TRANSACTIONS ON DATA PRIVACY 11 (2018) 199–217

Big Data Privacy Context: Literature Effects On Secure Informational Assets

Celina Rebello[a], **Elaine Tavares**[a,b]

[a]Federal University of Rio de Janeiro, Brazil - Coppead Graduate Business School

[a]Rua Pascoal Lemme, 355 - Ilha do Fundao - Cidade Universitaria, Rio de Janeiro - RJ

21941-918 - Brazil. E-mail: [a]rebello.celina@gmail.com,[b]elaine.tavares@coppead.ufjr.br

Received 24 August 2017; received in revised form 18 February 2018; accepted 23 April 2018

Abstract. This article's objective is the identification of research opportunities in the current big data privacy domain, evaluating literature effects on secure informational assets. Until now, no study has analyzed such relation. Its results can foster science, technologies and businesses. To achieve these objectives, a big data privacy Systematic Literature Review (SLR) is performed on the main scientific peer reviewed journals in Scopus database. Bibliometrics and text mining analysis complement the SLR. This study provides support to big data privacy researchers on: most and least researched themes, research novelty, most cited works and authors, themes evolution through time and many others. In addition, TOPSIS and VIKOR ranks were developed to evaluate literature effects versus informational assets indicators. Secure Internet Servers (SIS) was chosen as decision criteria. Results show that big data privacy literature is strongly focused on computational aspects. However, individuals, societies, organizations and governments face a technological change that has just started to be investigated, with growing concerns on law and regulation aspects. TOPSIS and VIKOR Ranks differed in several positions and the only consistent country between literature and SIS adoption is the United States. Countries in the lowest ranking positions represent future research opportunities.

Keywords. Privacy, Big Data, Bibliometrics, MCDM, TOPSIS, VIKOR

1 Introduction

Big Data and privacy studies promote market excitement due to its perceived potential in research, business economy and social activities [26, 41, 35]. Secure Internet Servers (SIS) are key data storage elements in big data's value chain [9, 51]. When individual's privacy enters the equation, frictional and also controversial effects show off like data misuse, user's overexposure, data breaches and many others [21, 36, 68, 6]. One of these effects lays on gaps between current big data privacy theory and its practical indicators on key informational assets adoption like SIS. However, no study has evaluated big data privacy relation through Systematic Literature Review (SLR), bibliometrics, text mining approach and multi-criteria decision making methods (MCDM) rankings.

This study main objective is the identification of research opportunities in the current big data privacy domain, giving a decision support alternative to researchers. First, the SLR provides the basis for bibliometrics mapping. Second, a theme and text mining analysis

is performed over selected documents. Finally, the ranking on paper-country-SIS is performed with Technique for Order Preferences by Similarity to an Ideal Solution (TOPSIS) and the VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR).

Big data privacy social sciences studies are focused on: users' concerns, awareness, self-management, self-disclosure, personality traits, privacy preservation. Variables as gender, age, education, and their relations have been explored but not exhausted [32, 7]. The computational studies focus on encryption algorithms and informational security. Questions like "how did they get my name?" [14] could be investigated with the support of research production versus informational assets indicators providing insights on big data privacy maturity level.

Data is contextual, stored in servers, and desired by many [38, 45, 62]. SIS are one of the data storage and transmission alternatives in the big data value chain. They are gateways to "personally identifiable information", which makes them privacy violations targets just as safes in banks. They also reflects the investment level in Privacy Enhancing Technologies (PETs), used to protect stored data from unauthorized access. World Bank's SIS indicator from 2002 to 2015, in table 2, provides a incipient, but valid, approximation to countries' big data privacy concerns. [54].

Servers' violations have consequences like: impersonation, record misuse, patents theft and many other types of frauds. These effects impact economy and countries' sovereignty [50, 13]. The first privacy computational studies alert for encryptions and control at servers level [72, 1, 77] and on social impacts [47] highlighting the relevance of stronger privacy regulations.

The 21st century computer was defined as "an invisible technology aiming current world's improvement, transparently enhancing human (inter)actions" [75]. People are drown by sensors embedded in almost everywhere: cars, wearables and in-home appliances monitoring technologies. Physical barriers are over [34] and monitoring efforts are transparent [28]. Ubiquitous computing became big data's necessary environmental condition [18, 40, 12].

In this study, section 2 presents the literature review that will provide SLR inputs for the search string. Section 3 describes the research method. Section 4 describes SLR results and TOPSIS and VIKOR ranks, eliciting research opportunities. Finally, section 5 discusses conclusions, limitations and implications.

2 Literature Review

2.1 Privacy Definitions

Privacy is conceived through many lenses. According to James F. Stephen, it is impossible to define privacy clearly. But, its violations are easy to point out. The intrusion of a stranger in someone's liberty be it by coercion, persuasion or even by the law are examples of privacy violations [64]. One of the first formal attempts to defy privacy is seen in Bentham's Inspection House. The Panopticon was designed as a place where an individual was set under custody and fully exposed [5]. Its violations and dilemmas were first analyzed in 1890, when photography was the new technological phenomenon, and its use by journalists was ethically questioned. Privacy was defined as the "right to be let alone"[74].

Other definitions are: "claim of individuals, or groups, or institutions to determine for themselves when, how and to what extent information about them is communicated to others"[76]; "the condition of not having undocumented personal knowledge about one possessed by others";"A person's privacy is diminished exactly to the degree that others

possess this kind of knowledge about him. Documented information consists on information that is found in the public record or is publicly available" [49]; "not simply the absence of information about us in the minds of others, rather is the control we have over information about ourselves" [22], and "the ability of the individual to control the terms under which personal information is acquired and used"[15].

Privacy is also defined in terms of protection from intrusion and information gathering, with individual control as choice, consent and correction [66]. Can also be stated as "a right to control access to places, locations, and personal information along with use and control rights to these goods"[43].Other privacy perspectives are: combination of secrecy, anonymity, and solitude [25], and still physical access, decisional, physiological and informational elements[65, 27]. . Privacy is conceived as a value that, when present at some level, improves society relations in each and every term [76, 60] [31]; as a right that ought to be protected [74]; as a need to ensure liberty and autonomy[52, 53].

Privacy can also relate to culture.This raises some questionings on its importance among all people, on what is inherently private or merely social conventions [69]. Privacy definitions compilation revealed an overlapping among: (1) the right to be let alone; (2) limited access to the self; (3) secrecy; (4) control of personal information;(5) person-hood; and (6) intimacy [60]. Privacy seems to be about everything, and therefore it is a vague concept. Still, a privacy taxonomy is based on informational processing, dissemination and violations[61]. In all cases, privacy faces limited protections by the law [74, 2].

