

# Sussex Research

## Rise of big data – issues and challenges

Bayan Alabdullah, Natalia Beloff, Martin White

### Publication date

06-06-2023

### Licence

This work is made available under the [Copyright not evaluated](#) licence and should only be used in accordance with that licence. For more information on the specific terms, consult the repository record for this item.

### Document Version

Accepted version

### Citation for this work (American Psychological Association 7th edition)

Alabdullah, B., Beloff, N., & White, M. (2018). *Rise of big data – issues and challenges* (Version 1). University of Sussex. <https://hdl.handle.net/10779/uos.23301014.v1>

### Published in

2018 21st Saudi Computer Society National Computer Conference (NCC)

### Link to external publisher version

<https://doi.org/10.1109/NCG.2018.8593166>

### Copyright and reuse:

This work was downloaded from Sussex Research Open (SRO). This document is made available in line with publisher policy and may differ from the published version. Please cite the published version where possible. Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners unless otherwise stated. For more information on this work, SRO or to report an issue, you can contact the repository administrators at [sro@sussex.ac.uk](mailto:sro@sussex.ac.uk). Discover more of the University's research at <https://sussex.figshare.com/>

# *Rise of Big Data – Issues and Challenges*

Bayan Alabdullah

Computer Science, Princess Nourah University; KSA  
Department of Informatics, School of Engineering and  
Informatics, University of Sussex, United Kingdom

Natalia Beloff, Martin White

Department of Informatics, School of Engineering and  
Informatics, University of Sussex  
United Kingdom

**Abstract** — The recent rapid rise in the availability of big data due to Internet-based technologies such as social media platforms and mobile devices has left many market leaders unprepared for handling very large, random and high velocity data. Conventionally, technologies are initially developed and tested in labs and appear to the public through media such as press releases and advertisements. These technologies are then adopted by the general public. In the case of big data technology, fast development and ready acceptance of big data by the user community has left little time to be scrutinized by the academic community. Although many books and electronic media articles are published by professionals and authors for their work on big data, there is still a lack of fundamental work in academic literature. Through survey methods, this paper discusses challenges in different aspects of big data, such as data sources, content format, data staging, data processing, and prevalent data stores. Issues and challenges related to big data, specifically privacy attacks and counter-techniques such as k-anonymity, t-closeness, l-diversity and differential privacy are discussed. Tools and techniques adopted by various organizations to store different types of big data are also highlighted. This study identifies different research areas to address such as a lack of anonymization techniques for unstructured big data, data traffic pattern determination for developing scalable data storage solutions and controlling mechanisms for high velocity data.

**Keywords-component; Privacy; Unstructured Big Data, Big Data Classification, Big Data Tools**

## I. INTRODUCTION

Big data is everywhere without certain origin. It is argued that the term ‘big data’ was coined as a result of lunch-table conversation at Silicon Graphic Inc. (SGI) in early 1990s. The usage of big data became widespread after “commercialism hype” created by technology companies in developing big data analytical markets.

The recent rapid rise in the availability of big data due to Internet-based technologies such as social media platforms and mobile devices has left many market leaders unprepared for handling very large, random and high velocity data. Big data is ubiquitous, for example, books in libraries are tagged and tracked, while smart phones are replete with large numbers of applications that collect huge amount data. Other devices, such as healthcare machines record heartbeat, blood pressure, hemoglobin and sleep habit data every minute. All of these examples are produce big data where companies are exploiting user preferences into commercial profits, which may compromise the privacy of users. Conventionally, technologies are initially developed and tested in labs and appear to the public through media as press releases and advertisements. These technologies are then adopted by the general public. Fast development and ready acceptance of big data by the public left little time to be matured by the academic domain. Although many books and electronic media articles are published by

professionals and authors for their work on big data, yet fundamental work is still lacking in academic publications [1]. This survey focused on issues and challenges in different areas of big data, and in particular we address the privacy problem in unstructured big data. This survey also highlights different privacy attacks and the loop holes in existing privacy preservation techniques.

