Journal of Computer Information Systems

ISSN: 0887-4417 (Print) 2380-2057 (Online) Journal homepage: http://www.tandfonline.com/loi/ucis20

Privacy and security in the big data paradigm

Zhaohao Sun, Kenneth David Strang & Francisca Pambel

JOURNAL OF COMPUTER INFORMATION SYSTEMS
https://doi.org/10.1080/08874417.2017.1418631

Taylor & Francis
Taylor & Francis Group

Privacy and security in the big data paradigm

Zhaohao Sun[a], Kenneth David Strang[b], and Francisca Pambel[a]

[a]Department of Business Studies, PNG University of Technology, Lae, Morobe, Papua New Guinea; [b]Plattsburgh, School of Business & Economics, State University of New York, Queensbury, NY, USA

ABSTRACT
Privacy and security in the big data age have drawn significant attention in the academia and industry. This article examines privacy and security in the big data paradigm through proposing a model for privacy and security in the big data age and a classification of big data-driven privacy and security. It extends the big data body of knowledge, highlights important research topics, and identifies critical gaps through statistical analysis of big data and its impacts on privacy and security based on literature data published from 1916 to 2016. It also presents the state-of-the-art privacy and security based on the analysis of SCOPUS data from 2012 to 2016. The result shows that privacy and security face new challenges and require new policies, technologies, and tools for protecting privacy in the big data paradigm. The proposed approach might facilitate the research and development of privacy and security, and big data-driven privacy and security in terms of technology, governance, and policy development.

KEYWORDS
Big data; big data paradigm; privacy and security; healthcare; exponential Weibull trend; big data analytics

Introduction

Big data is a new paradigm with almost all of the scholarly literature having emerged since 2012.[1,2]Privacy and security are important for everyone in the age of big data. For example, one concerns the abuse of medical privacy through leaking personal information to the employer so that the hiring decisions have been made unfairly based on the medical history. [2] Big data privacy and security risks are hot topics in the media as well as in academic research.[3–7] 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. [7–11] 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, 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]

BBC reported that a massive cyber-attack using ransomware struck organizations of 99 countries around the world in May 2017 [15] Among the worst hit was the National Health Service (NHS) in England and Scotland. Then about 40 NHS organizations and some medical practices cancelled operations and appointments. NHS staff shared screenshots of the WannaCry program, which is a ransomware demanding a payment of $300 (£230) in virtual currency Bitcoin to unlock the files for each infected computer. [15] This latest cyber-attack to hospitals implies that there is a need for more research into privacy and security in the big data paradigm. Furthermore, cyber-attacks against the hospital's power infrastructure and water supply and ransomware to shut down the hospital servers imply that healthcare big data privacy and security are an important yet unpopulated body of knowledge [11,17–19], because hospitals' servers could significantly impact patient care and healthcare. [16]

There have been a significant number of research publications on big data privacy and security.[2,7,11] However, the majority of recent literature states that we need more research on big data privacy and security (e.g. [8–10,13,15,17–21]). However, they have not further classified privacy and security in the big data age although ACM provided a classification of privacy and security in a general setting in 2012. [22]This is a compelling justification to examine the status of privacy and security research in the big data paradigm. Based on this above discussion, the following three issues are significant for privacy and security in the big data paradigm:

(1) What are the interrelationships among privacy, security, and big data?
(2) How can we classify big data-driven privacy and security?
(3) What is the status of privacy and security research in the big data paradigm and the state-of-the-art privacy and security in the big data age?

This paper will address these three issues. More specifically, it addresses the first issue by presenting a Boolean model for big data privacy and security based on the Boolean algebra. It addresses the second issue by presenting a classification of big data-driven privacy and security based on the characteristics of big data and literature review. It

CONTACT Zhaohao Sun zhaohao.sun@gmail.com Department of Business Studies, PNG University of Technology, Private Mail Bag, Lae 411, Morobe, Papua New Guinea

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/UCIS.

addresses the third issue by providing a statistical analysis of big data and its relationships with privacy and security based on the literature data published from 1916 to 2016. It also addresses the third issue by analysis of SCOPUS searched data (2012–2016).

The rest of this paper is organized as follows. Section 2 overviews the approach and methodology for this research. Section 3 proposes a Boolean model for big data privacy and security. Section 4 presents a classification of big data-driven privacy and security. Section 5 provides a statistical analysis of big data-driven privacy and security. Section 6 looks at the state-of-the-art privacy and security in the big data age. Section 7 discusses the implications and limitations of this research, and the final section summarizes the findings and proposes recommendations for future research.

Approach and methodology

To address the mentioned research issues, the research uses a logical approach (Boolean approach) to reveal the interrelationships among privacy, security, and big data at three levels: fundamental, interactional, and integrated levels. This research also uses a qualitative approach (i.e., literature review or secondary data analysis) to classify the big data-driven privacy and security based on the characteristics of big data. Then the research uses meta-analysis and quantitative approach (i.e., statistical analysis) to investigate privacy, security, and big data and reveals how important privacy and security are in the big data age. Then the research uses data analysis and statistical modeling to look at state-of-art privacy and security in the big data age, taking into consideration the classification of privacy and security of ACM in 2012.

From the viewpoint of research methodology, the proposed Boolean model of big data privacy and security reveals the interrelationships among big data privacy and security. The classification of big data-driven privacy and security reflects further investigation into privacy and security in big data in the proposed Boolean model, based on an in-depth analysis of big data and its characteristics. The statistical analysis of big data-driven privacy and security reflects the status of privacy and security research in the big data paradigm. The proposed state-of-the-art privacy and security in the big data age are based on the classification of privacy and security of the ACM in 2012 and searched document results of SCOPUS (2012–2016).

A Boolean model for big data privacy and security

This section proposes a Boolean model for privacy and security in the big data age and examines their interrelationships.

From a Boolean structure viewpoint [23], big data privacy and security and their interrelationships can be illustrated in Figure 1. This can be called a Boolean model for big data privacy and security.

In what follows, we will elaborate this model to some detail.

