Security and Privacy for Big Data: A Systematic Literature Review

Citation for the original published paper (version of record):
Nelson, B., Olovsson, T. (2016). Security and Privacy for Big Data: A Systematic Literature Review. 2016 IEEE International Conference on Big Data (Big Data): 3693-3702.
http://dx.doi.org/10.1109/BigData.2016.7841037

Security and Privacy for Big Data: A Systematic Literature Review

Boel Nelson
Department of Computer Science and Engineering
Chalmers University of Technology
Email: boeln@chalmers.se

Tomas Olovsson
Department of Computer Science and Engineering
Chalmers University of Technology
Email: tomasol@chalmers.se

***Abstract*—Big data is currently a hot research topic, with four million hits on Google scholar in October 2016. One reason for the popularity of big data research is the knowledge that can be extracted from analyzing these large data sets. However, data can contain sensitive information, and data must therefore be sufficiently protected as it is stored and processed. Furthermore, it might also be required to provide meaningful, proven, privacy guarantees if the data can be linked to individuals.**

To the best of our knowledge, there exists no systematic overview of the overlap between big data and the area of security and privacy. Consequently, this review aims to explore security and privacy research within big data, by outlining and providing structure to what research currently exists. Moreover, we investigate which papers connect security and privacy with big data, and which categories these papers cover. Ultimately, is security and privacy research for big data different from the rest of the research within the security and privacy domain?

To answer these questions, we perform a *systematic literature review* (SLR), where we collect recent papers from top conferences, and categorize them in order to provide an overview of the security and privacy topics present within the context of big data. Within each category we also present a qualitative analysis of papers representative for that specific area. Furthermore, we explore and visualize the relationship between the categories. Thus, the objective of this review is to provide a snapshot of the current state of security and privacy research for big data, and to discover where further research is required.

I. Introduction

Big data processing presents new opportunities due to its analytic powers. Business areas that can benefit from analyzing big data include the automotive industry, the energy distribution industry, health care and retail. Examples from these areas include analyzing driving patterns to discover anomalies in driving behaviour [1], making use of smart grid data to create energy load forecasts [2], analyzing search engine queries to detect influenza epidemics [3] and utilizing customers' purchase history to generate recommendations [4]. However, all of these examples include data linked to individuals, which makes the underlying data potentially sensitive.

Furthermore, while big data provides analytic support, big data in itself is difficult to store, manage and process efficiently due to the inherent characteristics of big data [5]. These characteristics were originally divided into three dimensions referred to as the three Vs [6], but are today often divided into four or even five Vs [2, 5, 7]. The original three Vs are *volume*, *variety* and *velocity*, and the newer V's are *veracity* and *value*. *Volume* refers to the amount of data, which Kaisler et al. [5] define to be in the range of 10^{18} bytes to be considered big data. *Variety* denotes the problem of big data being able to consist of different formats of data, such as text, numbers, videos and images. *Velocity* represents the speed at which the data grows, that is, at what speed new data is generated. Furthermore, *veracity* concerns the accuracy and trustworthiness of data. Lastly, *value* corresponds to the usefulness of data, indicating that some data points, or a combination of points, may be more valuable than others. Due to the potential large scale data processing of big data, there exists a need for efficient, scalable solutions, that also take security and privacy into consideration.

To the best of our knowledge, there exists no peer-reviewed articles that systematically review big data papers with a security and privacy perspective. Hence, we aim to fill that gap by conducting a *systematic literature review* (SLR) of recent big data papers with a security and privacy focus. While this review does not cover the entire, vast, landscape of security and privacy for big data, it provides an insight into the field, by presenting a snapshot of what problems and solutions exists within the area.

In this paper, we select papers from top security and privacy conferences, as well as top conferences on data format and machine learning for further analysis. The papers are recent publications, published between 2012 and 2015, which we manually categorize to provide an overview of security and privacy papers in a big data context. The categories are chosen to be relevant for big data, security or privacy respectively. Furthermore, we investigate and visualize what categories relate to each other in each reviewed paper, to show what connections exists and which ones are still unexplored. We also visualize the proportion of papers belonging to each category, and the proportion of papers published in each conference. Lastly we analyze and present a representative subset of papers from each of the categories.

The paper is organized as follows. First, the method for gathering and reviewing papers is explained in Section II. Then, the quantitative and qualitative results are presented in Section III, where each of the categories and their corresponding papers are further analyzed in the subsection with their corresponding name. A discussion of the findings and directions for future work is presented in Section IV. Lastly,

a conclusion follows in Section V.

II. Methodology

In this paper, we perform a *systematic literature review* (SLR) to document what security and privacy research exists within the big data area, and identify possible areas where further research is needed. The purpose of this review is to categorize and analyze, both in a quantitative and a qualitative way, big data papers related to security or privacy. Therefore, in accordance with SLR, we define the following research questions the review should answer:

- What recent security or privacy papers exists in the big data context?
- How many papers cover security or privacy for big data?
- Which security, privacy and big data topics are represented in the area?
- When a paper covers more than one category, which categories intertwine?

SLRs originate from medical research, but has been adapted for computer science, and in particular software engineering, by Kitchenham [8] in 2004. More specifically, a SLR is useful for summarising empirical evidence concerning an existing technology as well as for identifying gaps in current research [8]. We answer our research questions by performing the steps in the review protocol we have constructed, in accordance with Kitchenham's guidelines, displayed in Table I.