## II. BIG DATA CLASSIFICATION

The features of big data, such as volume and variety, are dominated by valuable information that created hype for the large-scale marketing efforts of software and hardware companies trying to sell their particular ‘big data solutions’. The commercial sector is more focused on developing big data solutions that largely target structured data. This leaves a large portion of big data ignored, such as text messages, videos and audio files captured from mobile devices — this largely ignored unstructured data is much harder to analysis, which makes it more difficult for companies to provide commercial big data solutions. A recent study shows that the largest portion of big data consists of unstructured data, while structured data is only a small subset of such data [1].

Big data exists in formats with various different characteristics. Classification of big data is important in order to understand the strengths and weaknesses of applications that process large dataset volumes. Big data can be classified based on its categories, such as data storage, contents formats and data staging [2]. All of these categories have their own characteristics and dependencies. The classification of big data is provided as follows:

### A. Data Sources

Data sources are usually referred as data production points. Among them, social media is one of the most relevant and representative sources of big data. Big data is generated by social media platforms through internet-based applications and websites such as Facebook, Twitter, Instagram, Flickr, YouTube, Google and Word Press [2]. These websites allow users to get connected and form a kind of virtual community where people share and collaborate on different topics. Since personal and inter-personal information is shared among the community, the misuse of such information can be consequential and influential [3]. Thus, the prevention of this information from different attacks is of extreme importance.

Another source of big data is the Internet of Things (IoT), which is based on large number of sensors that collaboratively operate

to generate huge amounts of data. Data is generated from sensing devices including mobile devices, satellites and other sensors related to healthcare and weather stations [4]. Recently, newly emerged smart phones, tablets, cameras and other sensing devices are being classified as part of this object-group. The connectivity of these devices over the Internet enables smart processing and provide services in many domains such as healthcare, banking and finance. The connectivity of large number of heterogeneous devices produce huge data [6], which includes features such as heterogeneity, variety, unstructured feature, noise, and high redundancy. Three different characteristics of data generated from IoT devices confirm it as big data. These characteristics are: (i) large amounts of data are generated from abundant terminals; (ii) IoT devices generate data that is not structured, and (iii) The data generated is only usable if it can be analysed in near real time. Acquiring, integrating, processing, storing and using IoT for these datasets becomes immediate and important research problems for enterprises to achieve their business goals [7].

### B. Content Format

Content formats of big data can also be used for classification. Different types of big data based on content format is as follows:

1. **Structured data:** Referred to as the data which can be input, stored, queried and analyzed easily. Such data is managed using a programming language, SQL, and stored in relational databases such as Oracle, DB2, Teradata, MySQL, PostgreSQL [8]. The transaction type of such data is Online Transaction Processing (OLTP). Examples of structured data include text, digits and dates [9].
2. **Unstructured data:** This data comes in various formats such as text messages over social media, cell phone location information, videos taken from CCTV cameras and other related social media contents from variety of sources such as mobile phones, tablets, IoT devices, social media applications and satellite images. Such data also exists in the form of web pages, images, audios and videos [8]. The size of this type of data is ever increasing due to increasing numbers of smart phones and social media applications, handling such data is a prominent challenge [9].
3. **Semi-structured data:** This type of data does not follow the rules of conventional database systems. This type of data could be stored in relational database tables. Data capturing for such type of data is different from fixed file format data and required usage of complex rule-based system for the next process to follow after data capturing process [9]. Such data needs dynamic processes for complex rules during operations on data. Thus, the complexity of handling such a versatile data source is an open research issue.

### C. Data Staging

Raw data is not in a valid format so it cannot be directly used for analysis. For example, consider unstructured data collected

from social media consisting of audio, video and images. All of the data is in a format that requires processing to clean and convert into a structured format so that it can be easily analyzed [8]. The rest of the data can be identified as garbage where a process of cleaning is conducted by identifying such data [10]. In another type of staging, new data is generated by transforming existing data from one format to other due to a business requirement. This staging process removes anomalies in data and is called normalization [11]. The techniques for cleaning data during this process requires further research efforts to minimize data loss.