Privacy is the "claim of individuals to be left alone, free from surveillance or interference from other individuals, systems or organizations"[[24], p. 168]. In the area of healthcare, privacy is defined as "an individual's right to control the acquisition, uses, or disclosures of his or her identifiable health data".[25] Privacy consists of two key points: confidentiality and fair use. [2] To protect confidentiality, the privacy-enhancing technologies and systems can be used to "enable users to encrypt email, conceal their IP address to avoid tracking by Web server, hide their geographic location when using mobile phones, use anonymous credentials, make untraceable database queries and publish documents anonymously". [2] However, dealing with the fair use of privacy has been a big issue although there are laws, regulations, and policies in every culture and country. [24]

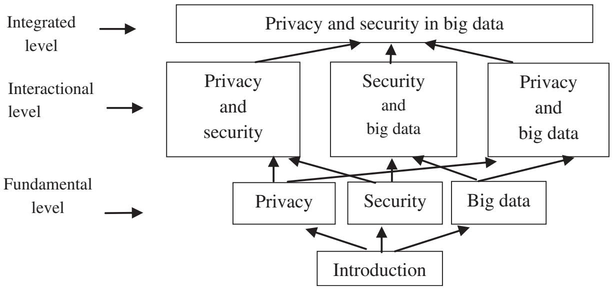

Figure 1. A Boolean model for big data privacy and security.

Security refers to the policies, procedures, and technical measures used to prevent unauthorized access, alternation, theft of data, or physical damage to devices and systems [[24], p. 338]. In healthcare, security is defined as "physical, technological, or administrative safeguards or tools used to protect identifiable health data from unwarranted access or disclosure".[25]Big data is defined as the data sets that are so large and complex that traditional data-processing applications are not sufficient. [7] Big data is generated from various instruments, billions of phones, payment systems, cameras, sensors, Internet transactions, emails, videos, click streams, social networking services, and other sources.[26] Big data has several big characteristics, for example, big volume, big velocity, big variety, and big veracity. [27,28] Big data has become a strategic asset for industry, business, and healthcare, and a strategic enabler of exploring business insights and economics of services as well. [2,29–31]

Big data brings about big issues to privacy and security. For example, privacy and security at least have technical, organizational, and environmental issues in big data. These issues result from technical deficiencies, organizational culture, and environmental factors [[24,32], p. 338]. How to secure queries over encrypted big data is a big issue on privacy and security in cloud services. [33] Mobile heath is an innovative approach to deliver healthcare in an accessible, portable, and cost-effective manner [25]; people in the big data age can easily use health apps (in mobile devices) to manage lifestyle, diet and fitness, drug reference, diagnosis, and treatment as well as chronic disease. However, customers are still concerned about the privacy on mobile healthcare, because smartphones pose the greatest privacy and security risks [[24], p. 341]. The smartness of mobile phones implies that mobile phones connect to the Internet. On the Internet, nearly every piece of data and information can be monetized by the data-dominated companies through deep data processing using deep learning.[2]

Big data also brings about big challenges and risks to privacy and security in the big data age. [32] These risks to privacy and security may possibly become bigger by big data's big volume, big variety, and big veracity for supporting big data applications in mobile cloud environments, where one's whereabouts, locations, habits, and friends have been tracked by marketing company at best and by terrorist tracking system at worst,[24] [p. 173, [11]]. Some individuals do not like to expose where they are to others, because they believe this is their privacy. However, iPhones and Android phones have been secretly sending information about users' locations to Apple and Google, respectively. [11] A few data companies can control and access most of the personal data of the world's population and almost all of the data on the Web. This is one of the biggest risks to privacy.[2]

Big data is the gold mine of the 21st century. [27] A significant number of big data companies have been tracing all the data of individuals online or offline, on public occasion or in private sphere. The individual does not know how to monetize the collected data. However, the big data companies really know what kind of data can be innovated, sometimes, as a secondary use,[29] [p. 153] to which kind of companies in order to monetize itoptimally. [2] Therefore, an individual is vulnerable in the face of global big data companies, because the traditional security solutions are not designed to protect individual privacy in the big data age. [11] The reason why European countries cannot have global big data companies might be because in Europe, privacy protection is much more stringent than in the USA,[24] [pp. 170–5]. That is, the European countries do not allow businesses or data companies to use personally identifiable information without the individual's prior consent. The companies must inform the individuals when they collect information about them and disclose how it will be stored and processed. This is an opt-in approach[24], a conservative and user-friendly data collection model. However, the USA allows businesses to gather information and use it for other marketing purposes without obtaining the informed consent from the individual whose information is being collected and then used. [24] This is the opt-out approach of data collection, an aggressive model. The opt-out data collection has produced a few global giants of big data, and it also makes the most individuals in a relatively disadvantaged position. [11] Meanwhile, the countries and organizations complying by the opt-in data collection principle have also been in a more disadvantaged position.

To resolve each of the above-mentioned issues encountered in the big data age, new laws, policies, regulations, safeguards, and industry standards and tools need to be developed for securing the privacy of individuals and organizations. [16,26] For example, Facebook recently announced new privacy tools to help its customers when intimate images are shared on Facebook without their permission, and help build a safe community of its billions of customers. [34] More authentication procedures should be put in place to address security. [26] The collection and transmission of big data, in particular, big sensitive data, through any communication networks will essentially introduce critical requirements for privacy and security in big data, for which the possible solution is to provide security technologies and solutions that are able to secure communication platforms and scale to the large size of data, taking into consideration the big velocity of big data. An organization must ensure that data are aggregated or anonymized to prevent any unauthorized access to personal identifiable information sets during the rapid transmission, creation, and processing of big data.

Almost every country has their own laws governing privacy and security to secure individuals' privacy and protect their private data.[3,24,35] For example, in the USA, federal legislation includes Health Insurance Portability and Accountability Act (HIPAA) of 1996[24], The Federal Trade Commission Act, and the Children Online Privacy Protection Act (COPPA) of 1998 for mobile health. [26] However, existing legislations are basically country-based, whereas the big data age is global in essence. Therefore, there are still gaps in the protection of an individual's privacy. Furthermore, the growing and widespread use of mobile or cloud services in the big data age has not been covered by the existing regulations of privacy and security in the existing legislation, because the existing security solutions have not considered the big volume, big velocity, variety, and veracity as well as complexity of big data properly.[11,16] Therefore, it is significant to look at the privacy and security taking into account the characteristics of big data.