1. **Data sources and search strategy:** Collect papers
2. **Study selection/study quality assessment:** Filter papers
3. **Data extraction:** Categorize papers, extract the novelty of the papers' scientific contribution
4. **Data synthesis:** Visualize papers and highlight the contributions

TABLE I: Review protocol

As the data source, we have used papers from top conferences, ranked A^* by the Computing Research and Education Association of Australasia (CORE)[i] in 2014. In total, twelve relevant conferences have been chosen, including all three of CORE's top ranked security and privacy conferences. There also exists several new, promising conferences in big data. However, none of these big data specific conferences are ranked yet, and thus they are not included in this review. Arguably, the highest quality papers should appear in the A^* ranked conferences, instead of in a not proven venue. Furthermore, it is our belief that new ideas hit conferences before journals, and thus journals have been excluded from the review. Consequently, we have chosen top conferences for closely related topics: machine learning and data format[ii]. Thus, the big data conferences are represented by seven conferences from the field of data format and two from machine learning. The chosen conferences are presented in Table II, and we further discuss the consequences of choosing these conferences in Section IV.

[i] http://portal.core.edu.au/conf-ranks/

[ii] Field of research code 0804: http://www.abs.gov.au/Ausstats/abs@.nsf/0/206700786B8EA3EDCA257418000473E3?opendocument

[iii] As labeled by CORE

Acronym	Conference Name	Field(s) of Research[iii]
DCC	Data Compression Conference	Data Format
ICDE	International Conference on Data Engineering	Data Format
ICDM	IEEE International Conference on Data Mining	Data Format
SIGKDD	Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining	Data Format
SIGMOD	Association for Computing Machinery's Special Interest Group on Management of Data	Data Format
VLDB	International Conference on Very Large Databases	Data Format
WSDM	ACM International Conference on Web Search and Data Mining	Data Format, Distributed Computing, Library and Information Studies
ICML	International Conference on Machine Learning	Artificial Intelligence and Image Processing
NIPS	Neural Information Processing System Conference	Artificial Intelligence and Image Processing
CCS	ACM Conference on Computer and Communications Security	Computer Software
S&P	IEEE Symposium on Security and Privacy	Computation Theory and Mathematics, Computer Software
USENIX Security	Usenix Security Symposium	Computer Software

TABLE II: Conferences the papers were collected from, including acronym and field of research

Step 1: To perform the first step from Table I, the collection of papers, we have constructed the following two queries:

- **Query A:** allintitle: privacy OR private OR security OR secure
  Sources: DCC, ICDE, ICDM, SIGKDD, SIDMOD, VLDB, WSDM, ICML and NIPS
  Timespan: 2012-2015
- **Query B:** allintitle: "big data"
  Sources: DCC, ICDE, ICDM, SIGKDD, SIDMOD, VLDB, WSDM, ICML, NIPS, CCS, S&P and USENIX Security
  Timespan: 2012-2015

Note that only the title of a paper is used to match on a keyword. The reason for this is to reduce the amount of false positives. For example, if the search is not limited to the title, a paper might discuss the keyword in the introduction or as related work, but it might not otherwise be included in the paper. Since the review is performed manually, it

would require a labor intensive analysis just to eliminate those irrelevant papers. Furthermore, we believe that the papers related to security or privacy would mention this in their title. Thus, we have focused on a smaller, relevant, subset.

Query A focuses on finding papers related to security or privacy in one of the big data conferences. This query is intentionally constructed to catch a wide range of security and privacy papers, including relevant papers that have omitted 'big data' from the title. Furthermore, query B is designed to find big data papers in any of the conferences, unlike query A. The reason to also include query B is foremost to capture big data papers in security and privacy conferences. Query B will also be able to find big data papers in the other conferences, which provides the opportunity to catch security or privacy papers that were not already captured by query A.

Step 2: After the papers have been collected, we manually filter them to perform both a selection and a quality assessment, in accordance with the guidelines for a SLR. First, we filter away talks, tutorials, panel discussions and papers only containing abstracts from the collected papers. We also verify that no papers are duplicates to ensure that the data is not skewed. Then, as a quality assessment we analyze the papers' full corpora to determine if they belong to security or privacy. Papers that do not discuss security or privacy are excluded. Thus, the irrelevant papers, mainly captured by query B, and other potential false positives, are eliminated.

To further assess the quality of the papers, we investigate each papers' relevance for big data. To determine if it is a big data paper we include the entire corpus of the paper, and look for evidence of scalability in the proposed solution by examining if the paper relates to the five V's. The full list of included and excluded papers is omitted in this paper due to space restrictions, but it is available from the authors upon request.

Step 3: Then, each paper is categorized into one or more of the categories shown in Table III. These categories were chosen based on the five V's, with additional security and privacy categories added to the set. Thus the categories capture both the inherent characteristics of big data, as well as security and privacy.

Category	V	Security or Privacy
Confidentiality[iv]		✓
Data Analysis	Value	
Data Format	Variety, Volume	
Data Integrity	Veracity	✓
Privacy[v]		✓
Stream Processing	Velocity, Volume	
Visualization	Value, Volume	

TABLE III: Categories used in the review, chosen based on the five V's. A checkmark in the third column means that the category is a security or privacy category.

In total, 208 papers match the search criteria when we run both queries in Google Scholar. After filtering away papers and performing the quality assessment, 82 papers remain. Query A results in 78 papers, and query B contributes with four unique papers that were not already found by query A. In Table IV the number of papers from each conference is shown for query A and query B respectively.