### D. Data Processing

Data can be classified based on the type of processing that generates the data. Such types of processing are given as follows:

1. **Batch:** Long-running jobs, which are also named as batch-processing jobs, are executed in systems such as MapReduce by many organizations [12]. The applications developed for such systems are scaled up to hundreds of nodes in the form of clusters. Efficient scaling of nodes during the process requires further research effort.
2. **Real time:** Data processing can be done using real-time systems such as the Simple Scalable Streaming System (S4) [13]. Continuous and unbounded streams are programmed by programmers on a S4 distributed computing platform with effective fault tolerance and scalable platform. Synchronization and results composition are issues that require further research effort in this domain [14].

### E. Data Stores

The analytics require clusters of data storage for effective and timely output from big data. Traditional relational database models are not designed for very large-scale datasets, thus performance issues arise during big data analytics. As a solution, No-SQL databases are preferred over SQL databases for processing due to the ability of horizontal partitioning of data, extensive processing capability and better performance [15]. Companies such as Google, Facebook, Amazon and LinkedIn use NoSQL for handling continuously increasing data streams. NoSQL databases can be classified into following three different data store formats:

1. **Document-oriented:** Documents such as PDF or MS Word and several different formats such as Java Script Object Notation (JSON) and Extensible Markup Language (XML) are stored in document-oriented data stores [2]. One document in a data store is equivalent to a row in a relational database where the query is applied to the contents of the document. Example of such data stores are MongoDB (open source, document-oriented storage system which stores documents as Binary JSON (BSON) objects), SimpleDB (distributed data storage system exposed via Amazon API as web services. Data can be stored

in different domains), and CouchDB (document-oriented database written in Erlang).

2. Column-oriented: These databases store data in columns along with attributes rather than in rows [16]. BigTable, Cassandra, HBase and HyperTable are examples of Column-oriented data storage engines. One of the challenges in Column-oriented databases is the difficulty in data profiling, which needs further investigation [17].
3. Graphs database: This database is designed based on graph theory. The nodes and edges represented properties of relations and their link to store data [18]. Dryad is an example of Graph database which is a general-purpose data processing engine for unstructured data. However, selection of appropriate graph platform for benchmarking is an open research challenge [19].
4. Key-value: Very large datasets are stored in Key-value based data stores. The data is accessed via a key using different algorithms [20]. One of the examples of such systems is Dynamo which is used by amazon.com for some services. Extension of such single key-value system is transactional multi-key systems [21].

### III. ISSUES AND CHALLENGES WITH BIG DATA

Traditional relational databases are obsolete and cannot store and process the data generated from recent business applications [18]. Daily life problems such as recording data, cost of data storage and synchronization problems prompt us to use NoSQL solutions [22]. Typical characteristics, diverse data types and patterns, complex relationships and greatly varied data equality [23] are other challenges in handling big data. The complex types, structure and patterns of big data make it difficult to perceive, represent, understand and compute using traditional computational models. Traditional data analysis operations such as data retrieval, semantic and sentiment analysis are complex operations in big data [24]. There exists a clear lack of understanding in the laws of distribution and big data association relationships. Another major issue is apparent in describing quantitatively different characteristics of the complexity of big data.

Performing operations on big data such as machine learning, data analytics and data mining [18] are important challenges to address. This data cannot be handled by traditional algorithms, past statistics and analysis tools which are basically designed for small datasets [25]. Since the older approaches are developed based on the assumptions of independent and uniform distribution of data which are further supported by reasonable samples with reliable statistics, thus big data computation requires re-validating these approaches considering their computational complexity and algorithms used.

Traditional computational frameworks, system architectures and processing systems are designed to handle structured data [24]. To perform scientific research on big data, systems and frameworks suitable for handling large and diverse data types

are mandatory. Huge volume, complex data structures and non-uniform distribution of data make computation complexity very high with long duty cycle and real-time requirements. These requirements not only require designs of computing frameworks, system architecture and processing systems but also constrain operational efficiency and energy consumption.

Existing methods for handling large scale data cannot handle big data and thus require new technologies such as cloud computing and grid computing to solve the scaling issue of big data [26]. A more focused and fine-grained problem is the processing time of large data, which is significantly higher than small data sets. This leads to delayed analysis in time critical applications such as robotics, space science and healthcare [27]. In areas such as the credit card industry and traffic management systems requires quick response from the analytics to take appropriate action [28]. Scheming such structures becomes a new challenge when the data volume is growing rapidly and queries require reduced response time.