A classification of big data-driven privacy and security

This section looks at privacy and security in the big data age by classifying it into four categories based on the characteristics of big data, i.e., big volume-driven privacy and security, big velocity-driven privacy and security, big variety-driven privacy and security, and big veracity-driven privacy and security. The key idea behind this classification is that the big volume, big velocity, big variety, and big veracity of big data increase new privacy and security challenges far beyond those individuals faced a decade ago. [7,16]

There are close relationships between various inherent characteristics of big data privacy and security from the perspective of data processing and management [11,24,27,32], as shown in Table 1, which leads to the classification of big data-driven privacy and security. The data processing and management include data collecting, copying, scanning, storing, analyzing, reusing, accessing, and sharing, to name a few.[22]

Table 1. Privacy and security and characteristics of big data.

Characteristics of big data	Privacy	Security
Big Volume[18,23]	Create big value Data is power! Data is money.	Contribute to a big number of cybercriminals.[11]
Big Velocity[18,23]	Real-time location data. [11]	Physical security risks.[11] Generate profile of the individual's click and position.[24]
Big Variety[18,23]	Cannot effectively manage data containing sensitive information.	Many companies have not properly secured and protected the unstructured data.[2]
Big Veracity[18,23]	Time variant data of individuals are concerns related to privacy [24]	Security breach related to a big number of credit cards.[24]

Table 1 implies that big data and its characteristics have a significant impact on privacy and security. This leads to an in-depth analysis of big data-driven privacy and security, taking into account each of big characteristics of big data.

Big volume-driven privacy and security

The big volume of big data reflects the size of the data set, which is typically in exabytes (EB, 2^{60}B) or zettabytes (ZB, 2^{70}B).[36] Nowadays, many giant data companies are working with petabytes (PB, 2^{50}B) of data daily. For example, Google processes over 20 PB of data daily.[38] Walmart collects more than 2.5 PB of unstructured data from its one million customers every hour. Big data repositories for future generation parallel and distributed systems currently exceed EB and are increasing at a rapid pace in size.[37] The volume of big data has been increased at the EB or ZB level. [7]

The big volume of big data has challenged privacy and security. For example, organizations may not securely store big amounts of data and manage the collected data during peak data traffic. The big volume of personal private data can be used to create big commercial value through multiple data-processing techniques, including big data analytics. [27] The big volume of personal private data in a data center or cloud platform is also a goldmine for cybercriminals, when confidential data and information are intercepted or disclosed, lawfully or otherwise.[11,34]Data is power, data is money. These have become reality in the big data industry. The big volume of big data would likely attract a big deal of cybercriminals. Juniper Research estimated in 2016 that the costs of cybercrime could be as high as 2.1 trillion by 2019[34]

Big velocity-driven privacy and security

Big velocity is related to big throughput and latency of data. [39,40] Big throughput means that at a big speed, data in and out from the networked systems in a real time. [36,37] In other words, it is the high rate of data and information flowing into and out of the interworked systems with real time. [30,38]

Latency is the other measure of velocity. Low latency is the requirement of modern business and individual. For example, Turn (www.turn.com) is conducting its analytics in 10 milliseconds to place advertisements in online advertising platforms[39], [p. 3, 40].

Big velocity of big data is more important than big volume for many real-world applications. [7] In some developing countries, many networked printers cannot be running sometimes because of low speed, to our knowledge.

Real-time location data is a key component of big velocity data, which is also a big privacy concern. [11] For example, big velocity data increases significant privacy risks because of real-time profiling, behavioral targeting, and tracking techniques on clickstream data and real-time position[24], [pp. 169, 434, [2]].

Big velocity data also increases physical security risks, because the real-time click behavior and position of individuals have been tracked [11] using big data analytics, which is used to collect, store, reuse, analyze, create profile, and visualize profile of the individual's click and position[24], [pp. 172–5]. On the contrary, big data analytics can protect the privacy of individuals related to click behaviors and positions.[27]

Big variety-driven privacy and security

Big variety means the big diversity or big different types of data sources with different structures from which it arrived, and the types of data available to nearly everyone.[35,36,41] Big data can be classified into three types at a higher level: structured, semi-structured, and unstructured. The data stored in relational database systems like Oracle are structured. The data available on the web are unstructured. In total, 80% of the world's data are unstructured.[35,39] Blogs and tweets on social media are not structured data, because they contain a large amount of slang words, with a mix of languages in a multiethnic, multi-language environment [[39], p. 41]. The big variety exists in the data on the Web. For example, in the WeChat world (https://web.wechat.com/), one can interact with his/her friends anywhere in the world using a variety of media such as texts, sound clips, photos, and videos.

Time-variant data of individuals such as credit card information are concerns related to privacy. [24] However, such information has been available in communication using smartphones via, for example, WeChat. Companies suffer high-profile security breach related to a significant number of credit and debit card accounts and personal data. [24]

Organizations cannot effectively manage unstructured data containing sensitive information. A big variety of sensitive information would make it difficult to detect security breaches. Many companies have not properly secured and protected the unstructured data, which later fall into dangerous hands, due to political pressure and other reasons. [2]

Big veracity-driven privacy and security

Veracity refers to the accuracy and truthfulness of big data. Veracity was introduced by IBM data scientists as the 4th V as a characteristic of big data. [42] There exists big ambiguity, incompleteness, and uncertainty of big data. [30,35,42] This might be the reason why vagueness has been considered as one of the 10 challenges of big data. [30] Accuracy and reliability are less controllable for many forms of big data, for example, in the case of Twitter posts with hash tags, abbreviations, typos, and colloquial speech. Big veracity is a big characteristic of big data, and this is particularly true in big data analytics for business decision-making.[27,43–46] Therefore, in order to get big veracity, we have to use big data technology to remove the ambiguous, incomplete, uncertain data.

