Mesopotamian journal of Big Data
Vol. **(2024)**, 2024, **pp**. 118–127
DOI: https://doi.org/10.58496/MJBD/2024/009 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Research Article

Using Data Anonymization in big data analytics security and privacy

Abdulatif Ali Hussain[1,*], Ismael Khaleel[2], Tahsien Al-Quraishi[3]

[1] *Al-Naji University, Iraq.*

[2] *Sunni Endowment Diwan, Iraq.*

[3] *Information Technology and Systems, Victorian Institute of Technology (VIT), Melbourne 3000, Australia.*

ARTICLE INFO

Article History

Received 11 May 2024
Accepted 16 Jul 2024
Published 10 Aug 2024

Keywords

Big Data
complex collection
data analytics
security and privacy
valuable insights

ABSTRACT

Big Data and Analytics mean an enormous and complex collection of very diverse information, which is processed with various technologies and methods to produce and deliver useful and valuable insights. Analytics is the science of using data, or information to extract useful and actionable insights, facts and knowledge from a collection of data it could be stated that Big Data Analytics is the best thing since every commercial data system ever built, although everybody with a more optimistic vision of technology would like to take note that there is a fine line where Everything Data crosses the boundary to something else, especially with regard to privacy and security of the world as we know it. Privacy and security are two distinct but closely related phenomena. Whereas privacy refers to the control over access to the individual, security refers to the stability or strength of controls designed to protect the individual's privacy. There are many obvious considerations and obstacles when attempting to securely share data. During big data analytics, many invasive techniques such as data fusion, cross-correlation, and algorithm training are often conducted over shared data, which can lead to severe privacy leaks. This means that every enterprise, organization, and individual maintaining large data repositories are in danger of being breached. Our study teaches us that security, privacy, and ethical concerns in big data analytics do not exist in parallel to the business cycle, but must be wisely and ethically managed in coherence throughout all emerging processes of the big data and information systems.

1. INTRODUCTION

The concept of Big Data Analytics is explained from a technical perspective: what it means, what it comprises, and why it means certain things. Bigger is better, or so the saying goes. In the 21st century, the world is becoming bigger, rounder and fuller, especially when it comes to collecting and keeping things. Collections of things like data are becoming larger and larger, more and more complex from all kinds of different sources [1]. In this context, the word big in Big Data Analytics might be better described as gigantic. Daily, zettabytes and should be considered beyond petabytes (10^{15}). Analytics is the science of using data, or information to extract useful and actionable insights, facts and knowledge from a collection of data.

Together, Big Data and Analytics means an enormous and complex collection of very diverse information, which is processed with various Byzantine technologies and methods to produce and deliver useful and valuable insights. On the one hand, this vision, for example, explains how something (usually very bad) happens here on earth and, as a consequence, that something else happens there (usually very good, and usually connected to the military, banking, fraud detection, or spying areas). On the other hand, it is also a very good introduction to the realm of Everything Data (the educational aspect), since it goes deep into what the words data, information and knowledge actually mean [2]. In this perspective, it could be stated that Big Data Analytics is the best thing since every commercial data system ever built, although everybody with a more optimistic vision of technology would like to take note that there is a fine line where Everything Data crosses the boundary to something else, especially with regard to privacy and security of the world as we know it. Here, privacy is defined as the degree to which personal information is shared, while security denotes the methods, mechanisms and technologies that protect privacy.

*Corresponding author. Email: lateef1960@yahoo.com

1.1. Definition and Scope of Big Data Analytics

Big Data Analytics is the process of examining large amounts of data of a variety of types to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. There are three popular definitions of Big Data: from 1) a business perspective, 2) a research perspective and 3) an ongoing public debate. (1) "Big Data is high-volume, high-velocity and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" [3]. This definition describes characteristics, i.e., volume, velocity, variety and value, but does not specify the technology (Technologies or business). (2) "Big Data is data that is too large, too fast, or too cheap to be handled by the current generation of technologies" [3]. This description indicates general technology-related challenges. However, it does not provide clear criteria for characterizing something as Big Data. (3) There are various ongoing public debates on the potential impact of Big Data on business competitiveness, privacy, ethics of data collection and analysis, etc.
Gartner's definition is widely accepted, but most of them originate in the fields of social and digital media, health, and sciences, where huge amounts of data are generated on an ongoing basis [3]. Big Data can be characterized in term of volume, velocity, variety and value. In general, Big Data can be seen as extremely large datasets that require a scalable, distributed hardware/software technology architecture to capture, store, manage, analyze and explore. In recent years Big Data has gained increasing attention from academia, industries and the media. Big Data has many potential applications and can benefit many fields and sectors. Businesses and public organizations are focusing on ways to harness data for real-time intelligence [1]. State services are investigating how to combine and analyze existing databases. Police organizations monitor social networks to predict where riots will occur. Banks seek novel technologies to detect possibly fraudulent claims in real time. Massive online stores adjust prices every second according to customer's locality, history and behavior. Besides its potential benefits, Big Data raises a number of critical issues.