Conference Acronym	Query A		Query B	
	Number of Papers	Percentage of Papers	Number of Papers	Percentage of Papers
DCC	0	0%	0	0%
ICDE	22	28%	0	0%
ICDM	4	5%	0	0%
SIGKDD	0	0%	0	0%
SIGMOD	21	26%	1	25%
VLDB	25	31%	1	25%
WSDM	0	0%	0	0%
ICML	5	6.3%	0	0%
NIPS	1	1.3%	0	0%
S&P	-	-	1	25%
USENIX Security	-	-	0	0%
CCS	-	-	1	25%
Total:	78	100%	4	100%

TABLE IV: The number, and percentage, of papers picked from each conference, for query A and query B

Step 4: Then, as part of the data synthesis which is the last step in the review protocol in Table I, the quantitative results from the queries are visualized. Both as circle packing diagrams, where the proportion of papers and conferences is visualized, and as a circular network diagram where relationships between categories are visualized. Thereafter a qualitative analysis is performed on the papers, where the novel idea and the specific topics covered are extracted from the papers' corpora. A representative set of the papers are then presented.

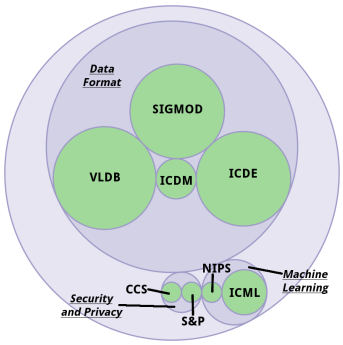
III. Results

In this section, we quantitatively and qualitatively analyze the 82 papers. Figure 1 (a) visualizes where each paper originates from, using circle packing diagrams. The size of each circle corresponds to the proportion of papers picked from a conference. As can be seen, most papers have been published in ICDE, SIGMOD or VLDB. Furthermore, the distribution of the different categories is illustrated in Figure 1 (b), where the size of a circle represents the amount of papers covering that category. Prominent categories are *privacy*, *data analysis* and *confidentiality*.

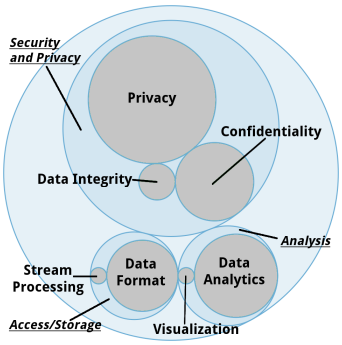
Furthermore, some papers discuss more than one category and therefore belong to more than one category. Therefore, the total number of papers when all categories are summed will exceed 82. To illustrate this overlap of categories, the relationship between the categories is visualized as a circular network diagram in Figure 2. Each line between two categories means that there exists at least one paper that discusses both categories. The thickness of the line reflects the amount of papers that contain the two categories connected by the line. *Privacy* and *data analytics* as well as *confidentiality* and *data format* are popular combinations. *Stream processing* and

[iv] As defined by ISO 27000:2016 [9]

[v] *Anonymization* as defined by ISO 29100:2011 [10]



(a) Conferences, grouped by research field



(b) Categories, grouped by similarity

Fig. 1: Circle packing diagrams, showing the proportion of papers belonging to conferences (a) and categories (b)

visualization are only connected by one paper, respectively, to *privacy*.

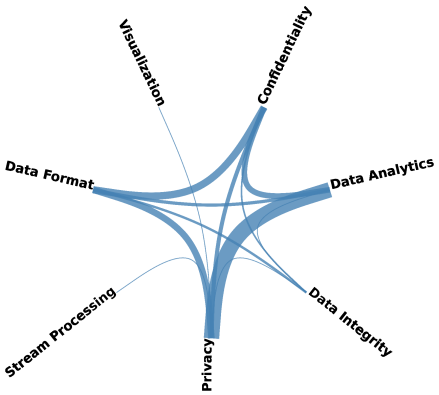


Fig. 2: Connections between categories, where the thickness of the link represents the amount of papers that connect the two categories

Since there is not enough room to describe each paper in the qualitative analysis, we have chosen a representative set for each category. This representative set is chosen to give an overview of the papers for each category. Each selected paper is then presented in a table to show which categories it belongs to. An overview of the rest of the papers are shown in Table **??**.

A. *Confidentiality*

Confidentiality is a key attribute to guarantee when sensitive data is handled, especially since being able to store and process data while guaranteeing confidentiality could be an incentive to get permission to gather data. In total, 23 papers were categorized as confidentiality papers. Most papers used different types of encryption, but there was no specific topic that had a majority of papers. Instead, the papers were spread across a few different topics. In Table V, an overview of all papers presented in this section is given.

Five papers use *homomorphic encryption*, which is a technique that allows certain arithmetic operations to be performed on encrypted data. Of those five papers, one uses fully homomorphic encryption which supports any arithmetic operation, whereas the rest use partial homomorphic encryption which supports given arithmetic operations. Liu et al. [11] propose a secure method for comparing trajectories, for example to compare different routes using GPS data, by using partial homomorphic encryption. Furthermore, Chu et al. [12] use fully homomorphic encryption to provide a protocol for similarity ranking.

Another topic covered by several papers is *access control*. In total, four papers discuss access control. For example, Bender et al. [13] proposed a security model where policies must be explainable. By explainable in this setting Bender et al. refers to the fact that every time a query is denied due to missing privileges, an explanation as to what additional privileges are needed is returned. This security model is an attempt to make it easier to implement the principle of least privilege, rather than giving users too generous privileges. Additionally, Meacham and Shasha [14] propose an application that provides access control in a database, where all records are encrypted if the user does not have the appropriate privileges. Even though the solutions by Bender et al. and Meacham and Shasha use SQL, traditionally not associated with big data, their main ideas are still applicable since it only requires changing the database to a RDBMS for big data that have been proposed earlier, such as Vertica [15] or Zhu et al.'s [16] distributed query engine.