2.2 Big Data

Big data definitions have evolved in time and perspectives, from 3V's (Volume, Velocity and Variety), passing through 4V's (Veracity), 5V's(Variability), 6V's(Value) [24] definitions. Big data analytics operates on statistical methods and semantics extraction processes from both structured and unstructured captured data.

When it comes to privacy, third parties data sharing and accessibility have a growing potential as investigation field [35]. Theories such as Communication Privacy Management(CPM) [10] and Privacy Calculus(PC) [19] are examples of data sharing studies which unveil new research opportunities focused on user self-exposure aspects.

Both analytics and big data's capacities are associated in many definitions: automation, search, aggregation and cross geo-referentiation of massive data volumes [37, 6]. Big data applications intersect economical, strategical, security and consumer welfare domains [8, 35, 39, 79], highlighting ethics as one of the most critical aspects [80, 57, 55, 31, 58, 67].

2.3 Big Data Practices

Privacy threats affect not only law and computer science but social, psychology, economics and media studies [14, 30, 79]. Data breaches generate economical effects that cannot be ignored [63].

Literature have documented data records selling and information misuse practices [15]. These effects are negative among individuals, ranging from value destruction to rights violations, potentially jeopardizing big data's environment. People get surprised when they discovered about their data being used with unexpected purposes [14]. Complaints are mainly related to uninformed, not-consented and out of context data usage, generating data-context distortions and information abuse [46].

2.4 Big Data And Privacy

Big data privacy also seen as a surveillance dilemma [68, 37, 71]. Big data constitutes a technological evolution with exponential scale effects, environmentally, constituted by ubiquitous computing [40, 79] in a need to preserve individuals' privacy. If such balance is not reached, this environment will be jeopardized.

3 Research Method

Privacy literature reviews are non-systematic, many of them cannot be reproduced[78, 59]. A few studies presented a systematic literature reviews related to security and privacy for big data, like [44]. Hopefully, big data research count on Systematic Literature Reviews (SLR)[23, 73, 9]. Just a few studies are focused on classical bibliometrics indicators. However, there have been no big data privacy literature reviews providing research production analysis and practical effects evaluation on computational assets.

This research is based on SLR method [33], added to a literature mapping exploring bibliometrics indicators. This work's objectives are: identify literature gaps ; analyze research themes and its evolution trough time; evaluate research opportunities per country. The first two objectives are covered by the SLR with support of bibliometrics mapping and text mining. To evaluate research opportunities per country, TOPSIS and VIKOR were applied.

This approach provides a precise, concise, technically reproducible and transparent evidence summary around a knowledge domain. The literature mapping elicits themes' evolution through time. Some of the mapped relations are: Most productive authors, countries, most related keywords, most cited authors, current research efforts and starting ones, themes' concentration areas and coupling relations.

3.1 Journal Database, Search String And Data Collection

Scopus was chosen as database because of its availability, broadness and reliability [42]. A peer-reviewed research paper database, such as Scopus, provides a consistent platform to disseminate scientific investigation results, fostering research opportunities and trends.

Chosen query parameters were "Title-Abstract-Keywords"; limited to "Articles" and "Conference papers", written in English. "Privacy" and "Big data" queries intersect directly constraining the result from 2002 to 2016. Research string terms were chosen based on literature selection exposed in section 2. Exclusion criteria were: inaccessible, non-authored, and/or redundant documents.

Search strings were constructed from Privacy and big data definition terms. First,privacy query revealed 83,657 publications (80,256 in English) while "Big Data" returned 27,111 document results (26,076 in English). Search strings such as "Priv*" included private funds and other themes that are out of the research scope. The intersection query retrieved 338 articles. The final string, which consists on the conjunction of privacy and big data search strings, returned 262 documents:

TITLE-ABS-KEY ((privacy AND (encrypt OR crypt* OR auth* OR signature OR steganograph* OR anonymization) AND (protect* OR secre* OR confident* OR "Polic*" OR control OR "self-management" OR preserv* OR hid*))))* ***AND*** *(TITLE-ABS-KEY ((("Big Data" OR "Ubiquitous Computing" OR "Penetrate Computing") AND ("Informed Consent" OR disclos* OR expos* OR shar* OR distribut* OR dissemination OR "Data Exchange" OR "Data Trade")*

))) AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cr")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (SRCTYPE , "p") OR LIMIT-TO (SRCTYPE , "j")).

In this sample, 13 documents had no authors and were excluded from the analysis. The final publication database has 226 publications. Articles with indexes 17, 199 and 249 had DOI errors. Thus, were not retrievable from Scopus. Articles 51 and 52 are redundant and so are 261 and 262. Articles 30, 53, 106 and 107, 120, 139, 157, 198, 216, 233, 241, 255, 250, 254, 256 were not accessible either. Article 147 had only abstract available, which was removed from corpus. Article 253 was a talk in a book chapter, and only one page was retrievable.

3.2 Information Retrieval And Classification Methods

First, the documents went through bibliometrics mapping and analysis with the support of VOSViewer [70]. Second, cluster and content analysis where employed to evaluate the key terms and their quantitative relevance. Hierarchical clusterization employed the *k-means* method. On clusterization algorithms word stemming was included due to semantic aspects and theme's extraction objectives. So, words such as "privacy" versus "private" were also treated as a single keyword and resolved using wildcards, e.g. "priva*".

Text mining and classification were based on article contents inspection through *Tf - IDf* matrix. All algorithms were operationalized in R language [56] with support of"tm" [20] and "bibliometrix" [3] packages. Terms relations, by similarities or distances, revealed concentration areas and relational gaps that were also confirmed by bibliometrics keyword analysis.

3.3 Multi Criteria Decision Making Methods Ranking

TOPSIS and VIKOR methods were chosen to check between practical literature effects on SIS. Both TOPSIS and VIKOR are widely considered as MCDM options[4]. Both methods could be applied to rank alternatives, propose a solution to the research question, having the decision criteria weights under decision-maker's discretion. Data convexity is not mandatory.

TOPSIS purely employs analytical methods based on applying Euclidean distance functions on normalized vectors of positive (outputs) and negative (inputs) criteria[11]. VIKOR determines the compromise ranking list, the trade-off solution, and the weight stability intervals of the obtained compromise solution [48].