1.2. Importance of Security and Privacy in Big Data Analytics

Big data represents a significant technological advance with the potential to yield insights in business, science, and healthcare that are impossible without it. It can be understood as an extremely large dataset or datasets for which established database management technology is inadequate. Big data analytics describes the technologies employed to glean insights from big data. The rapidly evolving field finds its current relevance in recent, vast increases in the ability to collect and process data from publicly available sources, including social media and publicly accessible websites. Surprisingly little attention has been devoted to the security and privacy implications of this emerging field. Throughout the history of computing, new architectures or technologies (hardware or software) uncovered unforeseen vulnerabilities or security threats, and Big Data analytics is no different. There is therefore concern that these novel vulnerabilities will go unaddressed or be inadequately addressed given the rapid pace of development of the field. The relevance of this question intensifies in light of the rapid growth in Big Data analytics capabilities and concerns about online monitoring or data collection by private companies [1].
Privacy and security are two distinct but closely related phenomena. Whereas privacy refers to the control over access to the individual, security refers to the stability or strength of controls designed to protect the individual's privacy. As control of personal data has shifted from individuals to organizations or other entities, the competencies required to safeguard privacy must also shift from the individual to these organizations. Moreover, privacy and security solutions must address structural inequities in power and control which have resulted in violations of privacy and abuses of secured data. Privacy must be understood as a context-appropriate, social good, and security must be understood as effective protection of privacy. There may be complex relationships between security and privacy, which may or may not be positive or protective safeguards [4].

2. SECURE DATA SHARING IN BIG DATA ANALYTICS

Big data analytics (BDA) is booming as a key industry driver and performance enhancer. Sharing big data among organizations is thought to boost its growth and bring superior benefits. However, there are serious security and privacy challenges due to the sensitivity of shared data. Sensitive data may contain personally identifiable information (PII) and confidential business information, making it attractive for misuse [4]. An entire big dataset needs to be security-protected when it is shared. Nevertheless, state-of-the-art techniques of data sharing seeking to protect data confidentiality generally consider data and its security protection separately. Indirect methods on security are employed, like restricting user access control mechanisms. This raises considerable concerns as there are numerous approaches for attacks on big datasets and analytical operations conducted on them. In the recent paradigm of BDA, there is a lack of adequate attention on integrity and confidentiality of shared data. Advanced security measures are required to protect big data from being misused after

being shared. A comprehensive overview of the threats and protection actions aspect is first presented. An innovative technique is then developed to enable BDA without knowing the shared (and sensitive) data. This approach exploits partially homomorphic Roth encryption and an additively homomorphic Paillier cryptosystem to provide both confidentiality and integrity protection on shared data [1].

2.1. Challenges in Secure Data Sharing

There are many obvious considerations and obstacles when attempting to securely share data. These considerations differ between local data stores and wide area networks (WANs) of data storage. Because external access through any WAN will either use the internet as a shared transport medium or utilize a private point-to-point link connection, security and privacy fears over storing data in the cloud revolve around how data can safely traverse the WAN and be stored outside of the control of the data owner. Many organizations have strict policies in place that prohibit storing any data on "the cloud." A common primary fear is the idea of data being stolen or compromised. Given very public data breaches at megacorporations like Adobe, J.P. Morgan, and Target, this fear is certainly warranted.

Additional concerns come into play when considering sharing sensitive data protected by laws and regulations like the Health Insurance Portability and Accountability Act (HIPAA), Federal Information Security Management Act (FISMA), and the Health Information Technology for Economic and Clinical Health (HITECH) Act. In light of these considerations, any solution involving sharing data with external organizations must be considered as a risk-reward trade-off, and the use of cloud storage only becomes feasible if this trade-off can be skewed heavily in favor of an organization [4]. At the core of these fears is the concern that a complete copy of the data being shared will be made and kept, a common issue with every externally serviced technology. In this circumstance, it is then very difficult to ensure that a given organization will have the ability to enforce a policy on who else can access the complete copy or how it can go through secondary sharing. The only options available to the data owner then are to trust the outside organization or to decide that the value of any data compromised is not worth taking the risk of sharing it.