Other topics covered were *secure multiparty computation*, a concept where multiple entities perform a computation while keeping each entity's input confidential, *oblivious transfer*, where a sender may or may not transfer a piece of information to the receiver without knowing which piece is sent, as well as different *encrypted indexes* used for improving search time efficiency. In total, three papers use secure multiparty computation, two use oblivious transfer and two use encrypted indexes.

B. *Data Integrity*

Data integrity is the validity and quality of data. It is therefore strongly connected to veracity, one of the five V's. In total, five papers covered data integrity. Since there is only a small set of data integrity papers, no apparent topic trend was spotted. Nonetheless, one paper shows an *attack* on integrity,

Author	Short Title	C	DA	DF	DI	P	SP	V
Akcora et al.	Privacy in Social Networks					✓		
Allard et al.	Chiaroscuro	✓	✓			✓		
Bonomi and Xiong	Mining Frequent Patterns with Differential Privacy		✓			✓		
Bonomi et al.	LinkIT					✓		
Cao et al.	A hybrid private record linkage scheme	✓				✓		
Chen and Zhou	Recursive Mechanism			✓		✓		
Dev	Privacy Preserving Social Graphs for High Precision Community Detection		✓			✓		
Dong et al.	When Private Set Intersection Meets Big Data	✓						
Fan et al.	FAST					✓		
Gaboardi et al.	Dual Query					✓		
Guarnieri and Basin	Optimal Security-aware Query Processing	✓		✓				
Guerraoui et al.	D2P					✓		
Haney et al.	Design of Policy-aware Differentially Private Algorithms					✓		
He et al.	Blowfish Privacy					✓		
He et al.	DPT		✓			✓		
He et al.	SDB	✓		✓				
Hu et al.	Authenticating Location-based Services Without Compromising Location Privacy	✓				✓		
Hu et al.	Private search on key-value stores with hierarchical indexes	✓						
Hu et al.	VERDICT	✓		✓				
Jain and Thakurta	(Near) Dimension Independent Risk Bounds for Differentially Private Learning		✓			✓		
Jorgensen and Cormode	Conservative or liberal?					✓		
Kellaris and Papadopoulos	Practical differential privacy via grouping and smoothing					✓		
Khayyat et al.	BigDansing			✓	✓			
Kozak and Zezula	Efficiency and Security in Similarity Cloud Services	✓				✓		
Li and Miklau	An Adaptive Mechanism for Accurate Query Answering Under Differential Privacy					✓		
Li et al.	A Data- and Workload-aware Algorithm for Range Queries Under Differential Privacy					✓		
Li et al.	DPSynthesizer					✓		
Li et al.	Fast Range Query Processing with Strong Privacy Protection for Cloud Computing					✓		
Li et al.	PrivBasis		✓			✓		
Lin and Kifer	Information Preservation in Statistical Privacy and Bayesian Estimation of Unattributed Histograms					✓		
Lu et al.	Generating private synthetic databases for untrusted system evaluation			✓		✓		
Mohan et al.	GUPT		✓			✓		
Nock et al.	Rademacher observations, private data, and boosting		✓			✓		
Oktay et al.	SEMROD	✓		✓				
Pattuk et al.	Privacy-aware dynamic feature selection		✓			✓		
Potluru et al.	CometCloudCare (C3)		✓	✓		✓		
Qardaji et al.	Differentially private grids for geospatial data					✓		
Qardaji et al.	PriView					✓		
Qardaji et al.	Understanding Hierarchical Methods for Differentially Private Histograms					✓		
Rahman et al.	Privacy Implications of Database Ranking					✓		
Rana et al.	Differentially Private Random Forest with High Utility		✓			✓		
Ryu et al.	Curso					✓		
Sen et al.	Bootstrapping Privacy Compliance in Big Data Systems	✓						
Shen and Jin	Privacy-Preserving Personalized Recommendation					✓		
Terrovitis et al.	Privacy Preservation by Disassociation					✓		
To et al.	A Framework for Protecting Worker Location Privacy in Spatial Crowdsourcing					✓		
Wong et al.	Secure Query Processing with Data Interoperability in a Cloud Database Environment	✓		✓				
Xiao et al.	DPCube			✓		✓		
Xu et al.	Differentially private frequent sequence mining via sampling-based candidate pruning		✓			✓		
Xue et al.	Destination prediction by sub-trajectory synthesis and privacy protection against such prediction		✓			✓		
Yang et al.	Bayesian Differential Privacy on Correlated Data					✓		
Yaroslavtsev et al.	Accurate and efficient private release of datacubes and contingency tables					✓		
Yi et al.	Practical k nearest neighbor queries with location privacy	✓	✓					
Yuan et al.	Low-rank Mechanism					✓		
Zeng et al.	On Differentially Private Frequent Itemset Mining		✓			✓		
Zhang et al.	Functional Mechanism		✓			✓		
Zhang et al.	Lightweight privacy-preserving peer-to-peer data integration	✓						
Zhang et al.	Private Release of Graph Statistics Using Ladder Functions					✓		
Zhang et al.	PrivBayes					✓		
Zhang et al.	PrivGene		✓			✓		

Author	C	DA	DF	DI	P	SP	V
Bender et al. [13]	✓						
Chu et al. [12]	✓	✓	✓				
Liu et al. [11]	✓	✓					
Meacham and Shasha [14]	✓	✓					

TABLE V: A set of confidentiality papers, showing categories covered by each paper. A checkmark indicates the paper on that row contains the category.

two papers are on *error correction and data cleansing* and two papers use *tamper-proof hardware* to guarantee integrity of the data. An overview of all papers covered in this section are shown in Table VI.

