Ann. Data. Sci. (2017) 4(1):1–17
DOI 10.1007/s40745-016-0096-6

Big Data Paradigm: What is the Status of Privacy and Security?

Kenneth David Strang[1] · **Zhaohao Sun**[2]

Received: 16 August 2016 / Revised: 15 October 2016 / Accepted: 26 December 2016 /
Published online: 21 January 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract We extended the big data body of knowledge by analyzing the longitudinal literature to highlight important research topics and identify critical gaps. We initially collected 79,012 articles from 1900 to 2016 related to big data. We refined our sample to 13,029 articles allowing us to determine that the big data paradigm commenced in late 2011 and the research production exponentially rose starting in 2012, which approximated a Weibull distribution that captured 82% of the variance ($p < .01$). We developed a dominant topic list for the big data body of knowledge that contained 49 keywords resulting in an inter-rater reliability of 93% ($r^2 = 0.89$). We found there were 13 dominant topics that captured 49% of the big data production in journals during 2011–2016 but privacy and security related topics accounted for only 2% of those outcomes. We analyzed the content of 970 journal manuscripts produced during the first of 2016 to determine the current status of big data research. The results revealed a vastly different current trend with too many literature reviews and conceptual papers that accounted for 41% of the current big data knowledge production. Interestingly, we observed new big data topics emerging from the healthcare and physical sciences disciplines.

✉ Kenneth David Strang
Kenneth.Strang@plattsburgh.edu

Zhaohao Sun
zsun@dbs.unitech.ac.pg; zhaohao.sun@gmail.com

[1] Regional Higher Education Center, School of Business and Economics, State University of New York, Plattsburgh, 640 Bay Road, Queensbury, NY 12804, USA

[2] Department of Business Studies, PNG University of Technology, Lae 411, Papua New Guinea

Keywords Big data body of knowledge · Literature review · Big data paradigm · Big data security · Big data privacy · Exponential weibull trend · Kappa interrater agreement

1 Introduction

Big data is a relatively new paradigm with almost all of the scholarly literature having emerged since 2012. Therefore it is a field that would benefit from literature reviews and additional studies. Chen et al. [1] conducted a review on big data state-of-the-art and concluded that there were several gaps in the literature. Their claims are worth quoting to emphasize the problems.

> There is compelling need for a rigorous definition of big data, a structural model of big data, formal description of big data, and a theoretical system of data science, etc. At present, many discussions of big data look more like commercial speculation than scientific research. This is because big data is not formally and structurally defined and not strictly verified. [1, p. 81].

They asserted that privacy, security and safety are key concerns which were not adequately investigated within the big data body of knowledge [1]. Their work was completed a couple of years ago so the research question here is see if their recommendations for more research have been pursued by scholars and practitioners. On that premise a research goal for this study was to determine what scholarly research has been published that extends our understanding of big data privacy, security and safety. The aim was also to examine what scholarly big data research topics were recently published.

Big data privacy and security risks are hot topics in the media as well as in academic research as proven by Goldfield [2], Kambatla et al. [3], Kim et al. [4], and Pence [5]. From an external risk perspective, there were multiple occurrences of cyber security hackers breaking into multiple private electronic databases, linking fields together, and subsequently leveraging that data to obtain confidential information [6–11]. For example, in the USA a "hacker believed to be tied to the Russian intelligence services made public another set of internal Democratic Party documents on Friday, including the personal cellphone numbers and email addresses of nearly 200 lawmakers" [12]. Some big data privacy dilemmas originated internally as revealed by several well-known American company misadventures, namely Orbitz, Netflix and Target [13,14].

Healthcare practitioner experience suggests there is a need for more research into big data privacy and security. When I worked for Blue Cross earlier in my career we considered our millions of health claim records (in the gigabyte range) to be big data because they were complex and distributed across several mainframe, mini and desktop network systems. It was difficult to understand exactly what was actually in the data but we developed methods to calculate important benchmarks to inform our actuarial estimates (set asides to pay future claims), to create group or individual premium quotes, and for other strategic decision making such as new product/service development or harvesting/retiring. The datum relationships were complex mainly due to the evolving regulations, products/services and our internal applications that necessitated

field additions, modifications, deletions or entity-relationship changes. Additionally, accuracy was a problem because the online claim systems sometimes allowed errors to be captured despite our rigorous overnight mainframe batch adjudication routines. Policy holder privacy and data base security concerns grew in importance because of regulations and due to the ability of hackers to somehow get access to our information (often through an employee accessing an Internet-based application or browser which inadvertently spawned off viruses to steal or in one case delete data). Several researchers concur that healthcare big data privacy is an important yet unpopulated body of knowledge [10,15–18].

There has been some contemporary research published about big data privacy and security. However, the majority of recent literature on big data privacy states that we need more research (e.g. [8–10,15,16,19–25]). This was compelling justification to examine the status of big data privacy and security research.

In this paper we examine the contemporary literature to determine which big data topics scholars have recently examined as well as the relative contribution towards big data privacy and security issues. This is important to know for two reasons. Firstly, researchers and organizations need to know what current literature is available across the disciplines about big data privacy and security. Secondly, since we know there are gaps in the literature about big data privacy and security, research fund granting institutions and universities ought to ensure the current production proportion of this critical topic is sufficient and not being displaced by a relative overproduction of less important or saturated big data topics.

It is difficult to validate how much research is sufficient in any topic but it makes sense to know the status of the big data body of knowledge. Consumers, scholars and decision makers of research funding need to know how much relative focus is being put on big data privacy and security studies as compared to other facets in the big data paradigm. The results from this study will inform grant funding and research decisions for big data privacy and security topics. There may be other generalizations from the results, such as what are the gaps or emerging topics in the scholarly big data body of knowledge literature.