Fuzzy logic and fuzzy sets have developed significant methods and techniques to address ambiguity and incompleteness of data and they will play an important role in overcoming big ambiguity and incompleteness of big data. [7,38]

Big veracity implicates big complexity of big data, for example, thousands of features per data item, the curse of dimensionality, combinatorial explosion, many data types, and many data formats.[28]

Health-related data such as disease symptoms, patients' prescriptions, and online purchases of medical supplies are big concerns related to privacy. A third-party company can collect, store, and analyze the health-related data to produce medicines and market a special area and group of patients.[1,16]

The big data age intensifies the information asymmetry between the individuals and the data companies, between the small organizations and global big data giants. [2] This puts the individuals and small organizations in a more disadvantaged position and creates significant vulnerabilities regarding privacy and security [47–48], because the individuals or small organizations cannot completely understand the related privacy and security policies[24], [p. 176]. They often lack the ability, skill, desire, and motivation to collect data and evaluate optimal amount of data; they also lack essential data-processing skills and advanced big data analytics tools[11,24], [p. 344]. They neither comprehend how to manage complex systems nor know how other companies and businesses are using their personal information, real-time location information.[11,47]

Statistical analysis of big data-driven privacy and security

This section provides a statistical analysis of big data and its impacts on privacy and security to address what is the status of privacy and security research in the big data paradigm.

Data collection and methods

Data were collected from publishers through their searchable literature indexes (N = 34). [1] Some sources embedded other indexes such as ABI/Inform, ACM Digital Library, Bacon's Media, Cabell's, DBLPDBLP, Index Copernicus, INSPEC, and SCOPUS. Although additional indexes existed, when searching them we found severe duplication, so based on the principle of diminishing returns, Google or other public domain search engines have not been used to obtain an accurate result with as few duplicates as possible to filter out.

The collected data can be considered big data. The sample had a high variety/complexity with presumed veracity (accuracy) and value although it did not meet the other two V's of big data, namely volume and velocity. [1] The collected full text data reached an estimated size of 100,000 MB when considering an article averages 1 MB including images and tables. The initial sample of big-data-related full text articles was N = 79,012 using the time span of 1916–2016, which is a century.

A positivistic ideology has been used with the rest of this section for processing the collected empirical qualitative data. Nonparametric statistics has been used for statistical analysis such as rank and distribution tests. Both linear and nonlinear techniques have also been adopted to evaluate the collected data. [1]

Text analytics was used on resulting full text articles to locate the highest frequencies of words used, and those were merged with the keywords given for the manuscript. Then one best-fitting keyword was designated to represent the dominant topic of every full text article. Various analytics were performed to highlight interesting patterns. Finally, the results were interpreted to answer the research questions and then draw implications.

Longitudinal big data literature analysis

Figure 2 is a columnar chart summarizing the top 10 sources for the collected data sample (1916–2016). It actually includes 11 indexes, along with the other category. The other column represents sources that individually held less than 1% of the sample. The production of big data articles by year was examined. Of the 79,012 articles collected since 1916, it seems that there had been no articles on big data prior to 2000, and most were dated after 2010. [1]

The sample data was then fitted into a Weibull distribution shape by creating an exponential trend equation.

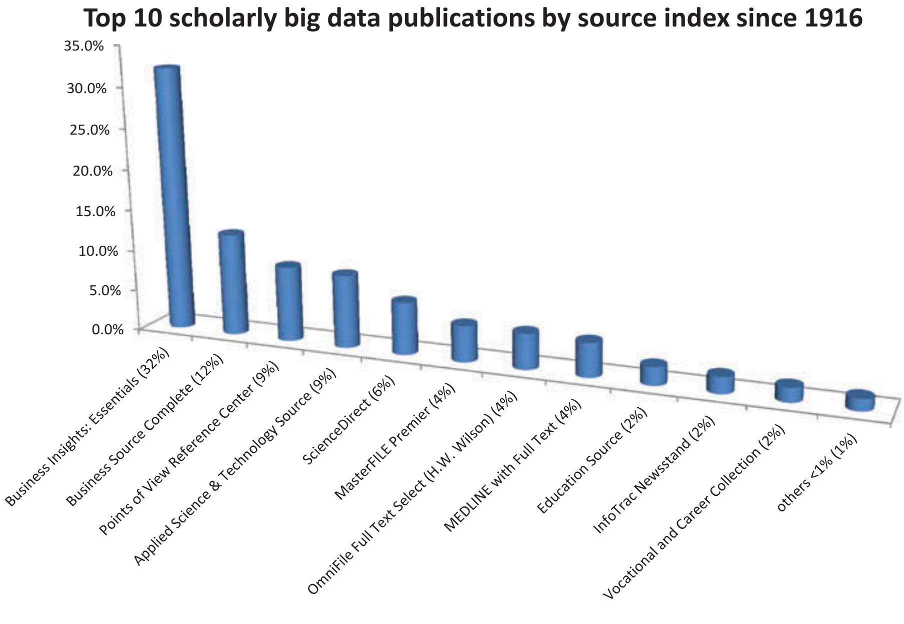

Figure 2. Top 10 big data publications(1916–2016, N = 34 indexes).

LinkManagerBM_REF_rotnKpzY The result was that the Weibull distribution significantly captured 83% of the frequency variation by year ($r^2 = 0.8256$, $p < .01$, N = 79012), as shown in Figure 3. 2016 data was not included because it was not over during the data collection and statistical analysis. Figure 3 clearly shows that the production of big data literature commenced late in 2011 with 1592 new articles, which increased almost fourfold to 7804 by 2012, although the rate of growth slowed during 2015. This means that the scholarly big data research production is accelerating at an exponential rate from 2011 on. [1]

The last column in Figure 3 is the reverse cumulative percent of the frequencies. This tests the Pareto principle to determine whether 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.[1] Now we look at big data production by outlet type, such as journals versus conferences or newspapers. This analysis includes all outlet types and all the data. The results are summarized in Figure 4, a bar chart ordered by the 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. This implies that social media, public media, as well as market media have played an important role in changing big data into a hot topic. It is 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, in particular in ICT fields.[1] This also means that big data has drawn increasing attention from a more widespread community, not only in the ICT community.

Privacy and security in the big data body of knowledge

The previous subsection demonstrated that big data research production started in 2011 and it is rising exponentially, which agrees with the research results of Salleh and Janczewski.[32] Next, the 2011–2016 literature data will be briefly examined to identify what the important big data topics are and what attention is being given to privacy-related research.