Privacy is one of the major concerns in big data. Yet, there are not much solutions that exist in this regard. It is believed that industry solutions to the privacy problem in big data will emerge from research outputs in this domain. Before proposing any solution in this domain, it is important to review existing solutions from both application and theoretical perspectives. Therefore, a review of existing solutions, frameworks, mathematical descriptions, measurements and modelling perspective is provided.

While there exists many research challenges in big data, this survey mainly covers the privacy problem.

#### A. The Privacy Problem in Big Data

Privacy has been largely studied in past decades. Primarily, previous studies cover cryptography, communication and information theory. Considering the very large size of big data, it is difficult to use existing cryptographic solutions effectively. Another limitation is imposed by limited processing and storage capacity of mobile devices, which make encryption and decryption a non-feasible solution [29]. Thus, conventional cryptographic solutions are not suitable for emerging requirements of big data.

Failure of simple anonymisation techniques increases challenges in the era of big data. There is no clear definition of privacy because it is a subjective concept [29]. Therefore, it is difficult to reach a global definition of privacy. Moreover, fast adoption of big data raises questions on the reliability of existing techniques. As a consequence, it becomes important to study existing privacy studies in terms of big data environment and work on new algorithms, models and frameworks to cope up with the privacy challenges in big data. Privacy studies in the domain of big data can be mainly divided into two categories as follows:

#### B. Data Clustering

One of the popular technique for data processing is clustering due to its ability to analyse un-familiar data. The basic idea of

clustering is to divide data without labels into different groups. However, major issues with existing clustering algorithms is their dependency on one data format that is in conflict with basic characteristic of unstructured big data which is variety. A Brief overview of different data clustering techniques and related privacy issues is provided as follows:

#### a) *K-anonymity*

Studies for effectively limiting the disclosure of identity of users in anonymised tables were conducted by Samarati and Sweeney [30], [31]. Data protection using  $k$ -anonymity is quite simple and easy to understand.  $K$ -anonymity can be defined as the property which distinguishes each record from  $k-1$  other records based on a quasi-identifier. It means that at least  $k$  records are required in each equivalence class to achieve anonymity. For example, if all records in a table satisfy  $k$ -anonymity condition, then for some value of  $k$ , a record can be identified with  $1/k$  confidence if quasi identifiers are known.  $K$ -anonymity focus on quasi- identifiers attributes and invest no effort on sensitive attributes. As a consequence, it is susceptible to many attacks such as homogeneity attack and background attack.

On the basis of  $k$ -anonymity, different algorithms, models and frameworks are proposed to solve different privacy attacks. A method of two-level vertex anonymization against a neighbourhood attack is proposed in [32]. In this method, anonymization is achieved in two steps. In the first step, a more generic label is applied to field of table instead of specific value and in second level, the edges are altered, but vertices are not

changed. In another technique,  $k$ -anonymity is applied to social

network graphs in two steps. In the first step, a neighbourhood of a vertex is identified using an encoding technique. The second step focuses anonymization by grouping together the vertices of same degree.

Another method of link anonymization was proposed in [33] based on neighbour randomization scheme. The process of anonymization of a sensitive link between two nodes is referred to as link anonymization. The key idea is to hide either source or destination node of a link to make it hard to find the exact existence of a link. The probability of a correct destination is ' $p$ ' and wrong destination is ' $1-p$ '. This approach is only suited to social-network data anonymization which is only addressed by one type of attack, i.e. the neighbourhood attack. However, there exist a wide variety of attacks such as Sybil attack [34], error tolerance of complex networks [35], attributes disclosure attack and background attack that cannot be prevented using  $k$ -anonymity [36]. A list of attacks on social networking sites to access user personal information is provided in [37]. Thus, it is evident that we need to discover more relevant and effective privacy preservation techniques.