2 Methodology

We took a positivistic ideology with this study so we collected empirical qualitative data. We applied nonparametric statistics such as rank and distribution tests to answer the research questions. We applied both linear and nonlinear techniques to evaluate the data.

We collected data from publishers through their searchable literature indexes (N = 34), as summarized in Table 1. The table in descending order by number of articles found. Some indexes contained other indexes such as ABI/Inform, ACM Digital Library, Bacon's Media, Cabell's, DBLPDBLP, Index Copernicus, INSPEC and others. Although additional indexes existed, when searched them we found severe duplication so based on the principle of diminishing returns we concentrated on exploring those listed in Table 1. We did not use Google or other public domain search engines

Table 1 Scholarly indexes for data collection in descending order (N = 34)

Data source	
Business Insights: Essentials	32.2%
Business Source Complete	12.4%
Points of View Reference Center	9.1%
Applied Science & Technology Source	8.9%
ScienceDirect (with proquest)	6.4%
MasterFILE Premier	4.4%
OmniFile Full Text Select (H.W. Wilson)	4.3%
MEDLINE with Full Text	4.1%
Education Source	2.2%
InfoTrac Newsstand	2.0%
Vocational and Career Collection	1.7%
Canadian Reference Centre	1.4%
Environment Complete	1.2%
CINAHL Plus with Full Text	1.2%
PsycINFO	1.1%
Library, Information Science & Technology	1.1%
Scopus®	1.1%
Professional Development Collection	0.8%
Military & Government Collection	0.8%
Opposing Viewpoints in Context	0.6%
Social Sciences Full Text (H.W. Wilson)	0.5%
Humanities Source	0.5%
Entrepreneurial Studies Source	0.4%
SocINDEX with Full Text	0.3%
Literature Resource Center	0.3%
Energy & Power Source	0.3%
GreenFILE	0.2%
Teacher Reference Center	0.2%
Criminal Justice Abstracts	0.2%
America: History & Life	0.1%
Historical Abstracts	0.1%
Art Full Text (H.W. Wilson)	0.1%
LexisNexis Academic: Law Reviews	0.1%
Government Printing Office Catalog	0.0%

since our goal was to focus on scholarly literature and we desired an accurate result with as few duplicates as possible to filter out.

The data we collected was big data from a pragmatic view point. From a theoretical perspective it was not big data although the sample had a high variety/complexity with presumed veracity (accuracy) and value. However, it did not meet the other two V's of big data, namely high volume and high velocity [25]. Our full text data

reached an estimated size of 100,000 MB when considering an article averages 1 MB including images and tables. Each article contained complex data, with some related to one another and many others being distinctly unique. Our initial sample of big-data-related full text articles was N = 79,012 using the timeframe of 1916–2016, which is one hundred years.

We analyzed the abstract, title and keywords for all publications available matching the keywords "big data". Due to the size of the data, we analyzed only the linear and nonlinear trends for all publications. We then determined the cutting point using nonparametric statistics to identify when the big data paradigm of publications significantly emerged in the literature, which was 2012.

We then analyzed the full text articles from scholarly academic journals from the cutting point onwards (2011–2016) to answer the research questions. We used 2011 as the starting point for this second stage because at the time of writing we were in the middle of 2016 so we wanted to include full year periods—thus we started in the middle of 2011 and ended in the middle of 2016 (therefore our annual intervals were not identical with the calendar year but we used the latter to simplify reporting and consumer interpretation). We use the term sample in this study although we assert our data represents the population of full text scholarly literature in the respective periods. Nonetheless we recognize there more big data articles beyond those which are full text and indexed so we will continue to apply the word sample here.

We used text analytics on resulting full text articles to locate the highest frequencies of words used, and merged those with the keywords given for the manuscript. We then designated one best-fitting keyword to represent the dominant topic of every full text article. From that we performed various analytics to highlight interesting patterns. Finally we interpreted the results to answer the research questions and then draw implications. We closed with recommendations for future study.

3 Longitudinal Big Data Literature Analysis

The most influential source of the big data literature from the EBSCO Business Insights: Essentials index at 32%, which was followed by Business Source Complete at 12.4%, Points of View Reference Center at 9.1%, Applied Science & Technology Source with 8.9% and ScienceDirect at 6.4%. These five sources amounted to almost 70% of the data. Figure 1 is a columnar chart summarizing the top 10 sources for our data sample, starting from 1916 and ending in 2016. We actually included 11 indexes plus the other category. The other column represents sources that individually held less than 1% of the sample.

Next we examined the production of big data articles by year. Of the 79,012 articles that we collected since 1916, it seemed clear that none existed prior to 2000 and most were dated after 2010. We analyzed the total production of big data articles by frequency probability, which is a technique that may be used to predict future trends (once the distribution shape is known). We compared the big data article distribution to several distribution types without censoring—Fig. 2 illustrates a panel of the four short-listed probability plots whereby lognormal and Weibull appeared to best-describe the sample distribution.