Xiao et al. [17] shows that it is enough to poison 5% of the training values, a data set used solely to train a machine learning algorithm, in order for feature selection to fail. Feature selection is the step where relevant attributes are being decided, and it is therefore an important step since the rest of the algorithm will depend on these features. Thus, Xiao et al. show that feature selection is not secure unless the integrity of the data can be verified.

Furthermore, Arasu et al. [18] implemented a SQL database called Cipherbase that focuses on confidentiality of data as well as integrity in the cloud. To maintain the integrity of the cryptographic keys, they use FPGA based custom hardware to provide tamper-proof storage. Lallali et al. [19] also used tamper-resistant hardware where they enforce confidentiality for queries performed in personal clouds. The tamper-resistant hardware is in the form of a secure token which prevents any data disclosure during the execution of a query. While the secure tokens ensures a closed execution environment, they posses limited processing power due to the hardware constraints which adds to the technical challenge.

Author	C	DA	DF	DI	P	SP	V
Arasu et al. [18]	✓		✓	✓			
Lallali et al. [19]	✓			✓	✓		
Xiao et al. [17]		✓		✓			

TABLE VI: A set of data integrity papers, showing categories covered by each paper

C. Privacy

An important notion is privacy for big data, since it can potentially contain sensitive data about individuals. To mitigate the privacy problem, data can be de-identified by removing attributes that would identify an individual. This is an approach that works, if done correctly, both when data is managed and when released. However, under certain conditions it is still possible to re-identify individuals even when some attributes have been removed [20, 21, 22]. Lu et al. [7] also point out that the risk of re-identification can increase with big data, as more external data from other sources than the set at hand can be used to cross-reference and infer additional information about individuals.

Several privacy models, such as k-anonymity [23], l-diversity [24], t-closeness [25] and differential privacy [26], can be used to anonymize data. The first three are techniques for releasing entire sets of data through privacy-preserving data publishing (PPDP), whereas differential privacy is used for privacy-preserving data mining (PPDM). Thus, differential privacy is obtained without processing the entire data set, unlike the others. Therefore, anonymizing larger data sets can be difficult from an efficiency perspective. However, larger sets have greater potential to hide individual data points within the set [27].

Out of a total of 61 privacy papers, one paper [28] uses *k-anonymity*, and another paper [29] uses *l-diversity* and *t-closeness* but also *differential privacy* to anonymize data. Furthermore, Cao and Karras [30] introduce a successor to *t*-closeness, called *β-likeness* which they claim is more informative and comprehensible. In comparison, a large portion, 46 papers, of the privacy oriented papers focuses only on differential privacy as their privacy model. Most of them propose methods for releasing differentially private data structures. Among these are differentially private histograms [31] and different data structures for differentially private multidimensional data [32].

An interesting observation by Hu et al. [33] is that differential privacy can have a large impact on accuracy of the result. When Hu et al. enforced differential privacy on their telecommunications platform, they got between 15% to 30% accuracy loss. In fact, guaranteeing differential privacy while maintaining high utility of the data is not trivial. From the reviewed papers, 15 of them investigated utility in combination with differential privacy.

One example of a paper that investigates the utility of differentially private results, and how to improve it is Proserpio et al. [34]. The work of Proserpio et al. is a continuation of the differentially private querying language PINQ [35], which they enhance by decreasing the importance of challenging entries, which induce high noise, in order to improve accuracy of the results.

The papers reviewed in this section can be seen in Table VII.

Author	C	DA	DF	DI	P	SP	V
Acs et al.[31]					✓		
Cao and Karras [30]					✓		
Cormode et al.[32]					✓		
Hu et al. [33]			✓		✓		
Jurczyk et al. [29]			✓		✓		
Proserpio et al. [34]			✓		✓		
Wang and Zheng [28]					✓		

TABLE VII: A set of privacy papers, showing categories covered by each paper

D. Data Analysis

Data analysis is the act of extracting knowledge from data. It includes both general algorithms for knowledge discovery, and machine learning. Out of 26 papers categorized as data analysis papers, 15 use *machine learning*. Apart from machine learning, other topics included *frequent sequence mining*, where reoccurring patterns are detected, and different versions of the *k-nearest neighbor (kNN) algorithm*, that finds the k

closest points given a point of reference. All papers from this section are shown in Table VIII.

Jain and Thakurta [36] implemented differentially private learning using kernels. The problem investigated by Jain and Thakurta is keeping the features, which are different attributes of an entity, of a learning set private while still providing useful information.

Furthermore, Elmehdwi et al. [37] implemented a secure kNN algorithm, based on partial homomorphic encryption. Here, Elmehdwi et al. propose a method for performing kNN in the cloud, where both the query and the database are encrypted. Similarly, Yao et al. [38] investigated the secure nearest neighbour (SNN) problem which asks a third party to find the point closest to a given point, without revealing any of the points to the third party. They show attacks for existing methods for SNN, and design a new SNN method that withstand the attacks.