TOPSIS and VIKOR focus on ranking alternatives selection in the presence of conflicting criteria. VIKOR provides a maximum "group utility" for the "majority" and a minimum of an individual regret for the "opponent", its ranking index is based on the particular measure of "closeness" to the ideal solution. TOPSIS rank has the "shortest distance" to the ideal solution, which is the best level for all considered attributes, and the "farthest distance" from the "negative-ideal" solution, which is the one with worst attributed values. So, TOPSIS returns two "reference" points, but it does not consider the relative importance of the distances from these points.

TOPSIS and VIKOR proposal analysis are applied to generate rankings presented here as a publication impact alternative measure for big data privacy research and also point countries' consistency between papers' impact and SIS installed capacity. Ranks provide a supported decision mechanism to big data privacy researchers.

The percentage of non-cited-publications (NCP) was assumed as the negative ideal while all other criteria are considered positive. These ranks reveal a new approach on publica-

tion's relevance, differently than classical bibliometrics. Rankings provide big data privacy research indicator.

4 Results and Analysis

As presented on Figure 1, the privacy and big data intersection had its first increase in 2008, with significant growth from 2013 to 2014, experiencing a Annual Percentage Growth Rate (APGR) of 18.614% in all subject areas, having APGR of 16.644% in computer science and 24.573% in non-computer science areas.

In the same period, privacy research had APGR of 21.762% in all subject areas, 23.721% in computer science alone and 17.035% in non-computer science domains. Big data in all fields had 36.403%, 34.670% in computer science and 53.781% in non-computer science areas. These rates reveals that other areas, different from computer science, turned their attention to big data. Big data privacy have taken other subject areas attention since 2005.

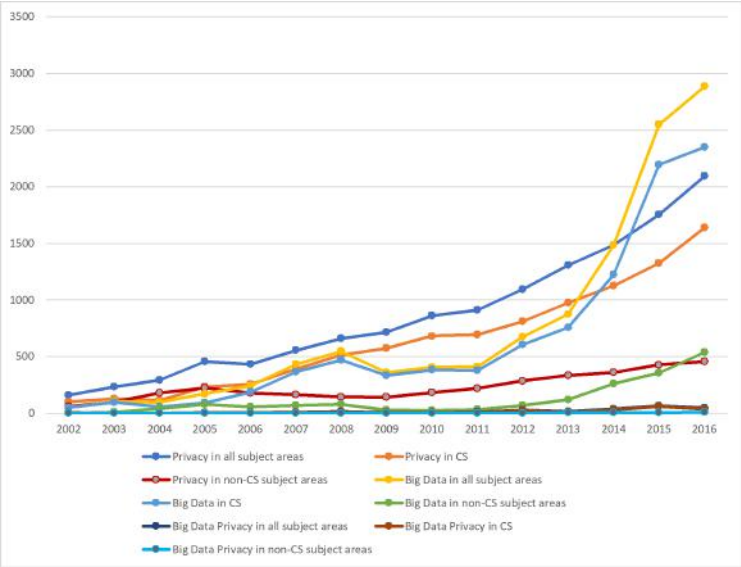

Figure 1: Big Data Privacy Themes And Research Evolution Per Year

Most productive authors are Chinese, associated to American institutions. Liu with 7 articles followed by Chen, Ma and Zhang with 5 articles. Most cited authors are Agrawal and Srinkant with 7 citations, and Weiser with 5 citations. Author's production shows Liu, and Zhang among most productive and cited authors, as seen in Figure 2.

Figure 3 shows Keyword Co-occurrence graph. Its edges reveal that"data mining" cannot reach "access control" directly. All minimum paths connecting these nodes have privacy or security related keywords, i.e. access control" can only be achieved if security as privacy aspects are considered. These relations indicate a new research challenge when non-digital aspects, such as "shoulder surfing", and other off-line information gathering techniques are present.

Most related keywords (Table 1), returned Hadoop and MapReduce, not present in SLR search string. This indicates a potential relation between storage and file access technologies to privacy and to big data's Volume dimension.

Figure 2: Bibliometrics Research Production Indicators

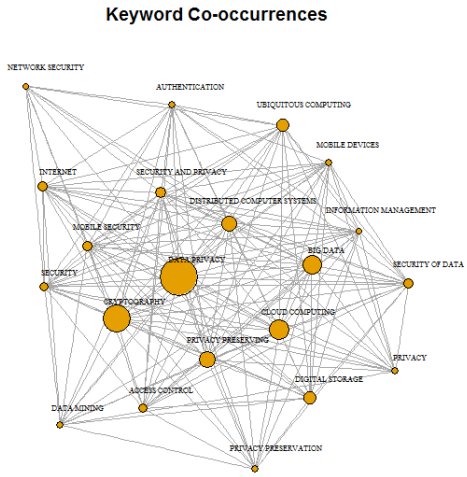

Figure 3: Top 20 Keywords Co-Ocurrences

Table 1: Most Related Keywords

	Author Keywords (DE)	Articles	Keywords-Plus (ID)	Articles
1	PRIVACY	61	DATA PRIVACY	92
2	BIG DATA	41	CRYPTOGRAPHY	61
3	CLOUD COMPUTING	27	BIG DATA	46
4	SECURITY	25	CLOUD COMPUTING	45
5	AUTHENTICATION	17	PRIVACY PRESERVING	41
6	ACCESS CONTROL	13	DISTRIBUTED COMPUTER SYSTEMS	32
7	UBIQUITOUS COMPUTING	13	DIGITAL STORAGE	29
8	PERVASIVE COMPUTING	12	UBIQUITOUS COMPUTING	23
9	ANONYMIZATION	10	INTERNET	22
10	DATA MINING	8	MOBILE SECURITY	22
11	PRIVACY-PRESERVING	8	ACCESS CONTROL	21
12	ANONYMITY	7	SECURITY OF DATA	21
13	CLOUD	7	SECURITY	20
14	ENCRYPTION	7	DATA MINING	18
15	HADOOP	7	DATA HANDLING	17
16	PRIVACY PRESERVATION	7	PRIVACY	17
17	CONFIDENTIALITY	6	SECURITY AND PRIVACY	17
18	DATA ANONYMIZATION	6	SENSITIVE INFORMATIONS	16
19	HOMOMORPHIC ENCRYPTION	6	PRIVACY PRESERVATION	15
20	MAPREDUCE	6	AUTHENTICATION	14

Figure 4 shows that initial research was focused on "access control policy", "schema","Proxy re-encryption". Access control policy defines which users or groups have permissions to access information. The proxy-re-encryption is a encryption process where a third-parties alter the previous encrypted cyphertext. These cryptosystems depends on "schemes" and are relevant to protect user keys. All of these terms are related to SIS.

