data analytics tool [32], predicting personal attributes (such as ethnicity, political views, personalities traits etc) from Facebook likes [41], and the use of deep neural networks in big data [44]. A key anomaly in Table 4 is the absolute and normalized citation score of [6]. The publication has a normalized score of 9.625 which is nearly 4 times less than the next entry.

4.3 Identifying research topics/trends

Unlike other scientific fields that have clear taxonomy and classification of the subject area [8], there is no classification scheme for big data. We use the frequency

Table 4 Top cited papers for each year

Year	Title	Citation count	Normalized score
2008	Big data: The future of biocuration	431	43.1
2009	The pathologies of big data	250	27.77
2010	Big science and big data in biology: From the international geophysical year through the international biological program to the Long Term Ecological Research (LTER) network, 1957-present	77	9.625
2011	Starfish: A self-tuning system for big data analytics	273	39
2012	Business intelligence and analytics: From big data to big impact	1098	183
2013	Private traits and attributes are predictable from digital records of human behavior	457	91.4
2014	Data mining with big data	681	170.25
2015	Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications	532	177.34
2016	Social big data: Recent achievements and new challenges	111	55.5
2017	A survey of deep neural network architectures and their applications	54	54

of keywords as metric to identify sub-areas and topics that gained the attention of researchers over the course of time. Our analysis found that the phrase “Big Data” is the most widely used keyword in our collection of records, which is unsurprising. The other keywords in order of occurrences are Data Mining (4995), Data Handling (4242) Digital Storage (3185), Cloud Computing (2921), Information Management (2795), Artificial Intelligence (2769), Distributed Computer Systems (2515), Learning Systems (2261), and Algorithms (2027). In order to make the comparisons more realistic, we removed the keyword “Big Data” and designed a word cloud using the statistical software *R* (see Fig. 1).

In the early phase of big data research, the researcher focused mainly on uses and applications of big data research. For example, Jacob [39] discussed the challenges posed by the big data and highlighted possible solutions to overcome the challenges. Cohen et al. [16] discussed that the cost of data acquisition and storage has reduced considerably, and sophisticated data analysis has become a norm. They introduced Magnetic, Agile, Deep (MAD) data analysis practice. The proposed approach, design philosophy, and techniques were used to provide MAD analytics for Fox Audience Network. In addition, some works such as that of Brinkmann et al. [10] designed systems for large-scale data acquisition, processing, and storage. The authors claimed to collect 3 terabytes of data per day by performing continuous electrophysiological recordings of patients undergoing evaluation for epilepsy surgery. The huge amount of data generated posed storage, and processing challenges. Authors designed a platform that facilitated the acquisition, compression, and storage of large amount of data.

Table 5 Top venues by most number of citations

Venue	Citation count	Publication count	Citations/publication
Lecture Notes in Computer Science	3064	2538	1.21
Proceedings of the VLDB Endowment	2369	99	23.93
Nature	1472	34	43.29
Proceedings of the National Academy of Sciences of the United States of America	1271	19	66.89
Future Generation Computer Systems	1195	129	9.26
Proceedings 2014 IEEE International Conference on Big Data, IEEE Big Data 2014	1172	295	3.97
MIS Quarterly: Management Information Systems	1111	5	222.2
IEEE Access	1108	131	8.46
Information Communication and Society	1085	14	77.5
Procedia Computer Science	1069	437	2.45
Proceedings of the ACM SIGMOD International Conference on Management of Data	1020	87	11.72
Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	1012	68	14.88
IEEE Transactions on Knowledge and Data Engineering	985	31	31.77
Information Sciences	908	45	20.17
IEEE Network	902	39	23.13

of the data reveals that the *MIS Quarterly: Management Information Systems* has received 1111 citations for 5 publications. Out of 1111 citations, 1098 citations are for [13], which means that the rest of 4 publications received only 13 citations. *Information Communication and Society* ranks 2 based on citations/publication; however, one publication [9] accrued 999 citations, whereas the remaining 13 publications acquired 86 citations. Other top venues based on the citation/publication include *Proceedings of the National Academy of Sciences of the United States of America* (66.9), *Nature* (43.3), and *IEEE Transactions on Knowledge and Data Engineering* (31.78).