Author	C	DA	DF	DI	P	SP	V
Elmehdwi et al. [37]	✓	✓	✓				
Jain and Thakurta [36]		✓			✓		
Yao et al. [38]	✓	✓					

TABLE VIII: A set of data analysis papers, showing categories covered by each paper

E. Visualization

Visualization of big data provides a quick overview of the data points. It is an important technique, especially while exploring a new data set. However, it is not trivial to implement for big data. Gordov and Gubarev [39] point out visual noise, large image perception, information loss, high performance requirements and high rate of image change as the main challenges when visualizing big data.

One paper, by To et al. [40], shown in Table IX, was categorized as a visualization paper. To et al. implemented a toolbox for visualizing and assigning tasks based on an individuals' location. In this toolbox, location privacy is provided while at the same time allowing for allocation strategies of tasks to be analyzed. Thus, it presents a privacy-preserving way of analyzing how parameters in a system should be tuned to result in a satisfactory trade-off between privacy and accuracy.

Author	C	DA	DF	DI	P	SP	V
To et al. [40]		✓			✓		✓

TABLE IX: All visualization papers, showing categories covered by each paper

F. Stream Processing

Stream processing is an alternative to the traditional store-then-process approach, which can allow processing of data in real-time. The main idea is to perform analysis on data as it is being gathered, to directly address the issue of data velocity. Processing streamed data also allows an analyst to only save the results from the analysis, thus requiring less storage capacity in comparison with saving the entire data set. Furthermore, stream processing can also completely remove the bottleneck of first writing data to disk and then reading it back in order to process it if it is carried out in real-time.

One paper, by Kellaris et al. [41] shown in Table X, combines stream processing with a privacy, and provides a differentially private way of querying streamed data. Their approach enforces *w* event-level based privacy rather than user-level privacy, which makes each event in the stream private, rather than the user that continuously produces events. Event-level based privacy, originally introduced by Dwork et al. [42], is more suitable in this case due to the fact that differential privacy requires the number of queries connected to the same individual to be known in order to provide user-level based privacy. In the case of streaming however, data is gathered continuously, making it impossible to estimate how many times a certain individual will produce events in the future.

Author	C	DA	DF	DI	P	SP	V
Kellaris et al. [41]					✓	✓	

TABLE X: All stream processing papers, showing categories covered by each paper

G. Data Format

In order to store and access big data, it can be structured in different ways. Out of the 19 papers labeled as data format papers, most used a *distributed file system*, *database* or *cloud* that made them qualify in this category. An overview of all papers from this section can be found in Table XI.

One example of combining data format and privacy is the work by Peng et al. [43] that focuses on query optimization under differential privacy. The main challenge faced when enforcing differential privacy on databases is the interactive nature of the database where new queries are issued in real-time. An unspecified number of queries makes it difficult to wisely spend the privacy budget, which essentially keeps track of how many queries can be asked, used to guarantee differential privacy, to still provide high utility of query answers. Therefore, Peng et al. implemented the query optimizer Pioneer, that makes use of old query replies when possible in order to consume as little as possible of the remaining privacy budget.

Furthermore, Sathiamoorthy et al. [44] focus on data integrity, and present an alternative to standard Reed-Solomon codes, which are erasure codes used for error-correction, that are more efficient and offer higher reliability. They implemented their erasure codes in the Hadoop's distributed file system, HDFS, and were able to show that the network traffic could be reduced, but instead their erasure codes required more storage space than traditional Reed-Solomon codes.

Lastly, Wang and Ravishankar [45] point out that providing both efficient and confidential queries in databases is challenging. Inherently, the problem stems from the fact that indexes invented to increase performance of queries also leak information that can allow adversaries to reconstruct the plaintext, as Wang and Ravishankar show. Consequently,

Wang and Ravishankar present an encrypted index that provides both confidentiality and efficiency for range queries, tackling the usual trade-off between security and performance.

Author	C	DA	DF	DI	P	SP	V
Peng et al. [43]			✓		✓		
Sathiamoorthy et al. [44]			✓	✓			
Wang and Ravishankar [45]	✓		✓				

TABLE XI: A set of data format papers, showing categories covered by each paper

IV. Discussion and Future Work

While this review investigates security and privacy for big data, it does not cover all papers available within the topic, since it would be infeasible to manually review them all. Instead, the focus of this review is to explore recent papers and to provide both a qualitative and a quantitative analysis, in order to create a snapshot of the current state-of-the-art. By selecting papers from top conferences and assessing their quality manually before selecting them, we include only papers relevant for big data, security and privacy.

A potential problem with only picking papers from top conferences is that, while the quality of the papers is good, the conferences might only accept papers with ground breaking ideas. After conducting this review, however, we believe most big data solutions with respect to security and privacy are not necessarily ground breaking ideas, but rather new twists on existing ideas. From the papers collected for this review, none of the topics covered are specific for big data, rather the papers present new combinations of existing topics. Thus, it seems that security and privacy for big data is not different from other security and privacy research, as the ideas seem to scale well.

Another part of the methodology that can be discussed is the two queries used to collect papers. Query A was constructed to cover a wide range of papers, and query B was set to only include big data papers. Unfortunately, query A contributed with far more hits than query B after the filtering step from Table I. This means that most papers might not have been initially intended for big data, but they were included after the quality assessment step, since the methods used were deemed scalable. Consequently, widening the scope of query B might include papers that present security or privacy solutions solely intended for big data.

Regarding the categories, *confidentiality* was covered by almost a third of the papers, but had no dominating topic. Rather, it contained a wide spread of different cryptographic techniques and access control. Furthermore, *privacy* was well represented, with 61 papers in the review. A large portion of these papers used differential privacy, the main reason probably being the fact that most differentially private algorithms are independent of the data set's size, which makes it beneficial for large data sets.