2.2. Techniques for Secure Data Sharing

After the massive data collection phase comes the big data analytic phase, during which data is normally shared with knowledgeable researchers and/or analysts for several reasons [5]. On one hand, it is intended to generate knowledge and answer relevant analysis questions that add business value. On the other hand, it is often intended by data owners to comply with legal and regulatory obligations. During big data analytics, many invasive techniques such as data fusion, cross-correlation, and algorithm training are often conducted over shared data, which can lead to severe privacy leaks.

To mitigate the risk of privacy breaches, the most straightforward approach is data anonymization. Different data sanitization techniques such as k-anonymization, l-diversity, and t-closeness were proposed to limit privacy leaks to an acceptable level. However, with the recent emergence of de-anonymization attacks and the rapid evolution of big data analytic techniques, previous pseudonymization methods have become insufficient for individual privacy protection. Out of concern of both privacy leeks and data utility loss, the total prohibition of data sharing after such a massive data collection is not a viable solution either. Therefore, it is of utmost importance to find new reasonable approaches for privacy-aware big data analytics.

This unique need tends to urge a paradigm shift towards the development of new privacy protection protocols which are natively designed for privacy-aware big data analytics based on specific and widely adopted common methodologies. These protocols would preferably be based on new cryptographic primitives, and data would be encrypted and stored on the cloud after being transformed by users' private data sanitization "defense" approaches, enabling trustworthy data sharing and analysis between untrusting parties using the sanitized data.

3. ANONYMIZATION TECHNIQUES IN BIG DATA ANALYTICS

Anonymization is a series of techniques that transform data in order to prevent the discovery of an individual's identity. The individual or user is the subject of the analysis and holds the personal information. While the data can be exploited to assess or discover personal attributes, it is generally processed as an anonymous aggregate in order to prevent an individual's detection. Anonymization is envisaged to settle the dispute between privacy and data-driven innovation [2]. On the one hand, it allows data holders to share their data in a privacy-preserving manner. On the other, it allows data analysts to gain insights from such data, helping them improve services that are beneficial for society (e.g., optimizing public transport).

Cohen et al. introduce the generic notion of anonymity, formalizing different types of anonymity and how they relate to one another. Then, cover the research already conducted on anonymity in the context of several complex network types,

from social networks to peer-to-peer overlays. The applicability of these research works is discussed by describing current anonymization tools and libraries, as well as potential privacy risks in the applications that rely on complex networks (e.g., dissemination of content in peer-to-peer systems or social data in social network sites). Finally, the challenges and future directions that research and application of anonymous complex networks might pursue are presented.

3.1. Purpose of Anonymization in Data Analytics

Data analytics enables the utilization of large amounts of data for the purpose of deriving insights in a way that was previously not possible. It has been successfully applied in various fields ranging from marketing to fraud detection and healthcare [6]. While the use of data analytics offers a plethora of advantages, often unintentionally a lot of private and sensitive data about individuals is also processed.

Consequently, there has been an increasing awareness of the security and privacy implications of big data architectures and analytics [2]. Following the increase in data breaches involving corporations such as Capital One, Facebook, Equifax, and Marriott, as well as government leaks, there is increased public concern about what should be considered acceptable privacy. Technically, data belonging to an individual can either not be processed, thus remaining always protected, or data can be processed and privacy might be lost. The concern with privacy emerges in the realm in between the two extremes, where approaches and techniques for good privacy protection are required to enable processing of sensitive data.

The overarching purpose of employing anonymization techniques is to achieve a balance between preserving, as much as possible, the utility of the data while protecting the privacy of individuals. The utility of the data involves supporting the knowledge extraction task over the data set, which can range from the performance of a more accurate and precise predictive model to an overall higher quality insight over the data. On the other hand, preserving the privacy of the individuals may involve using anonymization techniques that make it practically impossible to link an individual to the data that describes him/her.

3.2. Common Anonymization Techniques

Big Data Analytics involves extremely large complex data sets which are very difficult to manage with current database technologies. The complexity is from the vastness of growing data streams that will not be accumulated in a traditional relational database. The data is sourced from heterogeneous systems, various formats and with the involvement of multiple types of transactions [7]. The big data is usually located in a distributed environment. The data veracity is regarding how reliable and trustworthy is the given data. These challenges are managed by technologies such as NoSQL databases, MapReduce and its implementations like Hadoop and Spark etc., and data virtualization solutions. Other analytics technologies are here to consume and evaluate the processed data and enhance business intelligence [6].