#### b) *L-diversity*

The major feature of the  $k$ -anonymity technique is its strength against identity disclosure and neighbourhood attack. However, it is not enough to provide safety against attribute disclosure, Sybil attack, error tolerance and background attack [38]. To

address this limitation,  $l$ -diversity was proposed as a new privacy technique [39]. The  $l$ -diversity technique is based on dividing attributes into sensitive and non-sensitive attributes. In a relational database table, there are two factors that influence the privacy of a record. One of the factors is uniformity in the key attributes of a table. Another factor is knowledge of adversary about a particular record in the table. If a record is correctly identified by an attacker, it is called positive disclosure otherwise it is called negative disclosure. However, the assumption of a uniform global distribution of an attribute makes it sensitive for adversarial attack. The concept of  $l$ -diversity is generally adopted to overcome background and homogeneity attacks. However, it is insufficient to prevent attribute disclosure [36]. This limitation requires further research effort to develop more secure privacy preservation mechanism.

#### c) *T-closeness*

To address the global background knowledge problem associated with the  $l$ -diversity technique, the idea of  $t$ -closeness is proposed [38]. It is based on the notion that an attribute in equivalence class and table have a close distribution that is not more than a threshold  $t$ . However, a measurement of distance between distribution  $P$  and  $Q$  is challenging. The distance is computed using Earth Mover's Distance (EMD), which is a well-known problem in transportation. Assuming two different distributions,  $P$  and  $Q$  with defined set of elements in each and the ground distance between two elements of each set is  $d_{ij}$ , then the work can be defined as follows:

$$WORK(P, Q, F) = \sum_{i=1}^m \sum_{j=1}^m d_{ij} \quad (1)$$

All the models discussed so far suffer from the online availability of datasets. An attacker having enough knowledge of published datasets can easily estimate the status of sensitive attributes. Another important problem is that the  $t$ -closeness technique does not deal with the identity disclosure explicitly [36]. To overcome this issue, the idea of differential privacy is proposed in which some noise in the data is added through a query. The main goal is the query results about an individual record that are generated based on entire dataset.

#### C. *Differential Privacy*

Differential privacy is a new method for big data privacy preservation. It is based on the equal probability for all similar inputs and therefore, all outputs are insensitive to individuals [40]. Differential privacy can be defined as a random function  $K$  with  $\epsilon$ -differential privacy value if data values vary from each other for at most one row [41]. The value for ' $\epsilon$ ' depends on the dataset and query applied on it. The differential privacy plays important role by inserting additional layer between dataset and the user.

In terms of the privacy value ' $\epsilon$ ', differential privacy can be defined as:  $\epsilon$ -differential privacy can be achieved using a randomized function  $K$ , if for two given datasets  $D1$  and  $D2$  are differentiating from each other on at most one row ( $D1$  and  $D2$

are also called neighbouring datasets), and all  $S \in \text{Range}(K)$  [42].

$$\Pr[K(D1) \in S] \leq e^\epsilon \times \Pr[K(D2) \in S] \quad (2)$$

In other words, the difference is not greater than a factor  $e^\epsilon$  between two datasets after anonymization operation.

Adding noise in the output to achieve privacy is a common method in differential privacy. For this, there are two major techniques called Laplace mechanism [43] and Exponential mechanism [44]. Laplace distribution is followed for noise generation in real outputs. However, if outputs are not real then an exponential mechanism is followed, which assigns a higher probability to the desired outputs. As a result, the final output from the mechanism is close to the real world desired output. However, differential privacy completely depends on the amount of noise added by the curator. Compromised curator can also fail the whole system [45]. Thus, a mechanism to check the curator for its reliability needs to be developed.

#### IV. DIFFERENT TYPES OF PRIVACY ATTACKS

De-anonymisation is the reverse process of anonymisation. It is the process of exposing real information in a dataset by applying many different re-identification techniques. These techniques include linking datasets, background knowledge, pattern-matching and location information. The process of de-identification is usually performed by third parties who want to perform analytics of their own interest. A de-identification process is usually based on a quasi-identifier, which are highly correlated variables, but they still cannot de-identify data perfectly. Quasi-identifiers require all variables of a composite identifier to identify a record correctly. Different types of re-identification attacks are launched by attackers on different type of datasets. These attacks can generally be classified as information aggregation attack, re-identification attack, Graph/Node attack and location attacks [46].