While *privacy* was covered by a large portion of papers, only two papers use an existing privacy-preserving data publishing (PPDP) technique. Moreover, one paper introduces a new PPDP technique called β-likeness. A reason for why this topic might not be getting a lot of attention is the fact that PPDP is dependent on the size of the data set. Thus PPDP is harder to apply to big data, since the entire data set must be processed in order to anonymize it. Consequently, further work may be required in this area to see how PPDP can be applied to big data.

We have also detected a gap in the knowledge considering *stream processing* and *visualization* in combination with either *data integrity* or *confidentiality*, as no papers covered two of these topics. *Data integrity* is also one of the topics that were underrepresented, with five papers out of 82 papers in total, which is significantly lower than the number of *confidentiality* and *privacy* papers. However, it might be explained by the fact that the word 'integrity' was not part of any of the queries. This is a possible expansion of the review.

V. Conclusion

There are several interesting ideas for addressing security and privacy issues within the context of big data. In this paper, 208 recent papers have been collected from A^* conferences, to provide an overview of the current state-of-the-art. In the end, 82 were categorized after passing the filtering and quality assessment stage. All reviewed papers can be found in tables in Section III.

Conclusively, since papers can belong to more than one category, 61 papers investigate *privacy*, 25 *data analysis*, 23 *confidentiality*, 19 *data format*, 5 *data integrity*, one *stream processing* and one *visualization*. Prominent topics were *differential privacy*, *machine learning* and *homomorphic encryption*. None of the identified topics are unique for big data.

Categories such as *privacy* and *data analysis* are covered in a large portion of the reviewed papers, and 20 of them investigate the combination of *privacy* and *data analysis*. However, there are certain categories where interesting connections could be made that do not yet exist. For example, one combination that is not yet represented is *stream processing* with either *confidentiality* or *data integrity*. *Visualization* is another category that was only covered by one paper.

In the end, we find that the security and privacy for big data, based on the reviewed papers, is not different from security and privacy research in general.

Acknowledgements

This research was sponsored by the BAuD II project (2014-03935) funded by VINNOVA, the Swedish Governmental Agency for Innovation Systems.

References

[1] G. Fuchs et al. "Constructing semantic interpretation of routine and anomalous mobility behaviors from big data". In: *SIGSPATIAL Special* 7.1 (May 2015), pp. 27–34.

[2] M. Chen et al. "Big Data: A Survey". en. In: *Mobile Networks and Applications* 19.2 (Jan. 2014), pp. 171–209.

[3] J. Ginsberg et al. “Detecting influenza epidemics using search engine query data”. English. In: *Nature* 457.7232 (Feb. 2009), pp. 1012–4.

[4] O. Tene and J. Polonetsky. “Privacy in the Age of Big Data: A Time for Big Decisions”. In: *Stanford Law Review Online* 64 (Feb. 2012), p. 63.

[5] S. Kaisler et al. “Big Data: Issues and Challenges Moving Forward”. English. In: *System Sciences (HICSS), 2013 46th Hawaii International Conference on*. IEEE, Jan. 2013, pp. 995–1004.

[6] D. Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Tech. rep. META Group, Feb. 2001.

[7] R. Lu et al. “Toward efficient and privacy-preserving computing in big data era”. English. In: *Network, IEEE* 28.4 (Aug. 2014), pp. 46–50.

[8] B. Kitchenham. *Procedures for performing systematic reviews*. Joint Technical Report. Keele, UK: Software Engineering Group Department of Computer Science Keele University, UK, and Empirical Software Engineering, National ICT Australia Ltd, 2004, p. 26.

[9] International Organization for Standardization. *Information technology – Security techniques – Information security management systems – Overview and vocabulary*. Standard. Geneva, CH: International Organization for Standardization, Feb. 2016.

[10] International Organization for Standardization. *Information technology – Security techniques – Privacy framework*. Standard. Geneva, CH: International Organization for Standardization, Dec. 2011.

[11] A. Liu et al. “Efficient secure similarity computation on encrypted trajectory data”. In: *2015 IEEE 31st International Conference on Data Engineering (ICDE)*. 2015 IEEE 31st International Conference on Data Engineering (ICDE). 2015, pp. 66–77.

[12] Y.-W. Chu et al. “Privacy-Preserving SimRank over Distributed Information Network”. In: *2012 IEEE 12th International Conference on Data Mining (ICDM)*. 2012 IEEE 12th International Conference on Data Mining (ICDM). 2012, pp. 840–845.

[13] G. Bender et al. “Explainable Security for Relational Databases”. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’14. New York, NY, USA: ACM, 2014, pp. 1411–1422.

[14] A. Meacham and D. Shasha. “JustMyFriends: Full SQL, Full Transactional Amenities, and Access Privacy”. In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’12. New York, NY, USA: ACM, 2012, pp. 633–636.

[15] C. Bear et al. “The vertica database: SQL RDBMS for managing big data”. In: *Proceedings of the 2012 workshop on Management of big data systems*. ACM, 2012, pp. 37–38.

[16] F. Zhu et al. “A Fast and High Throughput SQL Query System for Big Data”. In: *Web Information Systems Engineering - WISE 2012*. Ed. by X. S. Wang et al. Lecture Notes in Computer Science 7651. DOI: 10.1007/978-3-642-35063-4_66. Springer Berlin Heidelberg, 2012, pp. 783–788.