4.5 Analyzing citation trends

The complete set of publications received 1,06,598 citations, which equates to 3.17 citations per publication. A vast majority (55%) has not received any citation. The number is significantly higher than reported in other studies [24,26]. Garousi and Fernandes [24] reported that 43% of research publications have zero citation count when analyzing the literature of software engineering. 15% of publications have received only a single citation which is in the same range as reported in [26]. Further analysis

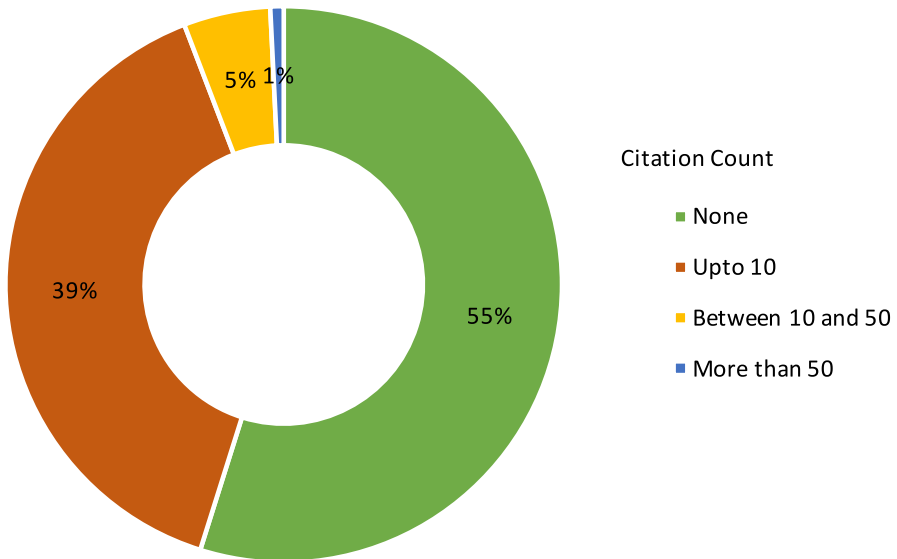


Fig. 2 Distribution of citation

of the data shows that 94% of publications have received no more than 10 citations, and 0.8% have received over 50 citations. Figure 2 presents an overview of the citation distribution.

An important question regarding the citation landscape of a set of publication is related to the applicability of power law [11]. It is often stated that citations of scientific publications follow heavy tail distribution (see [11] and references therein). In citation networks, it means that 80% of the citations are received by 20% of publications. Although the full investigation regarding the fitness of our data to power law is beyond the scope of this work, we will like to highlight that in our collected data, 80% citations are received by the top 12.74% publications (see Fig. 3), i.e., the vast majority of citations are received by relatively fewer number of publications.

4.6 Geographical and institutional contributions

Table 6 presents the yearly contribution by the top 10 countries. Geographically, China has contributed the most number of publications (8901), followed by the USA (8568), and India (2342). It is interesting to note that during the years 2008–2011 China has produced only 4 papers (zero publication prior to 2011), whereas USA has produced 36 publications. During the same period, the other top-10 countries (excluding China) contributed 23 publications, i.e., the cumulative sum of 9 countries (27) is less than that of USA (36). These numbers establish the fact that the USA played a pioneering role in big data research and is later joined by other countries. It is also observed that during the period 2008–2011 Japan (1 paper in each year) and South Korea (1 paper in 2010) are the only Asian countries to publish in the domain. It can also be seen that China leapfrogged USA in the last two years only.