Cross-site profile generation and profile cloning are two different, but important identity theft attacks that are important in the context of social media [37]. These attacks are profile cloning and cross-site profile cloning. A profile cloning attack is launched by cloning existing profiles of victims on a social network and sending friend request to all contact of the original profile user. Thus, the contacts of a legitimate user are forged and a second identical and forged profile is created. The sensitive personal information of the victim is easily revealed through the contacts of the victim [47]. In a cross-site profile cloning attack, users that exist on one social media platform but do not exist on other platform are identified [48]. The profile of the user is cloned from original site and then automatic attempt to rebuild forged identity is launched by sending friend requests to the connections that are identified on both platforms. This is an important attack because the targeted profile once existed on a social network. A game based model which is also called collusion attack model in proposed in [49]. In this model, the attacker attempts to acquire all encryption keys of data and wins

if a valid secret key is acquired. A background attack is an important attack on  $k$ -anonymized datasets that can be launched by an attacker having knowledge of a target dataset [50], [51]. Another attack related to  $k$ -anonymized dataset is the homogeneity attack, which happens when important attributes are homogenous and lack diversity of data [50]. If an attacker has some information about the neighbours of a target and relation among neighbours, the target can be re-identified using the information from the social network even if the targeted individual is anonymized using conventional privacy preservation techniques. This type of attack is called a neighborhood attack [52]. A different type of attack is launched by joining two different databases and thus called a joining attack [53]. A complementary release attack occurs when records are linked together on different releases of the same dataset [54]. This linking of records can compromise  $k$ -anonymity. Mixing of data from different data holders can solve the problem up to some extent. However, complete prevention from such an attack is difficult to achieve. Thus, it requires comprehensive research effort to ensure privacy of data against maximum attacks.

#### V. CONCLUSION AND FUTURE WORKS

This survey classified big data into different categories and identified related issues. Ensuring privacy of user data is identified as a big research challenge. A case of disclosure of Facebook<sup>1</sup> user data is raised recently. IoT devices has also recently emerged as new big data generators. Handling of issues such as heterogeneity of IoT devices, large and continuous streams of data are relevant research challenges. Similarly, designing a framework for handling very large scale data sets, optimal processing of semi-structured data, node scaling and fault tolerant storage structure are important areas to be considered. Typical characteristics, diverse data types and patterns, complex relationship and varying data equality are issues in data mining and data analytics that require uniform distribution of data.

Mainly, this survey focused on the privacy problem and the techniques that are used to handle the anonymity of users. It is observed that while all existing techniques work well for small-scale structured and uniform data, these are insufficient to ensure privacy in non-uniform, disturbed and very large volumes of unstructured data.

#### REFERENCES

- [1] A. Gandomi and M. Haider, "Beyond the hype : Big data concepts , methods , and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137–144, 2015.
- [2] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of 'big data' on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98–115, 2015.
- [3] J. N. Cappella, "Vectors into the Future of Mass and Interpersonal Communication Research: Big Data , Social Media , and Computational Social Science," *Human Communication Research*, vol. 43, no. 4, pp. 545–558, 2017.
- [4] U. S. Pushpa, "A Review of Big Data and Anonymization