First, the top 10 big data keywords between journals and conference proceedings are compared to determine whether certain topics are more likely to be studied in peer meetings at conferences or published in journals. The results are summarized in Figure 5, which shows the 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.[1]

An interesting finding is that there are 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 are conceptual in nature, and only 5% of the journal manuscripts are purely theoretical. Around 95% of the journal articles focused on specific big data topics such as data mining or cloud computing instead of conceptual frameworks, as shown in Figure 5.

Second, the research zoomed into the production of big data studies published in academic journals and found that most of the big data body of knowledge outcomes is in journals (2011–2016, N = 13029), almost double those in conference proceedings (2011–2016, N = 7445). The results

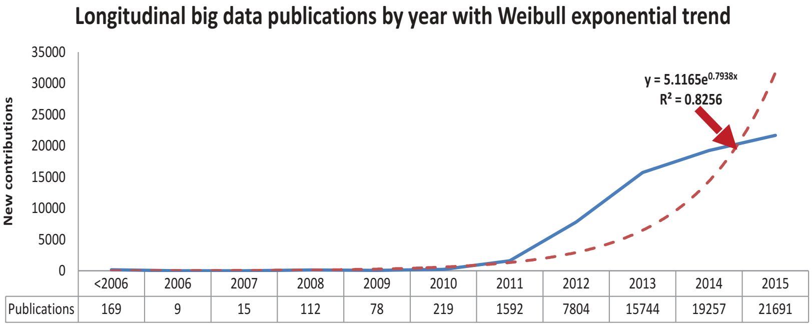

Figure 3. Frequency of new big data studies by year with dashed trend line (1916–2015, N = 79,012).

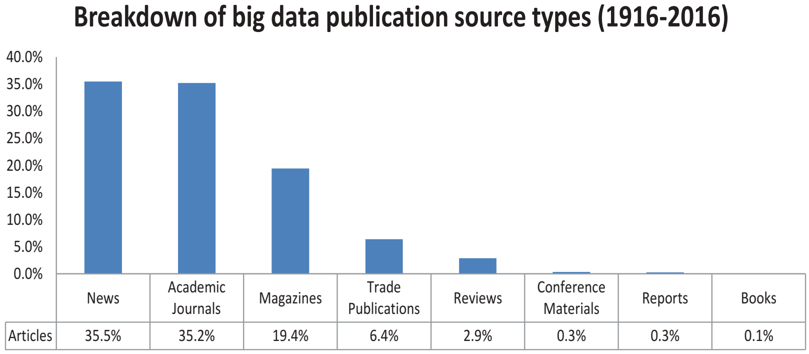

Figure 4. Big data production by outlet type (1916–2016, N = 79012).

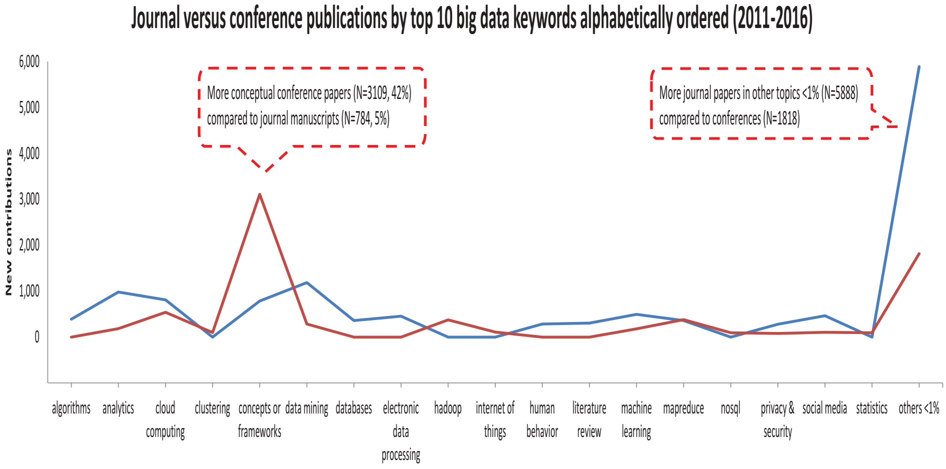
Journal versus conference publications by top 10 big data keywords alphabetically ordered (2011-2016)

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

had also revealed that more unique topics emerged in journals, and there were substantially less wholly theoretical papers in journals.[33] Then, the research downloaded and closely examined the 13,029 manuscript titles, abstracts, and keywords published during 2011–2016 in journals. The title, abstract, and keywords are used to nominate a dominant topic for every article to represent the big data body of knowledge. The finalized big data body of knowledge resulted in 49 topics consisting of one to three words like ‘data mining’, ‘artificial intelligence’, and ‘online social networks’. The keyword topics are edited to ensure that the spellings and meanings of the topics are consistent between the raters, such as changing plural to singular, and use the same underlying dominant topic; for example, ‘database mining’ and ‘mining analysis’ are changed to ‘data mining’. Individual statistical techniques are also recoded, for example, validity, correlation, regression, ANOVA, and so on, into the dominant topic ‘statistics’. The *Kappa* inter-rater statistical technique is selected to ensure that 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).[1] Since this technique requires an ordinal rating, the research rated the perceived ability of the keyword to fit the article on a scale of 1–5, where 5 was the highest. The research then performed an inter-rater agreement on the keywords, resulting in a Kappa index of 0.93 ($p > .05$), which is considered a reliable agreement.[52] This is comparable to a correlation of $r^2 = 0.8649$ or 87%, indicating that the researchers were significantly in agreement about the dominant topic for each article.