While Big Data Analytics generates enormous benefits for organizations, it brings issues with security and privacy as well. The collection and processing of huge amounts of personal data poses threats to the privacy of individuals. The right to privacy is put in danger by indiscriminate collection, mining, sharing or commercialization of sensitive information such as health records, financial transactions, geo-location data, and web browsing history. Big data can reveal individuals on a higher level than any other data mining technique. Consequently, large data sets can cause more harm if there is a breach because they can be used for re-identification and profiling purposes. Aggregate or anonymized data are not necessarily innocuous. The inferences or profiles that can be built based on these data and the services that can be offered are often so intrusive that they can lead to discrimination, victimization, tracking and harassment. The potential misuse of data can have catastrophic consequences for the individuals concerned.

4. PRIVACY-PRESERVING DATA MINING

As data-driven technologies become more mature, large SMEs and multinational enterprises increasingly adopt and implement Big Data-based systems. These systems collect, store, and analyze a diverse range of data, resulting in meaningful insights or value-added services. However, concerns over privacy from consumers and regulators are hindering the growth of Big Data technologies and services. This article explores these concerns, focusing on privacy-preserving data mining. For these tools to be accepted and widely adopted, simple, trustworthy, and non-prohibitive data processing must be established. As a by-product of data analysis, services also require transparent mechanisms for dealing with ownership and consumption rights of new assets emerging from it.

Inspired by advances in privacy-preserving computation, privacy-preserving data mining tools are proposed to facilitate a variety of data-driven applications. These are generic tools allowing to process and analyze data without revealing or exposing sensitive information. Privacy-preserving data mining refers to a broad class of problems, frameworks, and tools to extract knowledge from databases while preserving the privacy of the processed records. The privacy concern arises

during the extraction of valuable information from large amounts of data. From a technical perspective, privacy-preserving data mining includes several original results addressing different levels of privacy guarantees [8]. These include the definition of secure protocols for sophisticated data mining tasks and significant advances in the design of novel privacy-preserving data representation formats used to disclose statistics or mining results without compromising the underlying data.

4.1. Privacy Concerns in Data Mining

The role of data in the development of Big Data Analytics is examined with particular emphasis on its generation, processing, analysis, application, and storage. Potential security and privacy concerns that accompany a data mining process are also identified since the investigation of data mining actions may reveal sensitive information about individuals and organizations associated with the data [9]. As firms and government agencies alike are increasingly utilizing data mining processes to obtain insights that can benefit them, academic scholarship has similarly turned its attention to the implications of data mining [1]. Relying on recently published research, a number of widely-studied privacy concerns that accompany a data mining process are reviewed, such as data coarsening, value-chaining, linking, aggregation, and revealed presence. Furthermore, privacy-preserving alternatives to common data mining processes that mitigate these potential risks are also explored.

Big Data, comprising large datasets, is an emerging area for future research directions attend to in this study. One of the most crucial attributes of Big Data is not its huge size but the demand for real-time analytics applied to it. Data mining technologies are the only methods capable of analyzing such datasets with a high level of detail [2]. However, the performance of data mining actions on Big Data reveals privacy and security concerns that are substantially more sophisticated than those that arise in traditional data mining contexts, especially due to metadata presence.

4.2. Techniques for Privacy-Preserving Data Mining

The rapid growth of data analytics technologies, commonly referred to as the big data phenomenon, has transformed the way private and public organizations collect, store, and analyze personal data. Big data analytics has enabled organizations to collect and analyze vast volumes of personal data or metadata generated through smart devices and internet technologies. Such datasets offer opportunities to enhance productivity, increase revenue streams, and make organizations smarter. Nevertheless, this massive collection and sharing of data bring vulnerability and potential risks related to security and privacy issues. Consequently, the analysis of huge amounts of data to reveal personal or sensitive information about individuals has raised concerns regarding security and privacy, and the data analytics mechanisms used have come under scrutiny [8].

Preserving privacy while data mining is a crucial task, especially in large-scale datasets. A wide variety of data mining techniques have been proposed to live up to this challenge. Existing proposals can be categorized into three major trajectories. First, data mining itself can be transformed data-centrically in a way that attendant privacy is unobtrusively preserved. Several methods fall into this category. They either recover data or result in data that are associated with noise, distortion, or perturbation, such as in randomized methods. There are also techniques that transform the data to new data representation systems or spaces, e.g., through hashing, encryption, or projection. The second trajectory is that privacy-preserving data mining can be achieved through protocol-centric methods. In this case, privacy is upheld by specifications and arrangements regulating the conduction of data mining processes. Protocol-centric methods do not affect the overall accuracy of mining algorithms, and data can be shared and mined collectively while preserving privacy [10]. Finally, privacy-enhancing technologies (PET) can be employed to protect auxiliary datasets that are either generated or collected alongside the data at stake, such as in the case of telecom data.