<sup>1</sup><http://www.bbc.co.uk/news/technology-43465968>

- Algorithms,” *International Journal of Applied Engineering Research*, vol. 10, no. 17, pp. 13125–13130, 2015.
- [5] J. Gubbi, R. Buyya, and S. Marusic, “Internet of Things ( IoT ): A Vision , Architectural Elements , and Future Directions,” *Elsevier Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.
- [6] P. R. B. B. P. Saluja, N. Sharma, A. Mittal, and S. V. Sharma, “Cloud Computing for Internet of Things & Sensing Based Applications,” *Sensing Technology (ICST)*, pp. 374–380, 2012.
- [7] H. Cai, B. Xu, L. Jiang, and A. V. Vasilakos, “IoT-Based Big Data Storage Systems in Cloud Computing: Perspectives and Challenges,” *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 75–87, 2017.
- [8] S. Srivastava, “Appraising a Decade of Research in the Field of Big Data ‘The Next Big Thing,’” *Computing for Sustainable Global Development (INDIACom)*, no. 2014, pp. 2171–2175, 2016.
- [9] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.
- [10] E. Rahm and H. Hai Do, “Data Cleaning: Problems and Current Approaches,” *IEEE Data Eng.*, vol. 23, no. 4, pp. 3–13, 2000.
- [11] J. Quackenbush, “Microarray data normalization and transformation,” *Nature Genetics*, vol. 32, no. december, pp. 496–501, 2002.
- [12] J. P. Prathibha and E. . Dileesh, “Design of a Hybrid Intrusion Detection System using Snort and Hadoop,” *International Journal of Computer Applications*, vol. 73, no. 10, pp. 5–10, 2013.
- [13] Y. Chen, S. Alspaugh, and R. Katz, “Interactive analytical processing in big data systems,” *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 1802–1813, 2012.
- [14] R. Casado and M. Younas, “Emerging trends and technologies in big data processing,” 2014.
- [15] S. Venkatraman, S. Kaspi, Kiran Fahd, and R. Venkatraman, “SQL Versus NoSQL Movement with Big Data Analytics,” *International Journal of Information Technology and Computer Science*, vol. 8, no. 12, pp. 59–66, 2016.
- [16] D. J. Abadi, P. A. Boncz, and S. Harizopoulos, “Column-oriented Database Systems,” *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1664–1665, 2009.
- [17] F. Naumann, “Data Profiling Revisited,” vol. 42, no. 4, 2013.
- [18] M. Junghanns, M. Neumann, and E. Rahm, “Management and Analysis of Big Graph Data : Current Systems and Open Challenges,” in *Handbook of Big Data Technologies*, 2017, pp. 457–505.
- [19] T. Hegeman and P. Boncz, “Graphalytics : A Big Data Benchmark for Graph-Processing Platforms.”
- [20] M. Seeger, “Key-Value stores : a practical overview,” pp. 1–21, 2009.
- [21] S. Sharma, U. S. Tim, S. Gadia, and J. Wong, “Proliferating Cloud Density through Big Data Ecosystem , Novel XCLOUDX Classification and Emergence of as-a-Service Era,” 2008.
- [22] R. Estrada and I. Ruiz, “Big Data, Big Challenges,” *Big data SMACK*, pp. 3–7, 2016.
- [23] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, “Significance and Challenges of Big Data Research,” *Big Data Research*, vol. 2, no. 2, pp. 59–64, 2015.
- [24] S. Akter and S. F. Wamba, “Big data analytics in E-commerce : a systematic review and agenda for future research,” *Electronic Markets*, pp. 173–194, 2016.
- [25] H. Wang, Z. Xu, H. Fujita, and S. Liu, “Towards felicitous decision making: An overview on challenges and trends of Big Data,” *Information Sciences*, vol. 367–368, pp. 747–765, 2016.
- [26] H. Venkatesh, S. D. Perur, and N. Jaliha, “A Study on Use of Big Data in Cloud Computing Environment,” *Int. J. Comput. Sci. Inf. Technol. (IJCSIT)*, vol. 6, no. 3, pp. 2076–2078, 2015.
- [27] B. Baesens, R. Bapna, J. R. Marsden, J. Vanthienen, and J. L. Zhao, “Transformational issues of big data and analytics in networked business,” *MIS quarterly*, vol. 38, no. 2, pp. 629–631, 2014.
- [28] S. Amini and C. Prehofer, “Big Data Analytics Architecture for Real-Time Traffic Control,” *Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2017.
- [29] S. Yu, “Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data,” *IEEE Access*, vol. 4, pp. 2751–2763, 2016.