Research production evolved to "pervasive computing", "authentication", "privacy Protection", "google", "context aware resource management". Subjects like "authentication" relates to "servers" and also to "access control", since the former depends on the later. As a research theme deployment, "Privacy", "anonymity", "access control" "homomorphic encryption", "biometrics" became more relevant.

Keywords like "secure cloud computing", "incremental conceptual cluster" appeared as emerging research trends. Terms such as"law and regulation" are also under investigation and reasons are mainly because of ubiquitous computing effects. Other topics like "shoulder surfing" also called the attention because it is not related to big data itself, but as a information gathering off-line practice.

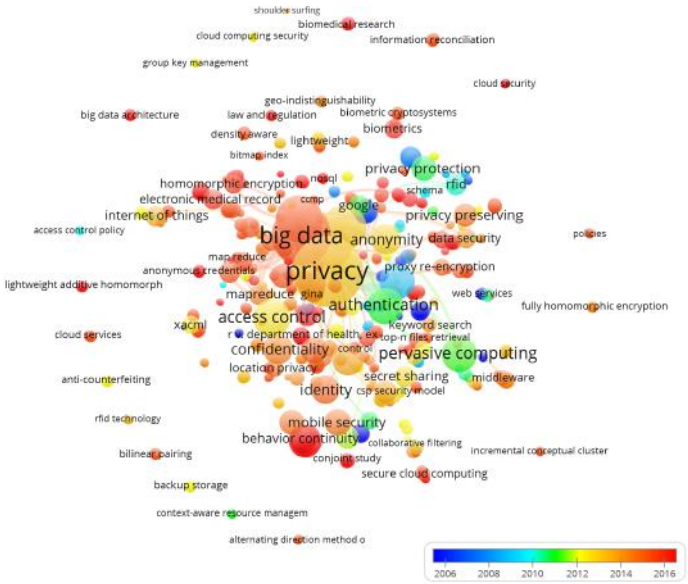

Figure 4: All Keywords Co-Ocurrences by Average Association Strength per Year

Keyword co-occurrence by association strength Figure 5 revealed research focus evolution from sensors, passing through "access control" and finally reaching big data and privacy aspects. Such evolution is in conformity with what is perceived in [40] when he states that privacy is never sufficient when computers are everywhere.

Keyword co-occurrence by association strength also showed the themes' evolution from 2006 to 2016. The first stage research revealed concentrated efforts on: "sensors", "wireless networks", "wireless sensor networks", "context aware", "semantics", "computer privacy" and "ubiquitous computing" as the most relevant among all of them.

Research evolved from "ubiquitous computing" to "access control" and "access control schemes" followed by "scalability","location" and "data storage systems". The third research stage gathers "data privacy", "sensitive information", "cryptography", "anonymization", reaching "big data", "cloud computing", "data handling" and "data mining".

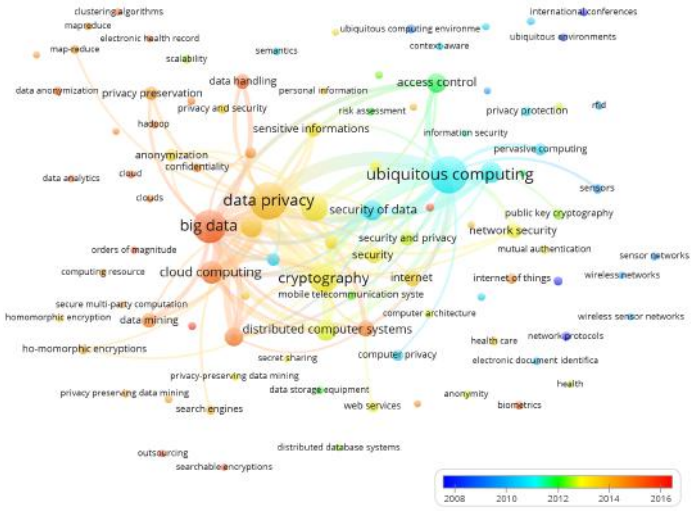

Figure 5: Top 100 Keywords Co-Ocurrences by Association Strength

Research themes started by sensors and networks, than evolved to scalability, storage and access issues. All these aspects are ubiquitous computing pillars. Later, works have focused on cryptography, privacy and security. Finally reaching "big data", "data mining", distributed and cloud computing caught researchers' attention. Privacy preservation aspects highlights the current research.

Figure6, revealed 8 clusters, with minimum of one document per country and minimum one citation per document, with average normalized citations method. Canada, Germany and Saudi Arabia are leading countries in this metric. India, China and United States are leaders in research production. India leads in author's with most recent publications.

Figure 7 shows that bibliometrics coupling per documents, represented as (Author,Year), has a network of 100 references listed only once where 71 nodes and 10 clusters network. Canny's "Collaborative Filtering with Privacy" and Al-Muhtadi's "Routing through the mist: privacy preserving communication in ubiquitous computing environments" as most referenced articles in privacy intersection with big data research domains. These documents reinforce the SIS's role as a critical element in big data privacy research.

Al-Muhtadi's work alerts to ubiquitous computing surveillance potential and proposes a *"mist"* between routers. Canny's work defines a server-based collaborative filtering systems to protect people from monopolies. In this model, users control all of their log data. Users can compute a public "aggregate' of all of their data without exposing individual users data. This model is based on homomorphic encryption with verification schemes distributed to all users. This is one of the first works to be proposed for untrusted servers. Both works propose privacy preservation through anonymization. Fabian's work on multi-cloud storage and sharing architecture is a natural evolution from both. This work focus on medical record anonymization shared among an cloud server array.

Figure 8 shows three clusters, all related to storage, encryption and information security. In the first one, in red, literature mentions hadoop, mapreduce, privacy preserving and anonymization. The green cluster relates to privacy preserving, privacy enhancement, anonymization, information classification. The blue one represents theme's convergence relation. Anonymization, privacy preservation are challenges in computer security.