5. PROTECTING SENSITIVE INFORMATION IN LARGE-SCALE DATA REPOSITORIES

With the proliferation of shared online resources and personal data, sensitive information could easily fall prey to cybercriminals seeking valuable data that can be misused. Cybercriminals extract sensitive information from large, accessible data repositories via malware, such as scripts or programs. This means that every enterprise, organization, and individual maintaining large data repositories are in danger of being breached. Just a few years back, enterprises had large databases containing customers' Personally Identifiable Information (PII), like Social Security Number (SSN), phone number, and addresses [1]. These pieces of information are highly valuable to malicious actors since they could be used for identity theft and financial fraud against unwitting victims.

Actively or passively, sensitive information will find its way into accessible databases on the internet. Large databases containing PII and sensitive information that has not been secured or encrypted have already been compromised.

Cybercriminals simply used scripts that continuously scanned for vulnerabilities in large data repositories, seeking these kinds of databases. Such attacks could be extremely disastrous for entities exposed in such breaches, dramatically affecting credibility and trust and potentially leading to major monetary losses or bankruptcy [11].

5.1. Types of Sensitive Information

In large-scale data repositories, the sensitive data can be in a variety of forms [3]. Data pertaining to one's thoughts, sentiments, emotional state, and behavioral patterns can be monitored either through direct readings of bodily states (biometrics) or by interpreting social media activities. Such data is so sensitive that if illicitly obtained, it can be abused in myriad harmful ways. Other highly sensitive information pertains to medical history, confidential records, and the social security numbers of individuals, which can all be used for illegal gains by scammers [2]. Therefore, the necessary protection measures to secure different types of sensitive data vary greatly and should be understood at an early stage.

5.2. Security Measures for Large-Scale Data Repositories

Securing sensitive data within large-scale volume repositories needs careful consideration of the effective mechanisms and protocols necessary for their protection. Sensitive information can be diverse, focusing on different types of personal or corporate aspects, requirements, and content-oriented domains. Thus, such requirements, on one hand, increase the complexity of security design processes, and on the other hand, turn the potential loss of sensitive information into attractive targets. Therefore, it is complicated to guarantee the security and privacy of such diverse types of information on para with other single type or less diverse data repositories [11]. Additionally, real-time and near-real-time volume of this data needs to be continuously processed, maintained, exposed, or updated by hackers. As a result, potential attackers can be both rogue employees or outsiders with no knowledge of internal operations. Batch and periodic update processes are less effective in combating these recently emerging threats.

Thus, suitable building guidelines for the design of such protection are first presented below, followed by the discussion of recently emerging large-scale data protection mechanisms based on them. As part of the approaches, typically loss prevention techniques are preferred, commonly implemented by encryption, watermarking, or both. Loss prevention approaches secure sensitive information from illegal possession or unauthorized alteration, assuring its integrity. Loss prevention mechanisms counter external threats, and hence both the internal and external attackers have to be above trust level [4]. Therefore, inside attack should be the most privileged type of attack and should be treated separately, leading to special and additional types of losses to information or processes stored within and covering sensitive information.

6. LEGAL AND ETHICAL CONSIDERATIONS IN BIG DATA ANALYTICS

In the big data analytics industry, the ethical commitment levels range from recognizing the importance of ethical considerations (Honest commitment) to establishing a formalized ethical review committee and process within the organization (Proactive commitment) [1]. There are also several ethical frameworks in the literature, addressing data protection, regulations, and guidelines. While some primarily note legal developments and recommendations outside the field of big data analytics [2] , others present ethical guidelines and tools that specifically target the technologies and processes involved in big data analytics.

6.1. Data Protection Regulations

In recent years, Big Data Analytics has gained immense popularity and is being adopted in various areas, such as political campaigns, marketing, health care, and social media, to derive valuable insights from the vast volume of data [1]. However, the existing data protection regulations are not adequate to deal with many of the data privacy concerns arising due to Big Data Analytics, as there is a high possibility of identification and inference of a person's information using unrelated datasets [2]. The necessity and requirements for data protection regulations in the context of Big Data Analytics are explored in the following sections.

Data Protection Regulations. Most of the countries following a civilized approach have written data protection regulations comprising of rules and rights related to the protection of individuals' data, which must be followed by organizations and government bodies. The data protection regulations are enforced in order to safeguard the anonymity and privacy of the individuals, as well as the organizations, regarding the use of their data by other organizations or bodies. Further, the data protection regulations state the rights of the individuals or organizations regarding their data such as the right to object to or refuse the processing of data. In the case of countries following a non-civilized approach, individuals or organizations usually adopt their own set of mutually agreed rules regarding the use of their sensitive data, sometimes without having

access to the contracted rules. This absence of data protection regulations keeps the individuals or organizations in a dark state concerning the processing of their data.