Finally, the research factored the journal big data from 2011 to 2016 into a displayable shortlist of 10–15 dominant topics using the frequency, and grouped all remaining low-count topics into a 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, the research does not display it at present (but will provide it upon request). The data labels in Figure 6 represent percentages, and 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. The results from Figure 6 reveal that the topmost of the 49 dominant big data topics published in journals during 2011–2016 is data mining (N = 1186), at 9.1%. The next three topics are data analytics (N = 979, 7.5%), cloud computing (N = 808, 6.2%), and literature reviews (N = 784,

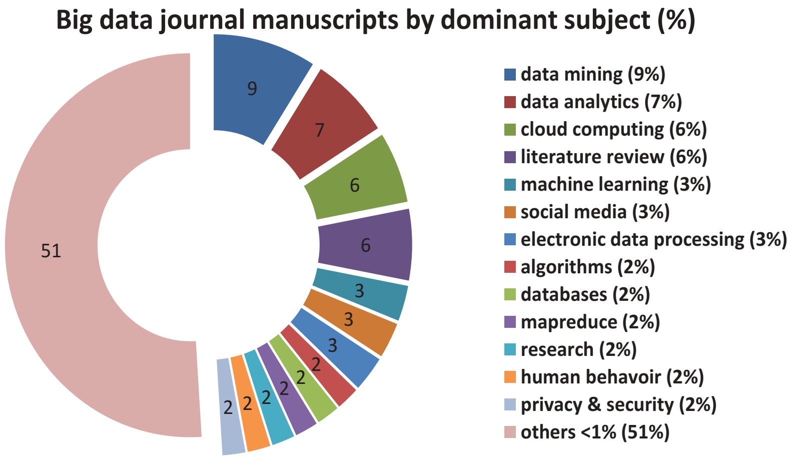
Big data journal manuscripts by dominant subject (%)

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

6.0%). 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%), algorithm (N = 388, 3.0%), database (N = 360, 2.8%), MapReduce (N = 358, 2.7%), research method (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.

The above analysis answers the third research question on what is the status of privacy and security research in the big data paradigm. 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.[1]

(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, and
(13) privacy& security.

The answer to the third research question is that the relative contribution of privacy and security topics within the scholarly big data body of knowledge is 2% and ranked at 13 out of 49 in our sample. This result implies that privacy and security topics are significant for the big data paradigm.

The state-of-art privacy and security in the big data age

The statistical analysis in the previous section explored the interrelationships between privacy and security and other areas such as data mining and data analytics in the big data age. In what follows, this section provides an in-depth analysis of privacy and security to reveal the state-of-the-art privacy and security in the big data age.

Based on the 2012 ACM Computing Classification System[55, 56], privacy and security have been formed into a three-level tree. For example, privacy and security–Cryptography–Key management, where privacy and security is the first-level node, cryptography is the second level node, directly linking the first-level node, and key management is the third node linking the second node, Cryptography. Now privacy and security are illustrated using the two-level structure using Table2 (first column), that is, the second-level topics (eight nodes) of privacy and security consist of Cryptography, Formal methods and theory of security, Security services, Intrusion/anomaly detection and malware mitigation, Security in hardware, Systems security, Network security, Database and storage security, Software and application security, and Human and societal aspects of security and privacy.

Table 2. Privacy and security of big data based on SCOPUS searched data.

Levels 1 & 2 in privacy and security	2012 No.	2013 No.	2014 No.	2015 No.	2016 No.
Cryptography	31	44	81	144	226
Formal methods and theory of security	0	0	2	1	1
Security services	49	74	158	260	372
Intrusion/anomaly detection and malware mitigation	3	3	12	11	11
Security in hardware	10	23	25	42	73
Systems security	103	182	331	583	818
Network security	64	89	209	363	551
Database and storage security	10	7	22	14	33
Software and application security	12	24	37	64	60
Human and societal aspects of security and privacy	0	0	0	1	2
Security and privacy	40	70	147	237	318

We searched the peer-reviewed literature using “article title, abstract, and key words” of SCOPUS (www.scopus.com), with a data range of 2011–2016. Data type: all. The search was executed on September 18, 2017.Our SCOPUS search rules are as follows, for example:

TITLE-ABS-KEY (security AND hardware AND big AND data) AND PUBYEAR > 2010 AND PUBYEAR < 2017.

The search result is shown in Table 2; the second column to the sixth column correspond to the publication number of SCOPUS from 2012 to 2016, respectively.

Table 2 shows that formal methods and theory of security, and human and societal aspects of security and privacy are two weak research areas in big data-driven privacy and security. This implies that these two areas should draw more attention for research and development, because formal methods and theory of security are important foundations for the healthy development of any other areas of big data-driven privacy and security.

Table 2 also shows that intrusion/anomaly detection and malware mitigation, security in hardware, security in hardware, database and storage security, and software and application security are relatively weak research areas. Cryptography, security services, systems security, and network security are four strong research areas in big data-driven privacy and security, because the number of publications in each of them (based on the research rules) has been exponentially increasing during 2012–2016, as shown in Figure 7.

Implications and limitations

This section looks at the theoretical and practical implications of this research for stakeholders.

There are at least two theoretical implications of this research for the stakeholders of privacy and security in the big data paradigm or related research fields.

(1) The proposed Boolean model reveals the interrelationships among privacy, security, and big data. This might help the stakeholders to deepen their understanding of privacy and security at three levels in the big data paradigm: fundamental, interactional, and integrated.
(2) The researchers and developers might conduct research and development of big data-driven privacy

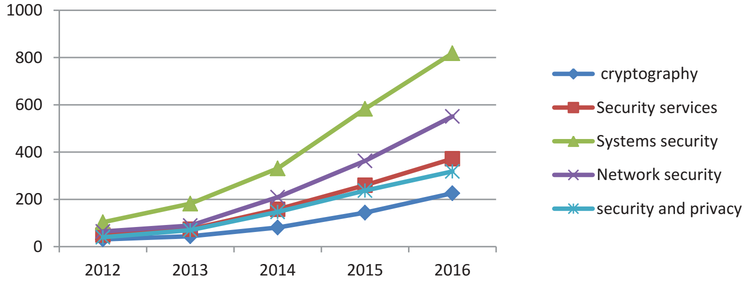

Figure 7. The state-of-the-art privacy and security in the big data age.

and security based on the proposed classification of big data-driven privacy and security. For example, they might delve into big volume-driven privacy and security, big velocity-driven privacy and security, big variety-driven privacy and security, or big veracity-driven privacy and security in the big data age.

This research has at least three practical implications for the stakeholders of privacy and security in the big data paradigm or related research fields.