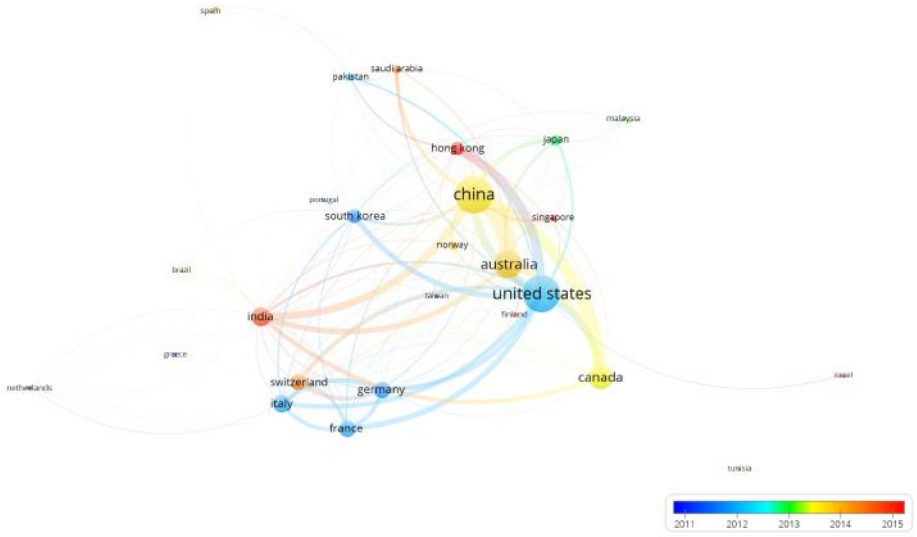

Figure 6: Bibliometrics Coupling per Country association method - average normalized citations

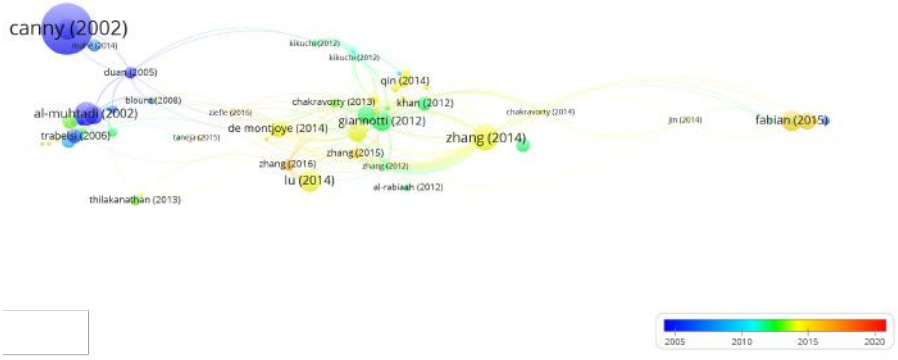

Figure 7: Bibliometrics coupling per documents

Figure 8: Co-Citations - with a minimum of 3 co-citations

4.1 TOPSIS and VIKOR literature effects evaluation

Table 2 shows several country research indicators. Countries research production, citations per paper and non-cited papers measure how relevant papers are and their impact. It presents research production frame against its practical effects and investments such as SIS per country. These differences reveal research opportunities on publication relevance, currently measured by citations. When these results are inserted on SIS adoption, there is an approximation between literature and its practical effects.

On citations per country the United States leads, followed by China, India and South Korea. United States also leads in average Citations per paper (CPP) indicator, followed by Canada and Germany. On SIS adoption rank results are different, with Switzerland leading with 3102 SIS, and China, India and Pakistan on the last positions.

North American research is the most diversified and cited among all countries. It covers privacy awareness and preserving, surveillance and its economical effects privacy metadata protection, network privacy architectures. Other topics are e-government policies, privacy usability challenges, self-disclosure, health, anonymization, geo-privacy, trust building and sensor networks.

China has the second place in publications and in TOPSIS evaluation, and third in citations, with only 10 SIS. However, China comes in second on publications. Internet in China is strongly regulated [29], indicating that government controls its SIS. This condition reflects some potential difficulties on big data privacy research. Chinese publications relate to privacy preserving, trust building, authentication protocols, anonymity, encryption scalability and efficiency.

South Korea is forth in publications; second in citations, VIKOR and third in TOPSIS with 2320 SIS. Both Countries offer good research opportunities. South Korea's high SIS indicator would be explained by the companies' contribution in GDP, such as Samsung. Korean research production is one of the most diversified among all countries in the ranking. It varies from authentication and encryption schemes to information policies and e-government. User behavior aspects are rarely investigated.

India is third in publications, forth in TOPSIS, last in VIKOR. India has the 19th position in SIS indicator. It is interesting to notice that many of Indian researchers are associated to American institutions, instead of Indian universities. This contributes to United Stated leading position. Big data privacy literature practical effects in India's are not as expressive as in other countries. Its research is concentrated on privacy preserving, anonymization algorithms, cloud computing, Internet of Things, wireless networks, health and trust building. User privacy awareness has not been investigated yet.

Australian research is focused on privacy preserving, cloud, green and ubiquitous computing. Surveillance, trust building and big data sharing integrate the research production. This literature has close relation to the Canadian and Asian on privacy preservation.

Canada's production is concentrated on privacy preserving through encryption algorithms and anonymization, access control and identity hiding schemes. Italian publications are on trust building in pervasive computing, anonymous mining, privacy preserving. Privacy law and regulations have not been investigated so far.

United Kingdom's research focus varies from intrusion detection, Radio Frequency Identification secure based protocols, privacy systems for context aware and ubiquitous computing, wearables, transparency, ethics and health care. Research has interesting aspects because it focus from ethics to digital attacks countermeasures like intrusion detection systems. User's behavior and regulations were not explored in their production yet.

French research is diversified and related to trust building, cryptography, biometrics,

cloud and ubiquitous computing. Literature is marked by transparency on social mining and sharing. The trust building proposed by these works are in conformity with privacy by design fundamentals. Trust related papers are the most cited among the French research production.

German studies include wearables, willingness to share, big data privacy health care management, social mining, trust in ubiquitous computing. These studies are mostly focused on building and ensuring trust. Their fundamentals are aligned to privacy by design principles [16].

Japanese papers focus on privacy preserving aspects and challenges. These studies may integrate a new framework that addresses changing business needs and fresh concerns over breaches of personal data. Asian are marked by a strong privacy preservation bias. Regulation development and privacy breaches effects are research opportunities in the country as well as in the Asian region.

Singapore's research production is on privacy preserving. Singapore enacted the Personal Data Protection Act (PDPA) in 2012, and in 2014 it became fully operational. PDPA's effects on corporate governance and data protection practices are also research opportunities in the country and region.

Malaysian publications are diversified from soft computing, privacy concerns on health records and intrusion detection. Malaysia invests in high end engineering research development. Privacy and data protection by design principles are also an opportunity with potential positive effects on country's strategy [17].