6.2. Ethical Guidelines for Data Use

Transparency: People should be aware of their data being collected and how it is used, shared, and monetized. Furthermore, it should be understood that data will be used to provide more personalized services and potentially for behavioral profiling that includes a prediction of activities and beliefs [1]. Advertisements and offers will be determined based on an understanding of consumers' interests, beliefs, and behavior due to the predictive behavioral analytics. Although concerning, the latter will not go beyond the normal scope of marketing. Still there are concerns about the data being used for social exclusion purposes using data-driven automation of decision making. There is a concern that unintended consequences may arise; for example, it is possible that because of being labeled as a possible fraudster by the bank, the consumer may not even be able to open an account with other banks. Hence the ability to justify and explain the decision of not granting an account should be a necessity.
Consent: There is a need for informed and explicit consent to collect and utilize data, proceeding from the position that data contain PII. Robust explicit consent would involve presenting data collectors with an understandable, comprehensive, unambiguous, and actionable option to agree for non-specific future assignments. The Tempus project advocates the deidentification of PII before its collection and utilization [12]. Further relevant consent measures could include the right not to be subjected to an automated decision-making process, unless the data range is not personal, as articulated in Art. 15 of GDPR, or if the consumer has explicitly consented, as articulated in Art. 22 of GDPR [13].
Data Access and Control: It must be ensured that consumers have access to their data and utilization decision at any time. The consumer should have the ability to understand what is stored, how data is being processed, and have an opportunity to articulate non-specific future authorizations. Furthermore, data has to be deidentified before monetization, and if consumers do not want their data to be used, their data must be removed from the databases.

7. CASE STUDIES AND REAL-WORLD EXAMPLES

Even as the usage of big data analytics is pervasive, there remains ambiguity either implicitly or explicitly within social, ethical, and political contexts [3]. This is true of commonplace user-facing services engaged in big data analytics. Examples include Internet advertisements, credit scoring, and insurance risk profiling, which are widely popularized and consequently contested. Though these are recent developments, there is a long history of suspicion surrounding the use of algorithmic profiling, surveillance, justification, and control. Other controversies arrive as moral panics surrounding emergent big data technologies, including major events such as the Snowden surveillance disclosures and ongoing worries about the social credit system in China [1]. The interaction of technological, political, social, economic, and cultural developments is ever complex. Nevertheless, widely used technologies such as big data analytics tend to raise similar social assumptions at a meta level.
Adverse social consequences may occur with or without active malfeasance, often taking on social forms unanticipated by designers or developers. With the rise of several political socioeconomic developments characterizing neo-liberalism, early efforts in securing data sharing have sought to engage a wide array of players such as government, industry, and civil society (non-governmental organizations). Operating in overlapping yet disparate silos, there is a tendency for emphases on competing contextual values focused on either risk and neglect, or conspiracy and skepticism. Consequently, there are possible tensions and discord between normative and social expectations leading to unrealized benefits and risks under contested or unsure prices, affecting stakeholder participation. Current implementations or developments of big data technologically necessitate a greater understanding of the significance of social values in shaping the course of emerging technologies in a more authentically democratic and accountable manner.

7.1. Successful Implementations of Secure Data Sharing

One case that will become vivid in the reader's mind is that of two American telecommunications companies who were locked in competition for a long time. After receiving a lucrative contract from the US government for the deployment of the next-generation telecommunications network, the two companies became interested in trading historical customer call detail records. This would level the playing field by allowing both companies to develop accurate models for optimizing network resource allocations. However, local laws limited the details of such a deal and only anonymized data could be exchanged. The full 9 billion records dataset was released to the other party, who then unveiled sensitive information such as social networks, customer names, and engineering plans. This very unfortunate security failure could have been avoided had the proper precautions been taken. Companies, organizations, and public institutions are all involved in such data-

related activities and models. In some cases, these entities wish to keep the results of their analysis to themselves, guarding them from competitors. In other cases, it is the raw data that must remain confidential, such as data containing sensitive personal information, trade secrets, or usage information linked to confidential processes.
Research Projects Underway. Anonymization and secure multiparty computation research efforts are upgrading computer scientist tools for preventing data-sharing security breaches. In the anonymization area, the expert is developing a broader range of new algorithms and applying them to large corporate datasets. Exploring underlying anonymization models and developing simpler high-level interfaces are also part of the research agenda. In the secure multiparty computation area, the long-term objective is to develop efficient systems that enable any of several mutually distrustful parties sharing data to run computations without revealing details about the data to the rival parties.