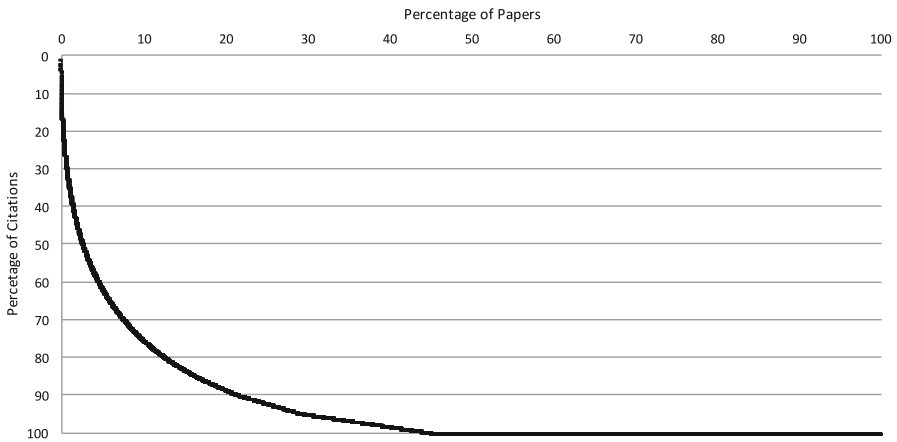


Fig. 3 Power law in big data citations

Table 6 Yearly contributions from top 10 countries

Country	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
China	0	0	0	4	51	310	741	1716	2946	3133
United States	3	5	5	23	209	575	1223	2213	2516	1796
India	0	0	0	0	15	45	173	507	887	715
United Kingdom	1	0	1	1	24	113	239	436	576	522
Germany	1	0	0	3	24	91	305	375	427	366
South Korea	0	0	1	0	19	55	136	350	406	372
Australia	0	0	0	2	10	48	141	301	340	367
Italy	0	0	0	2	12	53	114	262	320	328
Japan	1	1	1	2	34	71	145	243	309	275
Canada	0	1	1	4	6	63	115	238	321	250

In terms of author's affiliation *Chinese Academy of Sciences, China* has produced the maximum number of papers (698), followed by *Tsinghua University, China* (422), and *Ministry of Education China* (279). The top 10 institutes belong to China except *CNRS Centre National de la Recherche Scientifique, France* which occupies 9th position. The first entry from the USA is *Carnegie Mellon University* at position 15, followed by *Massachusetts Institute of Technology* at 16. Table 7 presents a summarized view of top-30 institutions in terms of the number of publications. It is important to note that a single publication may include authors from more than one country, and more than one affiliations (for a publication with all authors from the same country).

4.7 Authorship trends

For the complete time frame (2008–2017), the average number of authors per publication is 3.45 with a standard deviation of 2.34. In total, 13.67% of articles are published

Table 7 Top 30 Institutions Statistics

Country	# of Publications	# of Institutions	Paper/institution
China	3928	17	231.06
USA	705	5	141
Australia	415	3	138.34
India	269	2	134.5
France	193	1	193
Singapore	139	1	139
UK	134	1	134

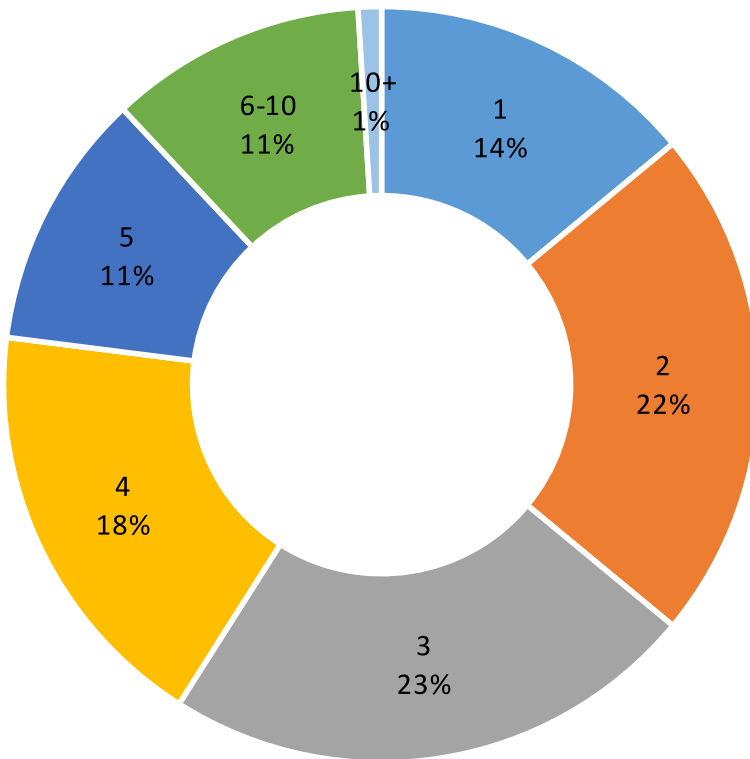


Fig. 4 Distribution of number of authors per publication

with a single authorship, and 41% of articles have more than the average number of authors, i.e., 41% of publications have at least 4 authors. The surprising aspect of the findings is the share of publications with more than 10 authors. We found that 1% of the papers have more than 10 authors. Figure 4 summarizes the number of authors per publication distribution.