Saudi Arabia presents works focused essentially on cloud computing frameworks and ubiquitous computing. Papers application domains are health and multimedia, with utility driven to policy making. These studies take into consideration current frameworks security flaws on their analysis. Ubiquitous computing plays a key role on Saudi Arabia diversification strategy: from oil to big data. Thus, further big data privacy research is a need.

Spanish papers vary from anonymization to PETs, passing through surveillance and authorization. Interesting to notice that user privacy awareness was absent as theme. This is also an aspect to be investigated as research opportunity.

Norwegian works are related to anonymization and privacy preserving. These works present as fundamentals some of the principles present in [16] like transparency and control. Further developments can be derived from these works, contributing to privacy by design philosophy through the proposed frameworks.

Pakistan's research production has a strong focus on protocols, frameworks and signature schemes to build trust. This country has no privacy laws nor policies. This affects country's social and economical development. This is also a research opportunity.

Switzerland appears 18th in publications, 4th in citations per paper, 8th in TOPSIS and 9th in VIKOR, and leads with 3102 SIS. It is interesting to notice that: Switzerland is considered a financial center, it has a high SIS indicator, and also a high Citation per Paper indicator. Swiss citations per paper is the forth highest in the rankings. A new research opportunity is on measuring if and how financial organizations influence affects this relation.

Brazilian research production focus on cloud computing security and privacy framework development. These efforts complement the PETs works and also protocols. Brazilian research did not focus on big data social aspects yet. Big data privacy effects on human behavior may become a prolific opportunity, like anti-fraud detection and also government transparency to fight corruption.

Netherlands is in 20th place regarding number of publications, but 5th in TOPSIS and 7th in VIKOR. Difference between TOPSIS and VIKOR proposed rankings can be due to nor-

malization method applied in these methods. Netherlands' high relation between Citations per Papers and SIS indicator is a research opportunity on how this relation is established. Netherlands' research production is mainly focused on information security. There are still several applications to be covered, specially in ubiquitous computing and trust building.

There is a discrepancy between number of publications and SIS on countries like Brazil, Spain, Malaysia, Saudi Arabia and United Kingdom. These countries have a small research production, less than 10 publications in the analyzed period, and represent new big data privacy local relations to explore. The small publication number indicates research venues, unexplored local opportunities. Big data privacy questions, specially in law and regulation are still concealed. Furthermore, both analysis help in identifying the publication efficiency and effects on SIS implementations and data breaches.

Table 2: Bibliometric Indicators Ranking Compared To TOPSIS And VIKOR

	Country	Pub	Cites	CPP	Std.Dev	NCP	Max.Cites	Pub.SIS	SIS	T.s	T.r	V.S	V.R	V.Q	V.r
1	United States	67	857	12.791	29.834	0.388	207	0.041	1650	0.667	1	0.157	0.090	0.000	1
2	China	46	197	4.283	10.114	0.500	54	4.600	10	0.428	2	0.722	0.142	0.872	12
3	India	33	33	0.971	2.634	0.618	15	4.714	7	0.341	4	0.909	0.143	1.000	20
4	South Korea	29	200	6.897	24.823	0.448	134	0.012	2320	0.419	3	0.482	0.111	0.412	2
5	Australia	16	102	6.375	13.266	0.375	54	0.011	1460	0.249	9	0.668	0.127	0.691	5
6	Canada	10	126	12.600	18.422	0.400	54	0.008	1309	0.298	7	0.590	0.127	0.641	3
7	Italy	10	81	8.100	11.120	0.300	31	0.035	289	0.203	13	0.727	0.131	0.764	6
8	United Kingdom	9	31	3.444	8.487	0.444	26	0.007	1383	0.165	16	0.794	0.139	0.887	13
9	France	8	57	7.125	11.180	0.250	31	0.010	813	0.207	12	0.712	0.135	0.791	8
10	Germany	8	93	11.625	14.481	0.125	36	0.005	1757	0.298	6	0.559	0.132	0.662	4
11	Japan	8	11	1.375	0.518	0.000	2	0.008	971	0.201	14	0.796	0.143	0.925	17
12	Singapore	6	41	6.833	14.825	0.500	37	0.006	932	0.196	15	0.753	0.137	0.844	10
13	Malaysia	5	29	5.800	11.323	0.400	26	0.048	104	0.155	18	0.809	0.139	0.900	14
14	Saudi Arabia	5	27	5.400	4.219	0.200	11	0.093	54	0.157	17	0.816	0.140	0.915	15
15	Spain	5	8	1.600	2.510	0.400	6	0.014	362	0.084	20	0.909	0.143	1.000	19
16	Norway	4	14	3.500	1.291	0.000	5	0.002	2033	0.238	11	0.724	0.142	0.867	11
17	Pakistan	4	22	5.500	1.915	0.000	8	2.000	2	0.244	10	0.787	0.143	0.919	16
18	Switzerland	4	35	8.750	14.953	0.500	31	0.001	3102	0.289	8	0.639	0.141	0.799	9
19	Brazil	3	13	4.333	6.658	0.333	12	0.039	77	0.124	19	0.852	0.143	0.962	18
20	Netherlands	3	25	8.333	10.214	0.000	20	0.001	2828	0.301	5	0.575	0.143	0.778	7

Pub: Number of Publications; Cites: Number of Citations
CPP: Citations per paper; Std.Dev: Citations Standard Deviation
Max.Cites: Maximum Citations
NCP: percentage of Non-Cited Papers
T.s:TOPSIS Score; T.r:TOPSIS Ranking
V.s:VIKOR Score; V.r:VIKOR Ranking
Pub.Sis: Publications/SIS
SIS: Secure Internet Servers in 2015
Source:http://data.worldbank.org/indicator/IT.NET.SECR.P6?view=chart
retrieved in 13/12/2016.

5 Conclusion

This study consists on a RSL, bibliometrics mapping and text mining analysis on big data privacy research evaluation. TOPSIS and VIKOR MCDM Methods were employed to evaluate research practical effects, identifying new research rankings and opportunities.

Privacy in big data is richly represented in the computer science domain, but non-computer science areas have started to investigate it. This study identified "access control", anonymization, authentication and PETs as recent concentration areas and also "ubiquitous computing" as a necessary environmental condition to big data. Non-computer science studies are concentrated on privacy preservation, trust building and privacy self-management. Computer Science studies are focused on encryption, anonymization, storage, cloud computing and data mining.

The TOPSIS and VIKOR Rankings revealed that United States leads on research impact and on the applying literature practical effects, which are represented by SIS. SIS ranking per country was the chosen criteria because it is an worldwide accepted computational asset indicator available from the World Bank. Another reason was SIS technical essence, which is secure data storage and transmission.