(1) The statistical analysis of big data-driven privacy and security can facilitate the researcher and practitioners to carry out research, because they understand that big data-driven privacy and security takes about 2% of all the research.
(2) The proposed state-of-the-art privacy and security in the big data age might be of interest to organizations and managers to invest in research and development on "formal methods and theory of security" and "human and societal aspects of security and privacy" because they are two weak research areas in big data-driven privacy and security based on our research.
(2) Computing practitioners might use the proposed Boolean model and classification of big data-driven privacy and security to develop a new information system for securing privacy in the big data age.

There are also at least three limitations of this research.

The proposed Boolean model only reveals the interrelationship among big data privacy and security at three levels: fundamental, interactional, and integrated. However, it has not revealed the deep relationship among them. This is the reason why we use two breakdowns to break the proposed Boolean model to address this limitation. The first breakdown is the classification of big data-driven privacy and security by breaking big data down using its characteristics. The second breakdown is the classification of privacy and security using the 2012 Taxonomy of ACM.

Another limitation of this research is addressing the third identified issue, we collected literature from 1916 to 2016. This seems to be important for the evolution of big data; however, it might not be significant because big data is a recent phenomenon since 2011, which has been verified in Figure 3. To address this limitation, we use literature data (2011–2016) to examine privacy and security in the big data body of knowledge in a subsection.

The third limitation of this research is that it has not examined the impact of big data and privacy on confidentiality, integrity, non-repudiation, availability, and reliability of information, where the latter are well-known in information security.[57] We will discuss it as a future work.

Conclusions

This article examined privacy and security in big data paradigm through proposing a Boolean model for big data privacy and security, and a classification of big data-driven privacy and security. It highlighted important research topics and identified critical gaps through the statistical analysis of big data and its relationships with privacy and security based on literature data published from 1916 to 2016. It analyzed a significant amount of big data literature and found that privacy and security-related topics accounted for about 2% of the total journal outcomes during 2011–2016. It also provided state-of-the-art privacy and security in the big data age by analyzing the document results searched using SCOPUS and the ACM classification for privacy and security in 2012. [22] It demonstrated that the number of publications in privacy and security in the big data age has been exponentially increasing in the past years since 2012. This implies that big data has a significant impact on the research and development of privacy and security.

In the future work, we will investigate big data-driven privacy and security through looking at the impact of big data on each of them in the proposed classification in terms of technology, governance, and policy development. We will also look at business's and consumers' perceptions of privacy and security issues in the healthcare industry in a developing country.[11]

References

1. Strang K, Sun Z. Big Data Paradigm: what is the Status of Privacy and Security?. *Ann Data Sci (Springer)*. 2017;4(1):1–17. doi:10.1007/s40745-016-0096-6.
2. Sun Z, Sun L, Strang K. Big data analytics services for enhancing usiness Intelligence. J Comput Inf Syst. 2016;1–8. doi:10.1080/08874417.2016.1220239,.
3. Hubaux J-P, Juels A. Privacy is dead, long live privacy. Communications of the ACM. 2016;59(6):39–41.
4. Goldfield NI. Big data—hype and promise. J Ambul Care Manage. 2014;37(3):195–96.
5. Kambatla K, Kollias G, Kumar V, Grama A. Trends in big data analytics. J Parallel Distrib Comput. 2014;74(7):2561–73. doi:10.1016/j.jpdc.2014.01.003.
6. Kim GH, Trimi S, Chung JH. Big data applications in the public sector. Commun ACM. 2014;57(3):78–85. doi:10.1145/2500873.