We also analyzed the evolution of number of authors per research publications over the selected time frame. We observe that during the initial years (2008–2012) 27%

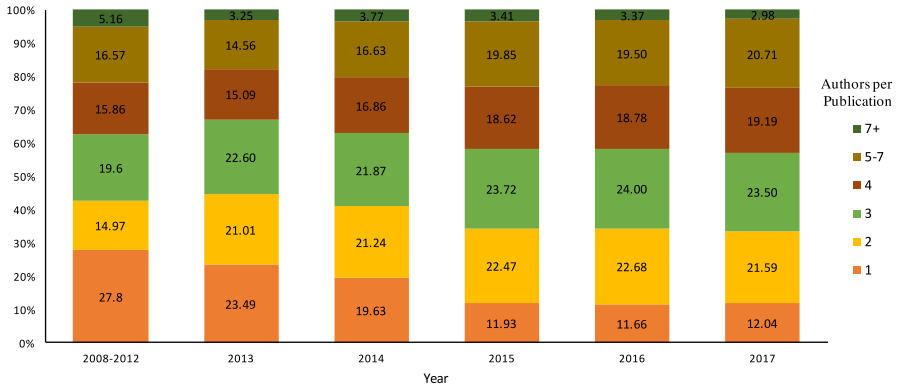


Fig. 5 Trend of authors per publication

of publications have single authorship. During the later years, the single authorship trend declines. In 2016, only 11.6% of papers have single authorship, the minimum for the selected time frame. Figure 5 depicts the authorship trends for the time frame. It should be noted that as the number of publications are significantly lower for initial years, the data for years 2008–2012 are combined for reporting purposes.

5 Limitations of the study

It is important to present the limitation of the study, as it might be possible that the results are not obtained if the same set of experiments is repeated again. We downloaded the data from Scopus on January 27, 2018. As Scopus has data download limits, the data were downloaded in an incremental manner and later combined. It is possible that the reader may find the number of citations for various articles different than the ones reported here. There can be several reasons for this. For example, the article might have accumulated more citations over time. It is also possible that various sources might report different statistics for the same article. A noticeable example is the case of publication “*Data mining with big data.*” The paper according to Scopus has received 681 citations, whereas Google Scholar has recorded 1320 citations for the publication till 2017. The publisher *IEEE Transactions on Knowledge and Data Engineering* recorded 524 citations for the same publication. The difference between the citation count of Scopus and other sources can be attributed to the fact that Scopus citations are based on Scopus-indexed publications only. It is also important to mention that Google Scholar citations also include non-academic citations. As stated earlier that a variety of past studies has validated the authenticity of Scopus, therefore, our results are also based on the statistics of Scopus only.

6 Conclusion

The study systematically analyzed the citations of publications in Big Data to answer a variety of research questions. Although our data start from 2008, we observe, rather surprisingly, that none of the top-10 publications are published between 2008 and 2011.

As expected, the highest cited publications include survey papers in the field. We also found that over 50% of publications are not cited, and 80% of citations are received by 12.74% of publications. Geographically, China has the most number of publications, followed by USA. However, in the initial years, China has no significant contribution. This confirms the leadership role played by USA in the field. We also observed that average number of citation per publication to be 3.17, whereas the average number of authors per publication is 3.45.

The current study is based on citation count as metric and treats every citation equally. In other words, it does not differentiate between a citation included for the sake of completeness of work, and the one which forms the foundation of a research work. It will be interesting to perform content-based citation analysis of the field and identify important works based on the context as well. Another research direction will be to analyze the citation counts differences between various sources, and compare the ranking.