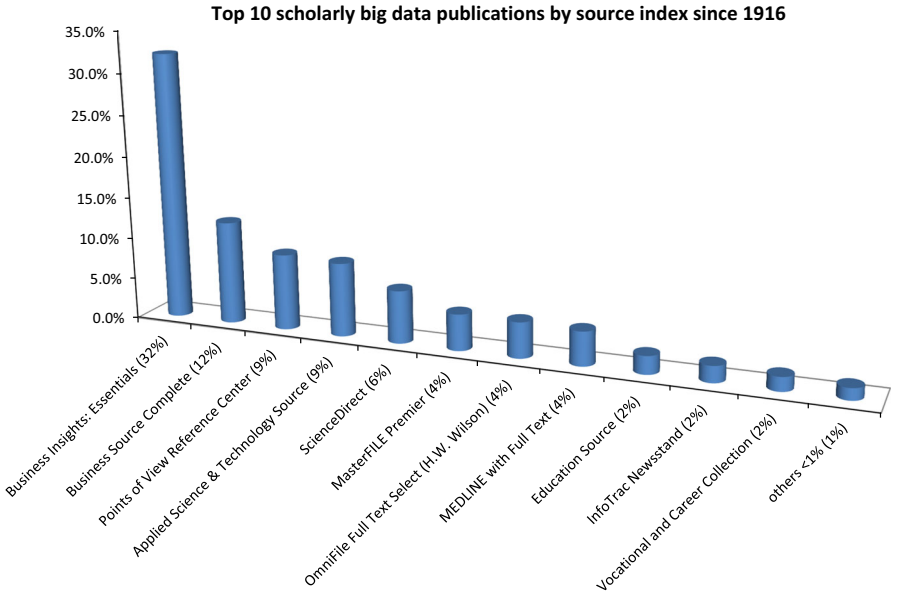

Fig. 1 Chart of top big data full text literature sources (1916–2016, N = 34 indexes)

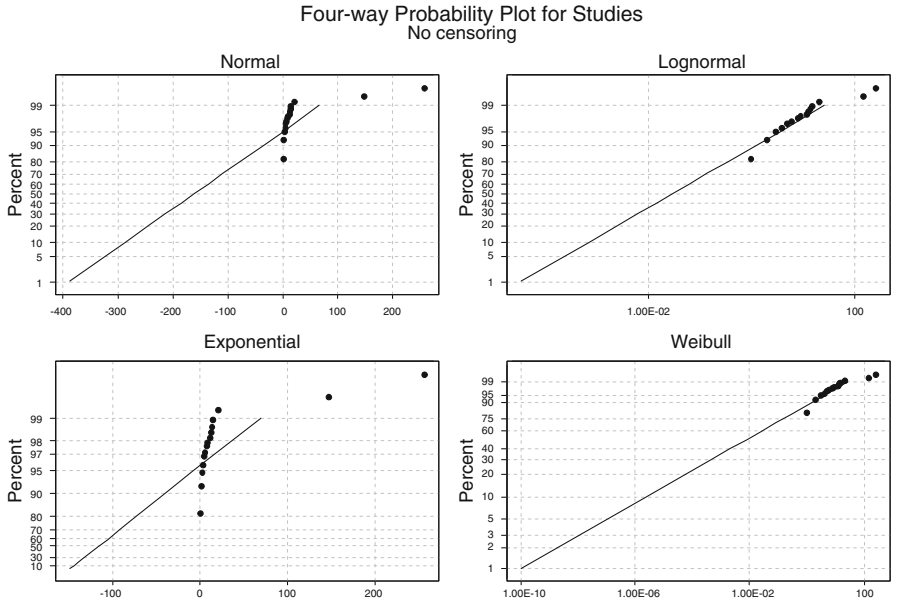

Fig. 2 Comparative distribution probability plots of big data studies (1916–2016, N = 79,012)

We then fitted the sample data into a Weibull distribution shape by creating an exponential trend equation. The result was that the Weibull distribution significantly captured 83% of the frequency variation by year ($r^2 = 0.8256,\ p < .01$, N = 79,012),

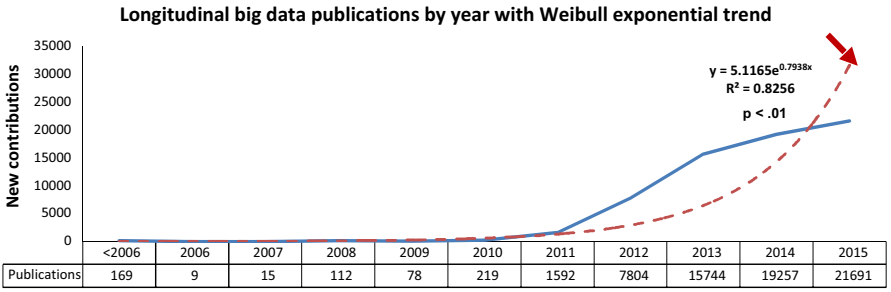

Fig. 3 Frequency of new big data studies by year with dashed trend line (1916–2015, N = 79,012)

Table 2 Big data articles chronologically by year (1916–2016)

Year	Publications	Percent	ReverseCum.%
<2006	169	0.2	100.0
2006	9	0.0	99.8
2007	15	0.0	99.8
2008	112	0.1	99.8
2009	78	0.1	99.6
2010	219	0.3	99.5
2011	1592	2.0	99.2
2012	7804	9.9	97.2
2013	15,744	19.9	87.3
2014	19,257	24.4	67.4
2015	21,691	27.5	43.0
2016*	12,322	15.6	15.6
Total	79,012		

* 2016 data consists of first six months

as shown in Fig. 3. We did not include 2016 data because it was midway through the year during our study. Figure 3 and Table 1 clearly show the production of big data literature commenced late in 2011 with 1592 new articles which increased almost 500% to 7804 by 2012, although the rate of growth slowed during 2015. We may interpret this as the scholarly big data research production is accelerating at an exponential rate.

The details are listed in Table 2. This table includes merges 1916–2006 data since there were only 169 articles produced and this marked one decade from the time of writing. The last column in Fig. 3 is the reverse cumulative percent of the frequencies. This tests the Pareto principle to determine if most (80%) of the big data research production was created in a few recent years. The test reveals that 97.2% of the big data research was produced since 2012 and 87.3% of the articles were published during 2013–2016.