[17] H. Xiao et al. “Is Feature Selection Secure against Training Data Poisoning?” In: *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 2015, pp. 1689–1698.

[18] A. Arasu et al. “Secure Database-as-a-service with Cipherbase”. In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’13. New York, NY, USA: ACM, 2013, pp. 1033–1036.

[19] S. Lallali et al. “A Secure Search Engine for the Personal Cloud”. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’15. New York, NY, USA: ACM, 2015, pp. 1445–1450.

[20] M. Barbaro and T. Zeller. “A Face Is Exposed for AOL Searcher No. 4417749”. In: *The New York Times* (Aug. 2006).

[21] A. Narayanan and V. Shmatikov. “Robust De-anonymization of Large Sparse Datasets”. In: *IEEE Symposium on Security and Privacy, 2008. SP 2008*. May 2008, pp. 111–125.

[22] P. Samarati and L. Sweeney. *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*. Tech. rep. SRI International, 1998.

[23] L. Sweeney. “k-anonymity: A model for protecting privacy”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570.

[24] A. Machanavajjhala et al. “L-diversity: Privacy beyond k -anonymity”. In: *ACM Transactions on Knowledge Discovery from Data* 1.1 (2007), 3–es.

[25] N. Li et al. “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity.” In: *ICDE*. Vol. 7. 2007, pp. 106–115.

[26] C. Dwork. “Differential privacy”. In: *Automata, languages and programming*. Springer, 2006, pp. 1–12.

[27] H. Zakerzadeh et al. “Privacy-preserving big data publishing”. In: *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*. ACM, June 2015, p. 26.

[28] Y. Wang and B. Zheng. “Preserving privacy in social networks against connection fingerprint attacks”. In: *2015 IEEE 31st International Conference on Data Engineering (ICDE)*. 2015 IEEE 31st International Conference on Data Engineering (ICDE). 2015, pp. 54–65.

[29] P. Jurczyk et al. “DObjects+: Enabling Privacy-Preserving Data Federation Services”. In: *2012 IEEE 28th International Conference on Data Engineering (ICDE)*. 2012 IEEE 28th International Conference on Data Engineering (ICDE). 2012, pp. 1325–1328.

[30] J. Cao and P. Karras. “Publishing Microdata with a Robust Privacy Guarantee”. In: *Proc. VLDB Endow.* 5.11 (2012), pp. 1388–1399.

[31] G. Acs et al. “Differentially Private Histogram Publishing through Lossy Compression”. In: *2012 IEEE 12th International Conference on Data Mining (ICDM)*. 2012 IEEE 12th International Conference on Data Mining (ICDM). 2012, pp. 1–10.

[32] G. Cormode et al. “Differentially Private Spatial Decompositions”. In: *2012 IEEE 28th International Conference on Data Engineering (ICDE)*. 2012 IEEE 28th International Conference on Data Engineering (ICDE). 2012, pp. 20–31.

[33] X. Hu et al. “Differential Privacy in Telco Big Data Platform”. In: *Proc. VLDB Endow.* 8.12 (2015), pp. 1692–1703.

[34] D. Proserpio et al. “Calibrating Data to Sensitivity in Private Data Analysis: A Platform for Differentially-private Analysis of Weighted Datasets”. In: *Proc. VLDB Endow.* 7.8 (2014), pp. 637–648.

[35] F. D. McSherry. “Privacy integrated queries: an extensible platform for privacy-preserving data analysis”. In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 19–30.

[36] P. Jain and A. Thakurta. “Differentially private learning with kernels”. In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 2013, pp. 118–126.

[37] Y. Elmehdwi et al. “Secure k-nearest neighbor query over encrypted data in outsourced environments”. In: *2014 IEEE 30th International Conference on Data Engineering (ICDE)*. 2014 IEEE 30th International Conference on Data Engineering (ICDE). 2014, pp. 664–675.

[38] B. Yao et al. “Secure nearest neighbor revisited”. In: *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. 2013 IEEE 29th International Conference on Data Engineering (ICDE). 2013, pp. 733–744.

[39] E. Y. Gorodov and V. V. Gubarev. “Analytical review of data visualization methods in application to big data”. In: *Journal of Electrical and Computer Engineering* 2013 (Jan. 2013), p. 22.

[40] H. To et al. “PrivGeoCrowd: A toolbox for studying private spatial Crowdsourcing”. In: *2015 IEEE 31st International Conference on Data Engineering (ICDE)*. 2015 IEEE 31st International Conference on Data Engineering (ICDE). 2015, pp. 1404–1407.

[41] G. Kellaris et al. “Differentially Private Event Sequences over Infinite Streams”. In: *Proc. VLDB Endow.* 7.12 (2014), pp. 1155–1166.

[42] C. Dwork et al. “Differential privacy under continual observation”. In: Proceedings of the forty-second ACM symposium on Theory of computing. ACM, 2010, pp. 715–724.

[43] S. Peng et al. “Query optimization for differentially private data management systems”. In: *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. 2013 IEEE 29th International Conference on Data Engineering (ICDE). 2013, pp. 1093–1104.

[44] M. Sathiamoorthy et al. “XORing elephants: novel erasure codes for big data”. In: *Proceedings of the 39th international conference on Very Large Data Bases*. VLDB’13. Trento, Italy: VLDB Endowment, 2013, pp. 325–336.

[45] P. Wang and C. V. Ravishankar. “Secure and efficient range queries on outsourced databases using Rp-trees”. In: *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. 2013 IEEE 29th International Conference on Data Engineering (ICDE). 2013, pp. 314–325.