References

1. Adam D, Kramer I, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proc Natl Acad Sci* 111(24):8788–8790
2. Ahmed E, Yaqoob I, Hashem IAT, Khan I, Ahmed AIA, Imran M, Vasilakos AV (2017) The role of big data analytics in Internet of Things. *Comput Netw* 129:459–471
3. Aksnes DW (2003) Characteristics of highly cited papers. *Res Eval* 12(3):159–170
4. AlZubi AA (2018) Big data analytic diabetics using map reduce and classification techniques. *J Supercomput*. <https://doi.org/10.1007/s11227-018-2362-1>
5. Anagnostopoulos I, Zeadally S, Exposito E (2016) Handling big data: research challenges and future directions. *J Supercomput* 72(4):1494–1516. <https://doi.org/10.1007/s11227-016-1677-z>
6. Aronova E, Baker KS, Oreskes N (2010) Big science and big data in biology: from the international geophysical year through the international biological program to the long term ecological research (LTER) network, 1957 present. *Hist Stud Nat Sci* 40(2):183–224
7. Bello-Orgaz G, Jung JJ, Camacho D (2016) Social big data: recent achievements and new challenges. *Inf Fusion* 28:45–59
8. Bourque P, Abran A, Garbajosa J, Keeni G, Shen B (2014) Guide to the software engineering body of knowledge (SWEBOK) version3.0. IEEE Press
9. Boyd D, Crawford K (2012) Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inf Commun Soc* 15(5):662–679
10. Brinkmann BH, Bower MR, Stengel KA, Worrell GA, Stead M (2009) Large-scale electrophysiology: acquisition, compression, encryption, and storage of big data. *J Neurosci Methods* 180(1):185–192
11. Brzezinski M (2015) Power laws in citation distributions: evidence from scopus. *Scientometrics* 103(1):213228
12. Chadegani A, Arezoo, Salehi H, Yunus M, Farhadi H, Fooladi M, Farhadi M, Ebrahim NA (2013) A comparison between two main academic literature collections: Web of Science and Scopus databases. *Asian Soc Sci* 9(5):18–26
13. Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: from big data to big impact. *MIS Q* 36(4):1165–1188
14. Chen Y, Alspaugh S, Katz R (2012) Interactive analytical processing in big data systems: a cross-industry study of mapreduce workloads. *Proc VLDB Endow* 5(12):1802–1813
15. Chianese A, Marulli F, Piccialli F, Benedusi P, Jung JE (2017) An associative engines based approach supporting collaborative analytics in the internet of cultural things. *Future Gener Comput Syst* 66:187–198
16. Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C (2009) MAD skills: new analysis practices for big data. *Proc VLDB Endow* 2(2):1481–1492

17. Crespo JA, Herranz N, Li Y, RuizCastillo J (2014) The effect on citation inequality of differences in citation practices at the web of science subject category level. *J Assoc Inf Sci Technol* 65(6):1244–1256
18. Culnan MJ (1978) An analysis of the information usage patterns of academics and practitioners in the computer field: a citation analysis of a national conference proceedings. *Inf Process Manag* 14(6):395–404
19. Davis PM (2009) Authorchoice openaccess publishing in the biological and medical literature: a citation analysis. *J Assoc Inf Sci Technol* 60(1):3–8
20. Ding Y, Zhang G, Chambers T, Song M, Wang X, Zhai C (2014) Contentbased citation analysis: the next generation of citation analysis. *J Assoc Inf Sci Technol* 65(9):1820–1833
21. Dou C, Cui Y, Wong R, Atif M, Li G, Ranjan R (2017) Unsupervised blocking and probabilistic parallelisation for record matching of distributed big data. *J Supercomput*. <https://doi.org/10.1007/s11227-017-2008-8>
22. Effendy S, Yap RHC (2017) Analysing trends in computer science research: a preliminary study using the Microsoft Academic Graph. In: Proceedings of the 26th International Conference on World Wide Web companion. International World Wide Web, Conferences Steering Committee, pp 1245–1250
23. Al-Fuqaha A, Guizani M, Mohammadi M, Aledhari M, Ayyash M (2015) Internet of things: a survey on enabling technologies, protocols, and applications. *IEEE Commun Surv Tutor* 17(4):2347–2376
24. Garousi V, Mäntylä MV (2016) Citations, research topics and active countries in software engineering: a bibliometrics study. *Comput Sci Rev* 19:56–77
25. Garousi V (2015) A bibliometric analysis of the Turkish software engineering research community. *Scientometrics* 105(1):23–49
26. Garousi V, Fernandes JM (2016) Highly-cited papers in software engineering: the top-100. *Inf Softw Technol* 71:108–128
27. Gingras Y, Wallace ML (2010) Why it has become more difficult to predict Nobel Prize winners: a bibliometric analysis of nominees and winners of the chemistry and physics prizes (19012007). *Scientometrics* 82(2):401–412
28. Gohar M, Ahmed SH, Khan M, Guizani N, Ahmed A, Rahman AU (2018) A big data analytics architecture for the internet of small things. *IEEE Commun Mag* 56(2):128–133
29. Goodrum, Abby A, McCain KW, Lawrence S, Giles CL (2001) Scholarly publishing in the Internet age: a citation analysis of computer science literature. *Inf Process Manag* 37(5):661–675
30. Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden A, Batcheller et al (2013) Big data and the future of ecology. *Front Ecol Environ* 11(3):156–162
31. Hashem, Targio IA, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of big data on cloud computing: review and open research issues. *Inf Syst* 47:98–115
32. Herodotou, Herodotos, Lim H, Luo G, Borisov N, Dong L, Cetin FB, Babu S (2011) Starfish: a self-tuning system for big data analytics. In *Cidr* 11(2011):261–272
33. Ho Y-S (2012) Top-cited articles in chemical engineering in Science Citation Index Expanded: a bibliometric analysis. *Chin J Chem Eng* 20(3):478–488
34. Ho Y-S (2014) Classic articles on social work field in Social Science Citation Index: a bibliometric analysis. *Scientometrics* 98(1):137–155
35. Hoonlor A, Szymanski BK, Zaki MJ (2013) Trends in computer science research. *Commun ACM* 56(10):74–83
36. Howe, Doug, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP et al (2008) Big data: the future of biocuration. *Nature* 455(7209):47
37. Ioannidis J, Boyack KW, Small H, Sorensen AA, Klavans R (2014) Bibliometrics: is your most cited work your best? *Nat News* 514(7524):561–562
38. Jabbar S, Malik KR, Ahmad M, Aldabbas O, Asif M, Khalid S, Han K, Ahmed SH (2018) A methodology of real-time data fusion for localized big data analytics. *IEEE Access* 6:24510–24520
39. Jacobs A (2009) The pathologies of big data. *Commun ACM* 52(8):36–44
40. Kalantari A, Kamsin A, Kamaruddin HS, Ebrahim NA, Gani A, Ebrahimi A, Shamshirband S (2017) A bibliometric approach to tracking big data research trends. *J Big Data* 4(1)
41. Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. *Proc Natl Acad Sci* 110(15):5802–5805
42. Leonelli S (2014) What difference does quantity make? On the epistemology of Big Data in biology. *Big Data Soc* 1(1)
43. Liao H, Tang M, Luo L, Li C, Chiclana F, Zeng X-J (2018) A bibliometric analysis and visualization of medical big data research. *Sustainability* 10(1)

44. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017) A survey of deep neural network architectures and their applications. *Neurocomputing* 234:11–26
45. Lladós J, Cores F, Guirado F (2018) Optimization of consistency-based multiple sequence alignment using Big Data technologies. *J Supercomput*. <https://doi.org/10.1007/s11227-018-2424-4>
46. Murdoch TB, Detsky AS (2013) The inevitable application of big data to health care. *JAMA* 309(13):1351–1352
47. Newman R, Tseng J (2018) Cloud computing and the square kilometre array. www.skatelescope.org/uploaded/8762_134_Memo_Newman.pdf. Last Accessed 8th May
48. Nobre GC, Tavares E (2017) Scientific literature analysis on big data and internet of things applications on circular economy: a bibliometric study. *Scientometrics* 111(1):463–492
49. Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2(1)
50. Rani S, Ahmed SH, Talwar R, Malhotra J (2017) Can sensors collect big data? An energy-efficient big data gathering algorithm for a WSN. *IEEE Trans Ind Inform* 13(4):1961–1968
51. Rodríguez-Mazahua L, Rodríguez-Enríquez C-A, Sánchez-Cervantes JL, Cervantes J, García-Alcaraz JL, Alor-Hernández G (2016) A general perspective of big data: applications, tools, challenges and trends. *J Supercomput* 72(8):3073–3113. <https://doi.org/10.1007/s11227-015-1501-1>
52. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al (2015) Big data: astronomical or genomical? *PLoS Biol* 13(7)
53. Thomson Reuters Using Bibliometrics: A guide to evaluating research performance with citation data www.ips.clarivate.com/m/pdfs/325133_thomson.pdf. Last Accessed 8 Feb 2018
54. Wohlin C (2005) An analysis of the most cited articles in software engineering journals—1999. *Inf Softw Technol* 47(15):957–964
55. Wohlin C (2007) An analysis of the most cited articles in software engineering journals—2000. *Inf Softw Technol* 49(1):2–11