Finally we looked at the big data production by outlet type, such as journals versus conferences or newspapers. In this analysis we included all outlet types and all the data. The results are summarized in Fig. 4, a bar chart ordered by highest frequency type, with the data values displayed below the axis. Journals and newspapers were similar, together accounting for approximately 70% of big data article production. It

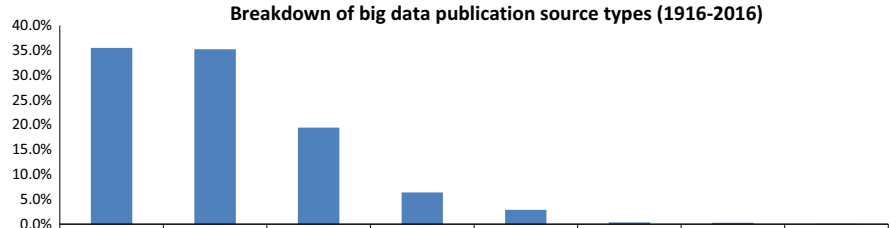

Fig. 4 Big data production by outlet type (1916–2016, N = 79,012)

was a surprise to see conference proceedings at such a low 0.3% contribution rate since new ideas often emerge through industry events and peer meetings.

4 Important Topics in the Big Data Body of Knowledge

In the previous section we established that the big data research production started in 2011 and it is rising exponentially. Next we briefly examined the 2011–2016 literature. We were focused on identifying what the important big data topics were and what attention was being given to privacy-related research.

First we compared the production of the top 10 big data keywords between journals and conference proceedings. Here we wanted to determine if certain topics were more likely to be studied in peer meetings at conferences or through projects published in journals. The results are summarized in Fig. 6, which shows frequency of the top 10 or so big data production topics for journals (blue line) versus conference proceedings (red line), in alphabetical order on the x-axis.

Generally, from the author's experience, it takes longer to publish a manuscript in a peer reviewed journal as compared with a peer reviewed conference. For this reason we assumed that there would be more unique keywords encountered in the conference proceedings. Ironically we found the opposite, as can be seen at the right of Fig. 6 where the majority of the journal topics were in the category '<1%' (N = 5888) as compared to conference in the '<1%' group (N = 1806). This means that more one-off unique topics were found in journals as compared to conferences during 2011–2016.

Another interesting finding was that there were more wholly theoretical generic type concept papers in the conference proceedings (3109 or 42%) as compared to journals (748 or 5%) during 2011–2016. This indicates that more than a third of the conference papers were conceptual in nature but only 5% of the journal manuscripts were purely theoretical. The deduction is that 95% of journal articles were focused on specific big data topics such as data mining or cloud computing instead of conceptual frameworks. We think big data studies require more time to complete and therefore it is more likely to see authors publish in journals rather than try to synchronize their projects with specific conference submission deadlines (Fig. 5).

Secondly, we zoomed into the production of big data studies published in academic journals. The previous analysis had revealed that most of the big data body of knowl-

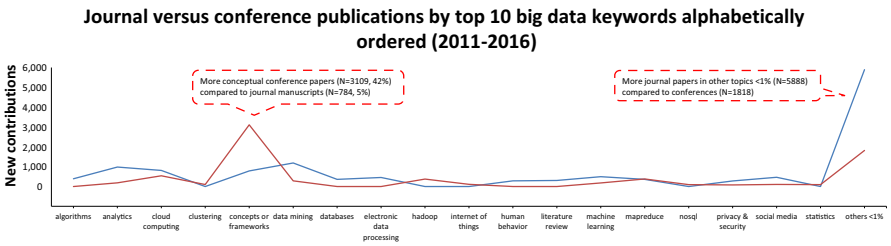

Fig. 5 Comparison of journal versus conference outcomes by top big data keywords (2011–2016)

edge outcomes were in journals (2011–2016, N = 13,029), almost double those in conference proceedings (2011–2016, N = 7445). The earlier results had also revealed that more unique topics had emerged in journals and there were substantially less wholly-theoretical papers in journals. When also considering practitioner experience that journal peer reviews are more rigorous than conferences and the journal submission deadline is not an arbitrary pressure to publish early, we narrowed our big data analytical lens to journals.

We downloaded and closely examined the 13,029 manuscript titles, abstracts and keywords published during 2011–2016 in journals. We used the title, abstract and keywords to nominate a dominate topic for every article to represent the big data body of knowledge. Our finalized big data body of knowledge resulted in 49 topics consisting of 1-3 words like 'data mining', 'artificial intelligence' and 'online social networks'. We edited the dominate keyword topics to ensure the spellings and meanings of the topics were consistent between the raters, such as changing plural to singular, and changing using the same underlying dominant topic; for example we changed 'database mining', 'mining analysis' to 'data mining'. We also recoded individual statistical techniques such as validity, correlation, regression, ANOVA, and so on, into the dominant topic 'statistics'. We selected the *Kappa* inter-rater statistical technique to ensure the resulting topic list was reliable this approach is considered rigorous since it removes the chance likelihood of agreement (i.e. the *Chi Square* expected probabilities). Since this technique requires an ordinal rating, the researchers rated the perceived ability of the keyword to fit the article, on a scale of 1–5 where 5 was the highest. We then performed an inter-rater agreement on the keywords resulting in a Kappa index of 0.93 ($p > .05$) which according to Cohen et al. [27] is considered a reliable agreement. This is comparable to a correlation of $r^2 = 0.8649$ or 87% indicating that the researchers were significantly in agreement about the dominate topic for each article.