- [30] P. Samarati, “Protecting respondents’ identities in microdata release,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.
- [31] P. Samarati and L. Sweeney, “Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression,” *Proc of the IEEE Symposium on Research in Security and Privacy*, pp. 384–393, 1998.
- [32] B. Zhou and J. Pei, “The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks,” *Knowledge and Information Systems*, vol. 28, no. 1, pp. 47–77, 2011.
- [33] A. Milani and F. Ke, “Neighborhood Randomization for Link Privacy in Social Network Analysis,” *World Wide Web*, vol. 18, no. 1, pp. 9–32, 2015.
- [34] R. Liu and H. Wang, “Privacy-preserving data publishing,” *Proc. Workshops of 26th Int. Conf. on Data Engineering*, vol. 42, no. 4, pp. 305–308, 2010.
- [35] P. Crucitti, V. Latora, M. Marchiori, and A. Rapisarda, “Error and attack tolerance of complex networks,” *Nature*, vol. 340, pp. 388–394, 2004.
- [36] J. Surana, A. Khandelwal, and A. Kothari, “Big Data Privacy Methods,” *IJEDR*, vol. 5, no. 2, pp. 979–983, 2017.
- [37] L. Bilge, T. Strufe, D. Balzarotti, E. Kirda, and S. Antipolis, “All Your Contacts Are Belong to Us : Automated Identity Theft Attacks on Social Networks,” *WWW ’09 Proceedings of the 18th international conference on World wide web*, pp. 551–560, 2009.
- [38] N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and L-Diversity,” in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference*, pp. 106–115, 2007.
- [39] A. Machanavajjhala, D. Kifer, and J. Gehrke, “L-Diversity : Privacy Beyond k -Anonymity,” vol. 1, no. 1, 2007.
- [40] H. Ye, X. Cheng, M. Yuan, L. Xu, J. Gao, and C. Cheng, “A Survey of Security and Privacy in Big Data,” in *16th International Symposium on Communications and Information Technologies (ISCIT)*, 2016, pp. 268–272.
- [41] C. Dwork, “Differential Privacy,” pp. 1–12, 2006.
- [42] S. Yu and S. Member, “Big Privacy : Challenges and Opportunities of Privacy Study in the Age of Big Data,” *IEEE Access*, vol. 4, pp. 2751–2763, 2016.
- [43] C. Dwork, “Calibrating Noise to Sensitivity in Private Data Analysis,” *Journal of Pharmaceutical Sciences*, vol. 100, no. 8, pp. 3441–3452, 2011.
- [44] F. McSherry and K. Talwar, “Mechanism Design via Differential Privacy,” in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS’07)*, 2007, pp. 94–103.
- [45] P. Jain, M. Gyanchandani, and N. Khare, “Big data privacy : a technological perspective and review,” *Journal of Big Data*, 2016.
- [46] J. Henriksen-bulmer and S. Jeary, “International Journal of Information Management Re-identification attacks — A systematic literature review,” *International Journal of Information Management*, vol. 36, no. 6, pp. 1184–1192, 2016.
- [47] J. Joshi and B. Palanisamy, “Towards Risk-aware Policy based Framework for Big Data Security and Privacy.”
- [48] F. S. Rizi, M. R. Khayyambashi, and M. Y. Kharaji, “A New Approach for Finding Cloned Profiles in Online Social Networks,” *arXiv:1406.7377*, vol. 6, no. April, pp. 25–37, 2014.
- [49] K. Liang, W. Susilo, and J. K. Liu, “Privacy-preserving ciphertext multi-sharing control for big data storage,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 8, pp. 1578–1589, 2015.
- [50] N. Victor, D. Lopez, and J. H. Abawajy, “Privacy models for big data : a survey,” *International Journal of Big Data Intelligence*, vol. 3, no. 1, 2016.
- [51] S. Kumar, N. Mishra, and S. Sharma, “Privacy in Social Networks : A Survey,” pp. 125–130.
- [52] J. P. Zhou, Bin, “Preserving Privacy in Social Networks Against Neighborhood Attacks,” in *ICDE*, 2008.
- [53] M. Mani, Papri, Bhattacharya, “A Brief Survey on Vertex and Label Anonymization Techniques of Online Social Network Data Papri Mani \*, Munmun Bhattacharya \*\*,” *Int. Journal of Engineering Research and Applications*, vol. 5, no. 6, pp. 38–42, 2015.
- [54] N. Maheshwarkar, K. Pathak, and V. Chourey, “Privacy Issues for K-anonymity Model,” *International Journal of Engineering Research and Applications (IJERA)*, vol. 1, no. 4, pp. 1857–1861, 1861.