7.2. Privacy Breaches in Big Data Analytics

Given the plethora of digital data and the rapid growth of internet traffic, especially online social networks (OSNs), a large amount of personal information is readily accessible. People tend to share everything with their friends via social networks, which has negative implications in terms of social relationships and privacy risks. These posted status updates, pictures, and videos can be left unnoticed by a huge number of other people in the network, which can be catastrophic. As a recent example, a social public figure posted a status update and a picture on a popular OSN in order to celebrate the birth of his baby. Nevertheless, the location metadata was still exposed because these posted contents were taken in a hospital room, which exposes the baby to unwanted terror [16].
Privacy is normally referred to as controlling the access of others to personal information. Recent developments in Big Data has enabled individuals and institutions to gather an unprecedented amount of data about people's behavior. With the advancement of data storage and processing technologies, these collected data can then be mined to uncover insights about individuals and their social groups. This gives rise to a new set of privacy concerns that are different and often larger in scale from the traditional concerns. In the traditional view of information privacy, privacy concerns arise only from the observation of an individual's behavior. In contrast, the Big Data view of privacy concerns can also arise from behavior inferences. This inferring process is done on the aggregate-level data and hence is independent of either direct observation or knowledge of the individuals. Nevertheless, such assumptions of anonymity tend not to hold in practice, especially when the data gathered about an individual is extensive [1].

8. FUTURE TRENDS AND EMERGING TECHNOLOGIES

Over the last few years, there has been a growing societal concern about the security and privacy implications of Big Data. Unfortunately, predicted long-term solutions, such as the control over that data by its originators, and the collection processes used, have yet to be effectively addressed [1]. But it is felt that short and medium term solutions could be implemented. Algorithms need to be reviewed and modified to, at a minimum, obfuscate personally identifiable information and limit the granularity of collected data. Aggregation processes that build upon social averages rather than individuals, whether people or objects, should be implemented. Hopelessly unregulated collection processes would be assisted through the implementation of more global and stringent regulation, such as the General Data Protection Regulation established by the European Union [2]. Unfortunately, there are vast, and largely impervious, transnational gaps in regulatory authority and enforcement. In the longer term, there is a need for more academic research into the security and privacy challenges posed by systems that operate on Big Data of individuals. In addition, as Big Data is a fundamental component of Artificial Intelligence, there is a growing need for academic research into the fundamental societal ethics of Artificial Intelligence. As understanding of this technology grows, a proactive rather than re-active approach may be possible.

8.1. Advancements in Privacy-Preserving Techniques

One of the major concerns of data users is that the data they are provided with have been properly anonymized or, in other words, the privacy of the individuals that they are based on is sufficiently guaranteed before the data is published. The term "data anonymization" is broadly used to refer to an array of technologies with the aim of transforming a database maintaining personal or sensitive data such that the risk of disclosing the identity and confidential information of the data subjects is reduced below a certain level. The most common privacy protection method is anonymization. With this method, any personal identifiers included in the data are removed and the data is published as a result [1]. However, based on the assumption that some sensitive or personal data is published, interested readers of a database may be able to reconstruct the identities of the individuals in it or to infer sensitive information of them.
There is a growing interest in data publishing methodologies focusing on the design of anonymization algorithms that protect the privacy of the individuals whose profiles are maintained in released databases. These algorithms sanitize the

data before publishing it by altering or removing certain values in such a way that sensitive information about individuals cannot be inferred while the usability of the data for statistical queries is preserved. The first framework proposed in this context is k-anonymity [14]. A data table is said to satisfy k-anonymity with respect to a set Q of quasi-identifiers if each tuple in the table is indistinguishable from at least k - 1 other tuples in Q in the sense that they have the same values for the attributes in Q. In other words, the shared characteristics represented by the values contained in the quasi-identifiers attribute of these tuples ensure a certain level of anonymity to the individuals related to them. Nevertheless and despite its widespread acceptance, k-anonymity turned out to have several limitations [15]. For instance, cm-inference attacks may lead to the identification or re-identification of individuals having particular sensitive attribute values even in k-anonymous databases [2].