We then factored the journal big data from 2011 to 2016 into a displayable short-list of 10–15 dominant topics using the frequency, and grouped all remaining low-count topics into a new category called '<1%'. This data reduction approach makes the information easily displayed but without compromising the relative frequencies. Since the full list of dominant topics (N = 49) was quite large, we elected to not display it at present (but will provide it upon request). Instead we included a revised dominant topic list in a later section of this article. We reformatted the pie chart into an exploded view to better illustrate the relative popularity of dominant big data topics published in journals during 2011–2016, as depicted in Fig. 6. The data labels in Fig. 6 represent

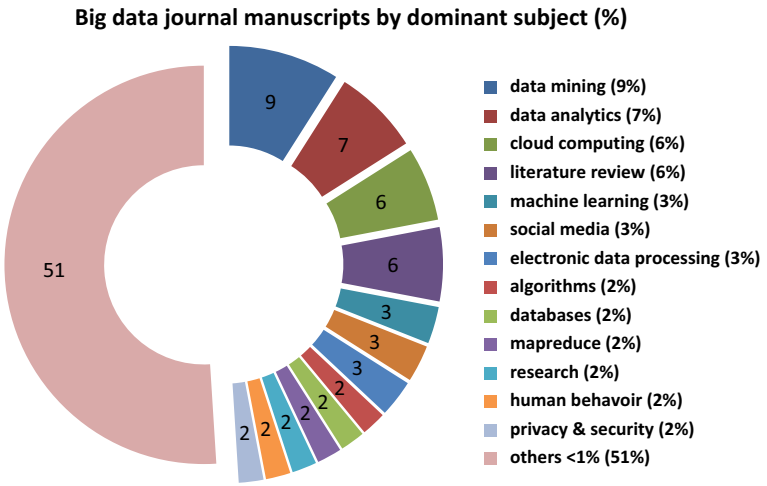

Fig. 6 Dominant big data topics published in journals by relative percentage (2011–2016, N = 13,029)

percentages not frequencies but note the last category of 'others <1%' means that over half of the 13,029 manuscripts had a unique dominant topic shared by only a few other researchers, or in some cases only there was only one study representing the keyword. We aimed to reduce the data to the top 10 big data topics but in reality we were able to accommodate 13 plus the 'others <1%' as a best-fit of the information.

The results from Fig. 6 reveal that the topmost of the 49 dominant big data topics published in journals during 2011–2016 was data mining (N = 1186) at 9.1%. The next three topics were similar in frequency and noticeably lower in application, namely data analytics (N = 979, 7.5%), cloud computing (N = 808, 6.2%), and literature reviews (N = 784, 6.0%). For reference purposes we would classify the current study as a big data literature review. Machine learning (N = 493, 3.8%) and social media (N = 466, 3.6%) came next but were a third less frequent than data mining. The following seven big data topics were somewhat equivalent in frequency: electronic data processing (N = 455, 3.5%), algorithms (N = 388, 3.0%), databases (N = 360, 2.8%), mapreduce (N = 358, 2.7%), research methods (N = 302, 2.3%), human behavior (N = 282, 2.2%) and privacy & security (N = 280, 2.1%). These 13 dominant topics represented 49% of the big data body of knowledge production in scholarly journals during 2011–2016.

As shown in Fig. 6, the remaining articles generated frequencies at or less than 1% so all were grouped into the '<1%' category which amounted to 6752 or 51% of the manuscripts during 2011–2016 in our sample. This other category included 36 topics like information technology, concepts or frameworks, hadoop, acquisition of data, computer algorithms, etc.

The above analysis answers our research question about the relative literature contribution towards big data privacy and security issues. The most important topics in the big data body of knowledge at least by frequency of journal manuscripts were as enumerated below. Note that all of these topics were specifically associated with big data and not a 'normal data' or unspecified type study.

1. data mining,
2. data analytics,
3. data cloud computing,
4. data literature reviews,
5. machine learning,
6. social media,
7. electronic data processing,
8. algorithms,
9. databases,
10. mapreduce,
11. research methods,
12. human behavior
13. privacy & security.

We can also answer the research question about the relative contribution of privacy and security topics within the scholarly big data body of knowledge which was 2% and ranked at 13 out of 49 in our sample. We argue that 2% is low in comparison to topics such as data mining at 9%. Furthermore we assert that less relative research is needed for specific software like mapreduce (2%) and hadoop (approximately 1%), which if reduced by half would provide more room to privacy and security, raising it to an expected percentage of 3.5%. We also propose that several topics related to the structure of the data, like databases and electronic data processing are similar enough to amount to duplication but if combined they would release another 3% of existing scholarly capacity to privacy and security, thus raising the latter expected frequency to 7.5%. There are additional 'other' topics that we question the relative importance of in the big data body of knowledge, such as why conceptual frameworks must be done separate of good literature reviews (combined they amount to 7.8%). In essence we argue that there should be a higher relative production of studies on privacy and security within the big data body of knowledge. To better explore that assertion we concentrated in more detail on reviewing contemporary journal manuscripts published this year (2016).

5 Current Research in the Big Data Body of Knowledge

In the previous section we identified the 13 dominant topics in the big data body of knowledge according to production frequencies in scholarly journals during 2011–2016. We found that privacy and security was in this top list but at 2% we argued that the relative contribution was lower than it ought to be based on the assumption that more effort was being put into less important topics. In this section we took a subsample of our data and examined in detail the content of the 2016 journal manuscripts. We were not able to conduct a meta-analysis because very few studies published quantitative benchmarks such as sample size of effect size.