7. Pence HE. What is big data and why is it important?. J Educ Technol Syst. 2015;43(2):159–71. doi:10.2190/ET.2143.2192.d].
8. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. J Big DataOpen Access. 2016;3(1):pp. doi:10.1186/s40537-016-0059-y,.
9. Bohannon J. Credit card study blows holes in anonymity. Science. 2015;347(6221):468. doi:10.1126/science.347.6221.468.
10. Eastin MS, Brinson NH, Doorey A, Wilcox G. Living in a big data world: predicting mobile commerce activity through privacy concerns. Comput Human Behav. 2016;58:214–20. doi:10.1016/j.chb.2015.12.050.
11. Ekbia H. Big data, bigger dilemmas: A critical review. J Assoc Inf Sci Technol. 2015;66(8):et al. doi:10.1002/asi.23294.
12. Gharabaghi K, Anderson-Nathe B. Big Data for Child and Youth Services?. Child Youth Serv. 2014;2014(7):193–95. doi:10.1080/0145935X.2014.955382.
13. Kshetri N. Big data's impact on privacy, security and consumer welfare. Telecomm Policy. 2014;38(11):1134–45. doi:10.1016/j.telpol.2014.10.002.
14. Lichtblau E, Weilandaug N, "Hacker Releases More Democratic Party Files, Renewing Fears of Russian Meddling," *New York Times*, pp. A12-A14, 2016, August 12.
15. BBC. Massive ransomware infection hits computers in 99 countries. 2017 13 5 . [Online]. Available: http://www.bbc.com/news/technology-39901382. [Accessed 17 9 2017].
16. Fu K, Kohno T, Lopresti D. Safety, Security, and Privacy Threats Posed by Accelerating Trends in the Internet of Things. 2017. [Online]. Available: http://cra.org/ccc/wp-content/uploads/sites/2/2017/02/Safety-Security-and-Privacy-Threats-in-IoT.pdf. [Accessed 15 4 2017].
17. B. L. Filkins, J. Y. Kim and B. Roberts. Privacy and security in the era of digital health: what should translational researchers know and do about it?. Am J Transl Res. 2016;8(3):1560–80
18. Hoffman S, Podgurski A. Big Bad Data: law, Public Health, and Biomedical Databases. J Law, Med Ethics. 2013;41:56–60. doi:10.1111/jlme.12040.
19. Thorpe JH, Gray EA. Law and the Public's Health. BIG DATA AND PUBLIC HEALTH: NAVIGATING PRIVACY LAWS TO MAXIMIZE POTENTIAL. Public Health Rep. 2015;130(2):171–75. doi:10.1177/003335491513000211.
20. Booch G. The Human and Ethical Aspects of Big Data. IEEE Software. 2014;31(1):20–22. doi:10.1109/MS.2014.16.
21. Leszczynski A. Spatial big data and anxieties of control. Environ Planning: Soc Space. 2015;33(6):965–84.
22. ACM. the 2012 ACM Computing Classification System. 2012. [Online]. Available: https://www.acm.org/publications/class-2012. [Accessed 18 9 2017].
23. Sun Z, Finnie G. Intelligent Techniques in E-Commerce: A Case-based Reasoning Perspective. Vol. 2004. Heidelberg Berlin: Springer-Verlag;2010.
24. Laudon KG, Laudon KC. Management Information Systems: Managing the Digital Firm. 14th. Harlow, England: Pearson; 2016.
25. Rothstein MA. Ethical Issues in Big Data Health Research: currents in Contemporary Bioethics. J Law, Med Ethics. 2015;43(2):425–29. doi:10.1111/jlme.12258.
26. Soumitra SB, Kim H, Oluwaseyi OI. Privacy and security issues in mobile health: current research and future directions. Health Policy Technol. 2017. doi:10.1016/j.hlpt.2017.01.004,.
27. Solove DJ. Introduction: privacy self-management and the consent dilemma. Harv Law Rev. 2013;126(7):1880–903.
28. Henke N, Bughin J, "McKinsey Global Institute," December 2016. [Online].
29. Sun Z, Strang K, Firmin S. Business Analytics-Based Enterprise Information Systems. J Comput Inf Syst. 2017;57(2):169–78. doi:10.1080/08874417.2016.1183977,.
30. Mayer-Schoenberger V, Cukier K. Big Data: A Revolution that Will Transform How We Live, Work, and Think. Boston, MA: Houghton Mifflin Harcourt Publishing Company;2013.
31. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev. 2012;61–68.
32. Minelli M, Chambers M, Dhiraj A, Data B. Big Analytics: Emerging Business Intelligenceand Analytic Trends for Today's Businesses. (Chinese Edition 2014). New York, NY: Elsevier; 2013.
33. Wang H, Jiang X, Kambourakis G. Special issue on Security, Privacy and Trust in network-based Big Data. Inf Sci (Ny). 2015;318:48–50. doi:10.1016/j.ins.2015.05.040.
34. Ou L, Qin Z, Yin H, Li K. Security and Privacy in Big Data (Chapter 12). Procedia Computer Science, Elsevier; 2016. pp. 285–308.
35. Salleh KA, Janczewski L. Technical, organizational and environmental security and privacy issues of big data: A literature review. Procedia Comput Sci. 2016;100:19–28. doi:10.1016/j.procs.2016.09.119.
36. IBM. The Four V's of Big Data. 2015. [Online]. Available: http://www.ibmbigdatahub.com/infographic/four-vs-bigdata.
37. Sun Z, Wang PP. A Mathematical Foundation of Big Data. J New Math Nat Comput. 2017;13(2):8–24.
38. VajjhalaN, Strang K, Sun Z.. Statistical modeling and visualizing open big data using a terrorism case study. Proceedings of the International Conference on Open and Big Data (OBD 2015); 2015 Aug 24- 26; Rome, Italy: IEEE Press.
39. Davis A. Using Technology to Protect Intimate Images and Help Build a Safe Community. 2017. [Online]. Available: https://newsroom.fb.com/news/2017/04/using-technology-to-protect-intimate-images-and-help-build-a-safe-community/. [Accessed 8 4 2017].
40. Sun Z, Wang P, Strang K. A mathematical theory of big data. *IEEE Transactions on Knowledge and Data Engineering*; 2017, IEEE Computer Society, p. Under Review. 2017.
41. Kumar B. An encyclopedic overview of 'big data' analytics. Int J Appl Eng Res. 2015;10(3):5681–705.
42. Wikipedia. Cybercrime. 5 9 2017. [Online]. Available: https://en.wikipedia.org/wiki/Cybercrime. [Accessed 15 9 2017].
43. Sathi A. Big data analytics: Disruptive technologies for changing the game. Boise, ID, USA: MC Press: IBM Corporation;2013.
44. Chen H, Chiang R, Storey V. Business intelligence and analytics: from big data to big imppact. Mis Q. December 2012;36(4):1165–88.
45. Betser J, Belanger D. Architecting the enterprise with big data analytics. In: Jay Liebowitz editor. Big Data and Business Analytics. Boca Raton, FL: CRC Press; 2013. p. 1–20.
46. Chen CP, Zhang C-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf Sci (Ny)*. 2014;275:314–47. doi:10.1016/j.ins.2014.01.015.
47. Borne K. Top 10 Big Data Challenges – A Serious Look at 10 Big Data V's. April 2014. [Online]. Available: https://www.mapr.com/blog/top-10-big-data-challenges-%E2%80%93-serious-look-10-big-data-v%E2%80%99s.
48. Gartner. Big data. 2016. [Online]. Available: http://www.gartner.com/it-glossary/big-data/.
49. Loshin D. Big Data Analytics: from Strategic Planning to Enterprise Integration woth Tools. Techniques, NoSQL and Graph. Amsterdam: Elsevier; 2013.
50. Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manage. 2015;35:137–44. doi:10.1016/j.ijinfomgt.2014.10.007.
51. Dana M. "On Orbitz, Mac users steered to pricier hotels. Wall Streat J. August 23 2012. pp. 1-3. Available: https://www.wsj.com/articles/SB10001424052702304458604577488822667325882.
52. Duhigg C. The power of habit: Why we do what we do in life and business. New York: Penguin Random House;2014.
53. Zadeh LA. Fuzzy sets. Inf Control. 1965;8(3):338–53. doi:10.1016/S0019-9958(65)90241-X.
54. Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. Hoboken, NJ: Wiley & IEEE Press;2011.
55. Wikipedia. Weibull distribution. 2 9 2017. [Online]. Available: https://en.wikipedia.org/wiki/Weibull_distribution. [Accessed 18 9 2017].
56. Cohen J, Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd. Mahwah, NJ: Lawrence Erlbaum Associates; 2003.
57. Wikipedia. Information security. 2017. [Online]. Available: https://www.wikipedia.com/wiki/Information_security. [Accessed 1 12 2017]