Rankings revealed that countries like Brazil, Spain may represent new opportunities according to both rankings. Saudi Arabia and United Kingdom, India, Japan and Pakistan according to VIKOR ranks. It is interesting to notice that Asia and Europe have research bias, driven to ethical aspects and privacy preserving, while United States drives efforts towards encryption, storage, and technical frameworks. Arabian countries investigate themes related to their economical growth. Latin countries like Brazil have just started to research big data privacy. Countries with a incipient research production is prolific in investigation opportunities because too little is known out their reality and matters.

Results may vary if inclusion and exclusion criteria are changed. Ranking may also change according to chosen MCDM method, criteria and weights adopted by decision makers. Since there was no previous study relating research production and MCDM methods, this work adds a contribution on a structured process where researchers should focus their efforts. SIS can provide non-exhaustive, but still relevant, measure of privacy concern per country. It is massively present in computer science research production and represents a key factor in data protection and application services.

The whole process had to be documented, including intermediary results to avoid inconsistency. Data retrieval depended on Scopus' search engine technical structure. Article's classification by publishers is a biased process and another recognized limitation. Documents exclusively available in other bases such as Web of Science and DBLP are excluded from the sample. Data extraction processed was limited to available articles and pdf conversion readability. Since each publisher has its own text template, data cleaning and text mining processes had increased in complexity. Text mining was performed on English-only article corpus. Such limitations can be surpassed with the addition of other languages' dictionaries, improving semantic broadness.

Future studies should target on big data privacy cultural aspects. User behavior, laws and regulations, and visual privacy are interesting topics that appeared on this analysis. Studies related to data breaches and practice versus theory evaluation on privacy governance would also be an interesting field to explore. Too little is known about privacy law and regulation causes and effects on people, organizations and government. These studies should be evaluated on their "intention to inform and evaluate" big data privacy practical effects. Would be desirable that these studies describe the big data privacy implications versus measurable protection practices, their benefits to policymakers, planners, researchers and

citizens.

References

[1] AM M, WS EJ (1970) Communications privacy. Telecommunications 4(10):34, 36.

[2] Ambrose ML, Ausloos J (2013) The Right to Be Forgotten Across the Pond. Journal of Information Policy Penn State University Press:1–23.

[3] Aria M., Cuccurullo C., (2016) bibliometrix: A r tool for comprehensive bibliometric analysis of scientific literature. http://wwwbibliometrixorg

[4] Behzadian M, Otaghsara SK, Yazdani M, Ignatius J (2012) A state-of the-art survey of topsis applications. Expert Systems with Applications 39(17):13,051–13,069

[5] Bentham J (2009) Panopticon or the inspection house (1791). Works, iv (Edinburgh: William Tait, and London: Simpkin Marshall, 1843)

[6] Boyd D, Crawford K (2012) Critical questions for big data. Information, Communication & Society 15(5):662–679

[7] Bozdag E (2013) Bias in algorithmic filtering and personalization. Ethics Inf Technol 15(3):209–227

[8] Buhl PDHU, Rglinger DM, Moser DKF, Heidemann DJ (2013) Big Data. Bus Inf Syst Eng 5(2):65–69

[9] Chen H, Chiang RH, Storey VC (2012) Business intelligence and analytics: From big data to big impact. MIS quarterly 36(4):1165–1188

[10] Child JT, Pearson JC, Petronio S (2009) Blogging, communication, and privacy management: Development of the blogging privacy management measure. Journal of the Association for Information Science and Technology 60(10):2079–2094

[11] Ching-Lai Hwang KYa (1981) Multiple Attribute Decision Making: Methods and Applications A State-of-the-Art Survey, 1st edn. Lecture Notes in Economics and Mathematical Systems 186, Springer-Verlag Berlin Heidelberg

[12] Crampton J (2015) Collect it all: national security, Big Data and governance. GeoJournal 80(4):519–531

[13] Cukier K, Mayer-Schoenberger V (2013) Rise of big data: How it's changing the way we think about the world, the. Foreign Aff 92:28

[14] Culnan MJ (1993) How did they get my name: An exploratory investigation of consumer attitudes toward secondary information use. MIS Quarterly 17

[15] Culnan PK Mary J; Armstrong (1999) Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. Organization Science 10

[16] D'Acquisto G, Domingo-Ferrer J, Kikiras P, Torra V, de Montjoye YA, Bourka A (2015) Privacy by design in big data: an overview of privacy enhancing technologies in the era of big data analytics. arXiv preprint arXiv:151206000

[17] Danezis G, Domingo-Ferrer J, Hansen M, Hoepman JH, Metayer DL, Tirtea R, Schiffner S (2015) Privacy and data protection by design-from policy to engineering. arXiv preprint arXiv:150103726

[18] Diebold FX (2012) A personal perspective on the origin (s) and development of'big data': The phenomenon, the term, and the discipline, second version

[19] Dinev T, Hart P (2006) An extended privacy calculus model for e-commerce transactions. Information Systems Research 17(1):61–80

[20] Feinerer I, Hornik K (2016) tm: Text Mining Package.https://CRAN.R-project.org/package=tm, r package version 0̇7-1

[21] Floridi L (2005) The Ontological Interpretation of Informational Privacy. Ethics Inf Technol 7(4):185–200

[22] Fried C (1984) Privacy: A moral analysis. philosophical dimensions of privacy: An anthology. schoeman, fd editor

[23] Frizzo-Barker J, Chow-White PA, Mozafari M, Ha D (2016) An empirical study of the rise of big data in business scholarship. International Journal of Information Management 36(3):403–413

[24] Gandomi A, Haider M (2015) Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35(2):137–144

[25] Gavison R (1980) Privacy and the limits of law. The Yale Law Journal 89(3):421–471

[26] Gotterbarn D (1999) Privacy lost: The Net, autonomous agents, and virtual information. Ethics and Information Technology 1(2):147–154

[27] Grodzinsky F, Tavani HT (2010) Applying the contextual integrity model of privacy to personal blogs in the blogoshere

[28] Hartzog W, Stutzman FD (2013) The case for online obscurity. California Law Review 101(1):1

[29] Harwit E, Clark D (2001) Shaping the internet in china. evolution of political control over network infrastructure and content. Asian Survey 41(3):377–408

[30] Heffetz O, Ligett K (2014) Privacy and Data-Based Research. The Journal of Economic Perspectives 28(2):75–98