Our earlier longitudinal analysis revealed that there were 12,322 manuscripts produced in 2016 related to big data keywords, with 1773 of those being published in journals (as compared to newspapers, conference proceedings and other outlets). We were able to access the full text manuscript of approximately half of the 1773 journal

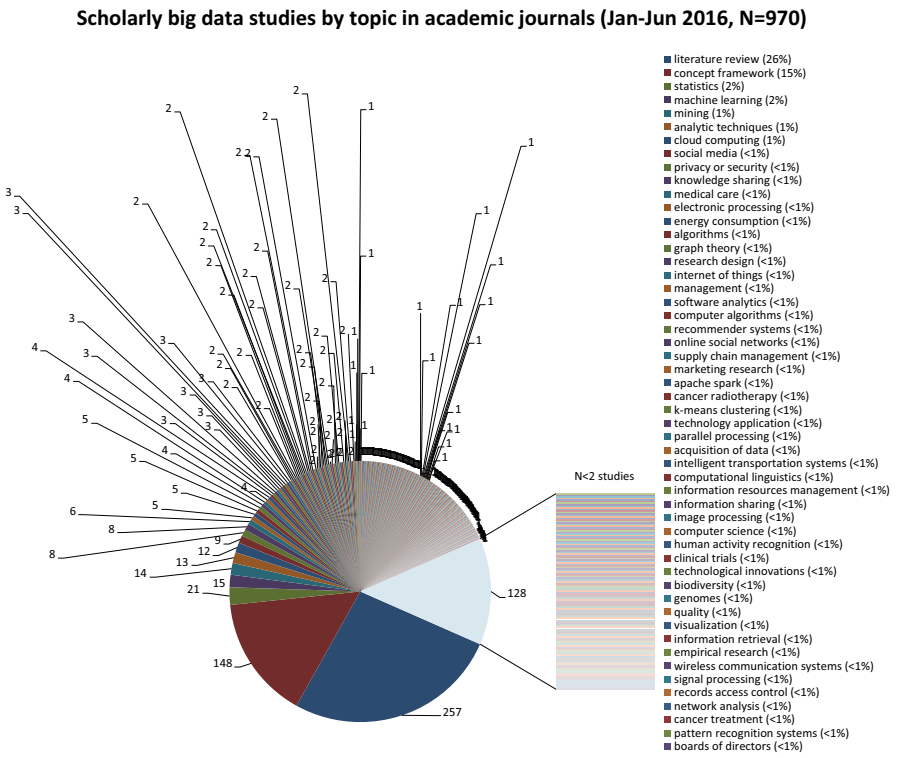

Fig. 7 Number of big data studies by best-fitting topic during January–June, 2016 (N = 970)

manuscripts due to the article simply not being accessible (private publisher, title and abstract indexing only, etc.). We realized the big data body of knowledge research is important only if it is shared so we concentrated on analyzing manuscripts that were available as full text.

We downloaded 970 journal manuscripts published during January-June 2016. We used carefully reviewed each manuscript, including the keywords (but not the title or abstract), using our dominant topic list produced earlier. We added, changed or deleted topics where needed. We calculated an inter-rater agreement of 100% on the resulting final topic list using the same approach as we discussed earlier. This was an acceptable level of reliability, which we attribute to the learning curve effect and not group think. We provide both the topic list and manuscript title in the appendix (it was too large to include in this paper). Additionally, we include the references for the current sample of 970 manuscripts. We will also provide these datafiles to other readers and researchers upon request.

Figure 7 is an explored pie chart customized to summarize our findings about the most current dominant big data topics published in scholarly journals within the last six months (January–June, 2016). In Fig. 7, the data labels represent the topic frequency with extender links connecting to the pie slice. The relative percentage of each big data topic is shown beside the label within the legend.

The current big data body of knowledge pie chart in Fig. 7 illustrates a vastly different trend as compared to the longitudinal or 2011–2016 sample. The 257 literature reviews at 26% of all production were the more frequent type of manuscript published in journals thus far in 2016. By contrast to Fig. 6 literature reviews constituted only 6% of the entire big data body of knowledge literature including this year (2011–2016). This reveals a relative increase of more than 400% towards literature reviews in the big data journal outlets. We also discovered a higher relative production of conceptual papers, at 148 or 15% which is a much larger share than the approximately 1% overall (see Fig. 6 to compare). This is almost 15 times more current focus on concepts or frameworks in journals during the last six months within the big data body of knowledge.

The next two dominant topics were 21 statistics articles (2% rounded) and 15 machine learning manuscripts (2% rounded). Statistics amounted to less than 1% overall while machine learning constituted 3.8% of the topics during 2011–2016. We believe the increased current focus on statistics is a result of our more accurately coding algorithms and analytics into recognized statistical techniques. This difference between 2011 and 2016 and the current subsample is not significant. However, machine learning has increased in interest during 2016. This could be due to researchers applying predictive data analytics using machine learning as opposed to using generic statistical routines like regression or forecasting. A close parable to this as an example only the premise of the well-known American TV show 'Person of Interest' where machine learning and artificial intelligence were used to build a computer that can predict a terrorist as well as identify an individual who is at risk of being killed.

There were only 14 articles focused on data mining (approximately 1%) as compared 9.1% of the overall relative production in the big data body of knowledge during 2011–2016. This is a substantial decrease of interest in data mining which we suggest is a result of the topic having become saturated by the end of 2015. Additionally, we think researchers are developing more specific less commercial approaches to examine big data, perhaps using open software, yet concentrating on the core topic of research and less on the technology aspect. For example we noted unique topics emerging such as software analytics focused on examining good versus bad code perhaps for viruses, as differentiated from hardware or data analytics. We also saw subjects like recommender systems (marketing discipline software) and topics like graph theory emerge in the current big data in addition to visual analytics.