8.2. Innovations in Secure Data Sharing

In the light of the Big Data solutions in use, some innovations are introduced as solutions to the problems surrounding data sharing. These solutions are the advanced technologies in which the needs and requirements for data improvement have been covered [1]. On suddenly compromising personal data against the sharing advantages, these simple new systems can first be used against the most efficient data sharing processes [4]. Following the changes of the currently used Big Data solutions into sharing conditions, the focus should be put on the physical placement of the data. In the currently in use databases, data in Big Data is usually stored in clusters/sets. In case of single clusters, this means that personal data will be located in one single physical location. These data locations could be full companies/countries facts against personal rules, if certain data sets get physically to certain countries (i.e., banking data to the USA against to be transferred there usage in Europe). The optimal physical data placement change resulting in a global sharing of the data across several clusters, countries, or companies ends with efficiency exchanges on advantage of low cost Big Data processing.

9. CONCLUSION AND RECOMMENDATIONS

Today, given both the increasing importance and omnipresence of big data and its analytics, it is crucial not only to stress the security and privacy requirements raised by big data analytics and its results, but also to proactively integrate these requirements into the entire design process of big data systems. Our study teaches us that security, privacy, and ethical concerns in big data analytics do not exist in parallel to the business cycle, but must be wisely and ethically managed in coherence throughout all emerging processes of the big data and information systems. Tailoring the level and scale of legal and ethical responsibility of the data controller, processor, and decision maker are among the most significant constructs that must be taken into consideration in the design of these systems.

Thus, we recommend that currently, authorities in anywhere in the world should seriously reconsider using and utilizing the newly developed set of principles, which in many parts, this paper suggests an outline. We successfully suggested a framework for evaluating and analyzing policy considerations when deciding on whether or not to anonymize big data in connection with government administration. Based on this framework, the current understanding is that data anonymity can be a potentially powerful tool in ensuring privacy and cybersecurity and should therefore be developed and adopted as part of a local set of policy considerations that governments should take into account in the holistic design of their administration. And so, consideration of it as an independent and new level for inclusion, hybrid is necessary in all papers relating to the security and privacy of big data..

Conflicts Of Interest

The author's paper explicitly states that there are no conflicts of interest to be disclosed.

Funding

The absence of acknowledgments or thank you notes to institutions or sponsors in the paper suggests no financial support was received.

Acknowledgment

The author would like to thank the administrative staff at the institution for their assistance and logistical support throughout the duration of this research.

References

[1] M. Dabab, R. Craven, H. Barham, and E. Gibson, "Exploratory Strategic Roadmapping Framework for Big Data Privacy Issues," 2018. [PDF]
[2] N. Dreyer, "Big data, Bigger privacy concern?," 2017. [PDF]
[3] B. Alabdullah, N. Beloff, and M. White, "Rise of big data – issues and challenges," 2018. [PDF]
[4] Z. HUO, H. HE, and R. WANG, "Personal Privacy Security Management in the Era of Big Data," 2015. [PDF]
[5] D. Thilakanathan, "Secure Data Sharing and Collaboration in the Cloud," 2016. [PDF]
[6] I. E. Olatunji, J. Rauch, M. Katzensteiner, and M. Khosla, "A Review of Anonymization for Healthcare Data," 2021. [PDF]
[7] D. W. Archer, B. de Balle Pigem, D. Bogdanov, M. Craddock et al., "UN Handbook on Privacy-Preserving Computation Techniques," 2023. [PDF]
[8] R. Onyemechi Oturugbum, "Preserving The Safety And Confidentiality Of Data Mining Information In Health Care: A literature review," 2023. [PDF]
[9] SS Sundar, J Kim, MB Rosson, MD Molina, "Online privacy heuristics that predict information disclosure," in Proceedings of the 2020, 2020, dl.acm.org. acm.org
[10] M. R. Keyvanpour and S. Seifi Moradi, "Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification-based Framework," 2011. [PDF]
[11] O. Soliman, "Big Data SAVE: Secure Anonymous Vault Environment," 2019. [PDF]
[12] S. Garmash, "Experience of participation in international projects in the aspect of interconnected processes (problems and prospects)–significance of the human factor," 2023. kpi.kharkov.ua
[13] A. Peralta, "Best Practices for Managing Privacy and Security of Cloud-Based Student Data in Primary and Secondary Education," 2024. adelphi.edu
[14] W. Mahanan, W. A. Chaovalitwongse, and J. Natwichai, "Data privacy preservation algorithm with k-anonymity," World Wide Web, 2021. springer.com
[15] V. Torra and G. Navarro-Arribas, "Attribute disclosure risk for k-anonymity: the case of numerical data," International Journal of Information Security, 2023. springer.com
[16] N. Andalibi, "Disclosure, privacy, and stigma on social media: Examining non-disclosure of distressing experiences," ACM Transactions on Computer-Human Interaction, 2020. acm.org