[31] Hull G (2015) Successful failure: what Foucault can teach us about privacy self-management in a world of Facebook and big data. Ethics Inf Technol 17(2):89–101

[32] Hull G, Lipford HR, Latulipe C (2010) Contextual gaps: privacy issues on Facebook. Ethics Inf Technol 13(4):289–302

[33] Kitchenham B (2004) Procedures for performing systematic reviews. Keele, UK, Keele University 33(2004):1–26

[34] Könings B, Schaub F (2011) Territorial privacy in ubiquitous computing. In: Wireless On-Demand Network Systems and Services (WONS), 2011 Eighth International Conference on, IEEE, pp 104–108

[35] Kshetri N (2014) The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns. Big Data & Society 1(2)

[36] Langheinrich M (2009) A survey of rfid privacy approaches. Personal and Ubiquitous Computing 13(6):413–421

[37] Lyon D (2014) Surveillance, snowden, and big data: Capacities, consequences, critique. Big Data & Society 1(2):2053951714541,861

[38] Margulis ST (1977) Conceptions of privacy: Current status and next steps. Journal of Social Issues 33(3):5–21

[39] Martin KE (2015) Ethical issues in the big data industry. MIS Quarterly Executive 14:2

[40] Matzner T (2014) Why privacy is not enough privacy in the context of ubiquitous computing and big data. Journal of Information, Communication and Ethics in Society 12(2):93–106

[41] McNeely CL, Hahm Jo (2014) The big (data) bang: policy, prospects, and challenges. Review of Policy Research 31(4):304–310

[42] Mongeon A Philippe; Paul-Hus (2015) The journal coverage of web of science and scopus: a comparative analysis. Scientometrics

[43] Moore A (2008) Defining privacy. Journal of Social Philosophy 39(3):411–428

[44] Nelson B, Olovsson T (2016) Security and privacy for big data: A systematic literature review. In: Big Data (Big Data), 2016 IEEE International Conference on, IEEE, pp 3693–3702

[45] Nissenbaum H (2004) Privacy as contextual integrity. Wash L Rev 79:119

[46] Nissenbaum H (2011) A contextual approach to privacy online. Daedalus 140(4):32–48

[47] OE D (1970) Social impact of computers 36:449–451, cited By 0; Conference of AFIPS Conf Proc, Spring Jt Comput Conf, NJ ; Conference Date: 5 May 1970 Through 7 May 1970

[48] Opricovic S (1998) Multicriteria optimization of civil engineering systems. Faculty of Civil Engineering, Belgrade 2(1):5–21

[49] Parent WA (1983) Privacy, morality, and the law. Philosophy & Public Affairs pp 269–288

[50] Parsons C (1979) Computers and the international flow of information. Computer Networks (1976) 3(3):171–173, cited By 0

[51] van de Pas J, van Bussel GJ (2015) 'Privacy Lost - and Found?' The information value chain as a model to meet citizens' concerns. Electronic Journal of Information Systems Evaluation 18(2):185–195

[52] Posner RA (1977) Right of privacy, the. Ga L Rev 12:393

[53] Posner RA (1978) Privacy, secrecy, and reputation. Buff L Rev 28:1

[54] Posner RA (1981) The economics of privacy. The American economic review 71(2):405–409

[55] Qiu JL (2015) Reflections on big data: just because it is accessible does not make it ethical. Media Culture and Society 37:1089–1094

[56] R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/

[57] Schroeder R (2014) Big Data and the brave new world of social media research. Big Data & Society 1(2)

[58] Selinger E, Hartzog W (2015) Facebooks emotional contagion study and the ethical problem of co-opted identity in mediated environments where users lack control. Research Ethics p 1747016115579531

[59] Smith HJ, Dinev T, Xu H (2011) Information privacy research: An interdisciplinary review. MIS Quarterly 35(4):989–1015

[60] Solove DJ (2002) Conceptualizing privacy. California Law Review 90 pp 1087–1155

[61] Solove DJ (2006) A taxonomy of privacy. University of Pennsylvania Law Review 154

[62] Soria-Comas J, Domingo-Ferrer J (2015) Big Data Privacy: Challenges to Privacy Principles and Models. Data Sci Eng pp 1–8

[63] Spiekermann S, Acquisti A, Bhme R, Hui KL (2015) The challenges of personal data markets and privacy. Electronic Markets 25(2):161–167

[64] Stephen JF (1991) Liberty, Equality, Fraternity: And Three Brief Essays. University of Chicago Press

[65] Tavani HT (2008) Informational privacy: Concepts, theories, and controversies. The handbook of information and computer ethics pp 131–164

[66] Tavani HT, Moor JH (2001) Privacy protection, control of information, and privacy-enhancing technologies. SIGCAS Comput Soc 31(1):6–11

[67] Taylor L (2016) No place to hide? The ethics and analytics of tracking mobility using mobile phone data. Environment and Planning D: Society and Space 34(2):319–336

[68] Tene O, Polonetsky J (2012) Big data for all: Privacy and user control in the age of analytics. Nw J Tech & Intell Prop 11:xxvii

[69] Torra V (2017) Data Privacy: Foundations, New Developments and the Big Data Challenge. Springer

[70] Van Eck NJ, Waltman L (2009) Software survey: Vosviewer, a computer program for bibliometric mapping. Scientometrics 84(2):523–538

[71] Van Dijck J (2014) Datafication, dataism and dataveillance: Big data between scientific paradigm

and ideology. Surveillance & Society 12(2):197

[72] Van Tassel D (1969) Advanced cryptographic techniques for computer. Communications of the ACM 12,

[73] Wamba SF, Akter S, Edwards A, Chopin G, Gnanzou D (2015) How big datacan make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics 165:234–246

[74] Warren SD, Brandeis LD (1890) The Right to Privacy. Harvard Law Review 4(5):193–220

[75] Weiser M, Gold R, Brown JS (1999) The origins of ubiquitous computing research at parc in the late 1980s. IBM systems journal 38(4):693

[76] Westin A (1970) Privacy and freedom. Atheneum, New York

[77] Young D (1970) A computerized face identity-verification system (fais). Proceedings of SPIE - The International Society for Optical Engineering 22:119–126

[78] Yu S (2016) Big privacy: Challenges and opportunities of privacy study in the age of big data. IEEE Access 4:2751–2763

[79] Zuboff S (2015) Big other: Surveillance capitalism and the prospects of an information civilization. Journal of Information Technology 30(1):75–89

[80] Zwitter A (2014) Big Data ethics. Big Data & Society 1(2)