Following our earlier practices, we allocated an 'other' category to group less used topics with less than 2 studies published, of which there were 128 articles amounting to 13% of the relative current production in the big data body of knowledge. Some of these unique topics included keywords like anemia (<1%), air quality (<1%), estrogen receptors (<1%), agricultural industries (<1%), environmental auditing (<1%), stratospheric aerosols (<1%), emergency service administration (<1%), raman spectroscopy (<1%), electron microscopy (<1%), speciation (<1%), electromyography (<1%), and zooplankton (<1%). We assert this proves that other disciplines, particularly healthcare and the physical sciences (agriculture-aquaculture, chemistry and geography).

Previously popular topics were less likely to be found in 2016. Table 3 is a comparison of the top 13 big data topics during 2011–2016 with the current findings for this

year. The last column is an interpretation of the topic from the perspective of the current period. The data is double counted in the first column since the earlier sample of 13,029 articles included the current subsample of 970 manuscripts. This amounts to 7% of the original sample but it was too difficult to eliminate the subsample retroactively.

There were significant increases of literature reviews (26%) and conceptual frameworks (15%) which impacted all other topics given that N = 970 for the current year as compared to N = 13,029 for the earlier analysis. Machine learning (2%) and statistics (2%) garnered about the same relative attention in the current period. The remaining big data topics decreased in frequency during the current sample frame, namely: data mining (1%), data analytics (1%), cloud computing (1%), social media (<1%), privacy or security (<1%), electronic processing (<1%) and algorithms (<1%).

There was a significant reduction of several topics from the current big data body of knowledge although they were still in the sample, namely: social media (<1%), databases (<1%), mapreduce (<1%), research methods (<1%) and human behavior (<1%).

An interesting observation from Table 3 was the emergence of three new popular terms in the current big data body of knowledge: knowledge sharing (<1%), medical care (<1%), and energy consumption (<1%). We attribute this to the healthcare and education disciplines where knowledge sharing and medical care are now becoming more involved in the big data field. Additionally, since there has been an incredible awareness of global climate change, the physical and social sciences have begun to publish their studies about improved energy methods and pollution reduction, due to the availability of large data samples. All of these topics were popular in the general literature but 2016 has marked the point when many disciplines are entering the big data paradigm as researchers are acquiring larger and more complex data for their empirical studies.

6 Conclusions

In this paper, we extended the big data body of knowledge by analyzing the longitudinal and current literature to highlight important research topics and identify critical gaps. Our motive was that big data is a relatively new paradigm and numerous researchers had claimed that more research was needed. Many researchers has also asserted that privacy and security were not receiving sufficient attention in the scholarly big data literature.

We analyzed a large amount of big data literature to confirm the above assentation's. We initially collected meta-data on 79,012 articles from all sources during 1900–2016 related to big data to analyze longitudinal research production trends. We determined that the big data paradigm commenced in late 2011 and the research production exponentially rose starting in 2012. From there, we refined our sample to the timeframe of 2011–2016. We confirmed a Weibull exponential trend of big data knowledge production that captured 82% of the variance increase ($p < .01$). Next we examined the abstract and keyword content of the 13,029 articles in the sample to produce a dominant topic list for the big data body of knowledge. The result was 49 dominant big data topics which we calculated was valid for the sample with an inter-rater reliability of

Table 3 Comparison of 2011–2016 versus current top big data topics published in journals

Top big data topic in 2011–2016	Top current big data topic	Interpretation of current year
Data mining (9%)	Literature review (26%)	Significant increase
Data analytics (7%)	Concept framework (15%)	Significant increase
Cloud computing (6%)	Statistics (2%)	Similar trend
Literature review (6%)	Machine learning (2%)	Similar trend
Machine learning (3%)	Data mining (1%)	Significant decrease
Social media (3%)	Data analytics (1%)	Significant decrease
Electronic data processing (3%)	Cloud computing (1%)	Significant decrease
Algorithms (2%)	Social media (<1%)	Small decrease
Databases (2%)	Privacy or security (<1%)	Small decrease
Mapreduce (2%)	Knowledge sharing (<1%)	New topic
Research (2%)	Medical care (<1%)	New topic
Human behavior (2%)	Electronic processing (<1%)	Small decrease
Privacy & security (2%)	Energy consumption (<1%)	New topic
Others <1% (51%)	Algorithms (<1%)	Small decrease

93% ($r^2 = 0.89$). We found there were 13 dominant topics that captured most (49%) of the big data production research in journals. Privacy and security related topics accounted for approximately 2% of the total journal outcomes during 2011–2016.

We then took a more detailed subsample of 970 journal manuscripts produced during the first of 2016 and we analyzed these in detail. This data illustrated a vastly different trend as compared to the longitudinal 2011–2016 sample. Literature reviews increased 400% from 6 to 26%, and conceptual papers went from less than 1–15% of the total current big data knowledge production in journals. Together, these types of non-empirical studies accounted for 41% of the current big data knowledge production, which we asserted was far too high. There was too little concentration of current research into other important big data topics as discussed below.

Several big data topics are currently on par with expectations but require no further attention, such as machine learning and statistics. Other formerly popular big data topics have significantly decreased in application, namely data mining, data analytics, cloud computing, social media, privacy or security, electronic processing and algorithms.

Interestingly, new topics emerged in the current big data literature, including anemia, air quality, estrogen receptors, agricultural industries, environmental auditing, stratospheric aerosols, emergency service administration, raman spectroscopy, electron microscopy, speciation, electromyography, and zooplankton. We attributed this increase of novel big data topics to the healthcare and education disciplines where knowledge sharing and health/medical care researchers are now becoming more involved in the big data paradigm. Additionally, since there has been an incredible awareness of global climate change, the physical and social sciences have begun to publish their studies about improved energy methods and pollution reduction, as well as owing to the availability of larger more complex data samples. We believe other

disciplines will enter the big data field and further expand the body of knowledge in the coming year.

We answered the research question concerning the status of big data privacy and security research. We determined that only 2% of the big data body of knowledge production was focused on the privacy and security topic during 2011–2016. We developed a unique expected frequency of 7.8% for big data privacy and security production by relative comparison to other topics. The determined that the current big data privacy and security output in 2016 was slightly less than 1% and well below our expected frequency. Thus, we recommend more production of big data privacy and security related studies.

In conclusion, we have presented a novel list of 49 important topics representing the status of the big data body of knowledge. These were empirically developed from the scholarly literature, primarily from peer reviewed journal articles, and weighted towards the current trend since we are early in the production cycle for the big data paradigm (it has evolved during only 5 years from 2011–2016). We also confirmed the status of privacy and security related research production in the big data field to be low (recently dropping from 2 to 1%) and below the expected relative frequency of 7.8%.

Privacy and security is an important social topic and a critical dimension of big data in most disciplines and industries. It is difficult to find a discipline or industry where privacy or security is not important, but since big data is everywhere, therefore we ought to see more studies using big data. Based on our findings, we call for consumers, scholars and decision makers to address this gap in the big data body of knowledge by asking for and funding more research about privacy and security. We believe the way to do that is to encourage special issues in journals on the topic of big data privacy and security, targeted at specific disciplines, and encouraging applications or empirical studies. It is clear that we have seen enough literature reviews and conceptual models in the big data body of knowledge because these account for 41% of the topics current journal production (January–June 2016). More applied and empirical big data studies are needed. In closing we will make our data available to anyone by request.

References

1. Chen M, Mao S, Zhang Y, Leung VC (2014) Open issues and outlook in big data. In: Chen (ed) Big data: related technologies, challenges and future prospects, vol 1. Springer, New York, pp 81–89
2. Goldfield NI (2014) Big data–hype and promise. J Ambul Care Manag 37(3):195–196
3. Kambatla K, Kollias G, Kumar V, Grama A (2014) Trends in big data analytics. J Parallel Distrib Comput 74(7):2561–2573
4. Kim G-H, Trimi S, Chung J-H (2014) Big data applications in the public sector. Commun ACM 57(3):78–85. doi:10.1145/2500873
5. Pence HE (2015) What is big data and why is it important? J Educ Technol Syst 43(2):159–171. http://journals.sagepub.com/doi/abs/10.2190/ET.43.2.d
6. Bohannon J (2015) Credit card study blows holes in anonymity. Science 347(6221):468
7. De Zwart M, Humphreys S, Van Dissel B (2014) Surveillance, big data and democracy: lessons for Australia from the US and UK. Univ N South Wales Law J 37(2):713–747
8. Eastin MS, Brinson NH, Doorey A, Wilcox G (2016) Living in a big data world: predicting mobile commerce activity through privacy concerns. Comput Hum Behav 58:214–220

9. Ekbia H, Mattioli M, Kouper I, Arave G, Ghazinejad A, Bowman T, Suri VR, Tsou A, Weingart S, Sugimoto CR (2015) Big data, bigger dilemmas: a critical review. J Assoc Inf Sci Technol 66(8):1523–1545
10. Gharabaghi K, Anderson-Nathe B (2014) Big data for child and youth services? Child Youth Services, pp 193–195
11. Kshetri N (2014) Big datas impact on privacy, security and consumer welfar. Telecommun Policy 38(11):1134–1145
12. Lichtblau E, Weilandaug N (2016) Hacker releases more democratic party files, renewing fears of Russian Meddling. New York Times, New York
13. Dana M (2012) On Orbitz, Mac users steered to pricier hotels. Wall Streat J 1–3. http://www.wsj.com/articles/SB10001424052702304458604577488822667325882
14. Duhigg C (2014) The power of habit: why we do what we do in life and business. Penguin Random House, New York City
15. Filkins BL, Kim JY, Roberts B, Armstrong W, Miller MA, Hultner ML, Castillo AP, Ducom J-C, Topol EJ, Steinhubl SR (2016) Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Trans Res 8(3):1560–1580
16. Hoffman S, Podgurski A (2013) Big bad data: law, public health, and biomedical databases. J Law Med Ethics 41:56–60
17. Thorpe JH, Gray EA (2015) Law and the Public's Health. Big data and public health: navigating privacy laws to maximize potential. Public Health Rep 130(2):171–175
18. Ward JC (2014) Oncology reimbursement in the era of personalized medicine and big data. J Oncol Pract 10(2):83–86
19. Booch G (2014) The human and ethical aspects of big data. IEEE Softw 31(1):20–22
20. Leszczynski A (2015) Spatial big data and anxieties of control. Environ Plan D Soc Space 33(6):965–984
21. Rothstein MA (2015) Ethical issues in big data health research: currents in contemporary bioethics. J Law Med Ethics 43(2):425–429
22. Shull F (2014) The true cost of mobility? IEEE Softw 31:5–9
23. Solove DJ (2013) Introduction: privacy self-management and the consent dilemma. Harvard Law Rev 126(7):1880–1903
24. Vaidhyanathan S, Bulock C (2014) Knowledge and dignity in the era of big data. Ser Librarian 66(1–4):49–64
25. Wang H, Jiang X, Kambourakis G (2015) Special issue on security, privacy and trust in network-based big data. Inf Sci 318:48–50
26. Jovanovi U, Stimec A, Vladusi D (2015) Big-data analytics: a critical review and some future directions. Int J Bus Intel Data Mining 10(4):337–355
27. Cohen J, Cohen P, West SG, Aiken LS (2003) Applied multiple regression/correlation analysis for the behavioral sciences, 3rd edn. Lawrence Erlbaum Associates, Mahwah