Privacy and Security of Big Data in THE Cloud

Narges Naderi
Islamic Azad University
E-Campus

Hasan Alizadeh
Islamic Azad University
E-Campus

Abstract:
Big data has been arising a growing interest in both scientific and industrial fields for its potential value. However, before employing big data technology into massive applications, a basic but also principle topic should be investigated: security and privacy. One of the biggest concerns of big data is privacy. However, the study on big data privacy is still at a very early stage. Many organizations demand efficient solutions to store and analyze huge amount of information. Cloud computing as an enabler provides scalable resources and significant economic benefits in the form of reduced operational costs. This paradigm raises a broad range of security and privacy issues that must be taken into consideration. Multi-tenancy, loss of control, and trust are key challenges in cloud computing environments. In this paper, the recent research and development on security and privacy in big data is surveyed and reviews the existing technologies and a wide array of both earlier and state-of- the-art projects on cloud security and privacy. First, the effects of characteristics of big data on information security and privacy are described. Second, topics and issues on security are discussed and reviewed. Third, privacy-preserving trajectory data publishing is studied due to its future utilization, especially in telecom operation. Forth, present an overview of the battle ground by defining the roles and operations of privacy systems. Fifth, we review the milestones of the current two major research categories of privacy: data clustering and privacy frameworks. Finally, we discuss the effort of privacy study from the perspectives of different disciplines, respectively.

I. INTRODUCTION

Big data has emerged to a new paradigm for data applications. Due to significant benefits, big data arises a growing interest in many industry fields, such as telecom operation [1], [2], healthcare [3], [4] and so on. Many efforts on big data have been working on the data storage, data mining, and data application. However, the widespread usage of big data relies not only on the promising solutions and mechanisms of data analysing, but also on security protection and privacy preserving.

Information security can be improved by big data technology, which is beneficial from security tools such as network monitoring, security information, and event management [5], [6]. However, on the down-side, there are additional security challenges brought by the big data technology, including cryptography algorithms, data provenance, secure data storage, access control, real time monitoring and so on [7]. Identifying and analysing the security issues will bring a better usage of big data. Thus, in this paper, we will first survey existing research on security and privacy. Then, we will focus on an essential type of data: trajectory.

Trajectory data represents the mobility of moving objects, such as people, vehicles, and so on. Spatio-temporal trajectories provide significant and valuable information, and foster a broad range of applications, such as intelligent transportation system, commercial site planning and so on. Therefore, trajectory data mining has become an increasingly interesting research topic, attracting attentions from numerous fields, especially in telecom operation. As a trustful data owner, telecom operators are authorized to have large amounts of location data of mobile phone customers. Making a good use of trajectory data can help telecom operators optimize the network and promote social services as well. However, location and trajectory data can be sensitive for individuals. Attackers may infer individuals' privacy such as personal habits or personal details from trajectories. In order to preserve privacy, techniques and algorithms should be applied in the case that trajectories are released to third party for data analysis or data mining results of trajectories are published.

During recent years, data production rate has been growing exponentially [11,12]. Many organizations demand efficient solutions to store and analyze these big amount data that are preliminary generated from various sources such as high throughput instruments, sensors or connected devices. For this purpose, big data technologies can utilize cloud computing to provide significant benefits, such as the availability of automated tools to assemble, connect, configure and reconfigure virtualized resources on demand. These make it much easier to meet organizational goals as organizations can easily deploy cloud services.

However, privacy protection has become one of the biggest problems with the progress of big data. Human privacy is usually challenged by the development of technology. The record of individuals for tax and draft purpose was a great threat to personal privacy in the 11th century in England, and photographs and yellow page services significantly threatened people's privacy in the late 19th century.

The main objective of this paper is to survey the literature related to security and privacy in big data to provide a comprehensive reference of the challenges and risks to which a big data application chain is facing. Then, we focus on the research and industry approaches to trajectory preserving issues in big data.

In this work, we provide a context to the work by introducing the security and privacy challenges triggered by characteristics of big data in Section II. In Section III, we present big data system security and privacy analyses. Research work on trajectory publishing is then highlighted in Section IV. In Section V, gives an overview of big data, cloud computing concepts and technologies. Section VI, reviews the existing security solutions that are being used in the area of cloud computing. Section VII describes research on privacy-preserving solutions for big sensitive data.

We discuss the different roles and operations of privacy systems in Section VIII. The major developments of modern privacy study are presented in Section IX. In Section X, we survey the privacy study from different disciplines. Finally, we summarize the paper in Section XI.

II. Security and Privacy Challenges Triggered by 5Vs.

In this section, the discussion begins with understanding the impacts of big data characteristics on security and privacy. According to the definition and principle of big data [9], the characteristics of big data are summarized as "5Vs", i.e., Volume, Variety, Velocity, Value, and Veracity:

1) "Volume" points to the size of data. There is a huge amount of data generated by organizations, individuals and sensors every second in every fields. It is nearly impossible for data providers to supervise or control all the data they "actively" or "passively" provide to others[13].
2) "Variety" indicates the diversity of data formats and sources. The data format includes structured, semi-structured, and unstructured ones, while the filetype consists of texts, figures, and videos[14].
3) "Velocity" shows the continuousness and high frequencies of data. This feature makes information security and privacy issues even more severe[15].
4) "Value" refers to the outputs that gains from huge data sets. The highly potential value and intensely integrated data attracts hackers [16].
5) "Veracity" refers to the trustworthiness, applicability, noise, bias, abnormality and other quality properties of the data [9].

III. Security and Privacy Issues in Big Data

Given the big data characteristics and the impacts triggered by the characteristics on security and privacy in section II, existing security and privacy topics and issues are discussed and surveyed in this section.

The authors of [7], [17] have proposed some conceptual and operational taxonomies of security and privacy to introduce vulnerabilities of big data system:

- Infrastructure Security
- Data Privacy
- Data Management

In the architecture dimension, the research objective is considered as a big data management platform. There are four layers, to which physical or logical entities correspond: storage layer, including secure storing, distributed equipment, monitoring and control etc. For each type of user, the concerns and the methods adopted for each user role can be diverse [18]. In the data value chain dimension, it considers the stages in the process to obtain data value. For each stage, the risks, requirements as well as methods are different [19]. In the industrial fields dimension, [20] analyses the most concerning security issues in different application area.

At the beginning of the research, surveys focus on analyse vulnerabilities and risks, which are brought or increased in big data era [10],[13],[16],[18]. However, the more recent work also analyse more specific technologies or mechanisms to enhance the security [21]-[22],[23], [24]. The security and privacy analyses are comparatively presented in Table I in terms of their consideration of topics and big data process layer/interfaces discussed.

In [18], authors first illustrate the privacy-preserving issues of four types of users in data life cycle from their unique perspectives. Then, they mainly focus on how privacy-preserving data publishing(PPDP) is realized in two emerging applications, i.e., social networks and location-based services. Rather than portraying the whole application chain, some researches focus on depicting typical issues. In [16], authors discuss big data security management platform, information security system and relevant laws and regulations. In [10], authors present related research work on five subjective: Hadoop security, cloud security, monitoring and auditing, key management and anoymization.

Data collecting and storing are essential phases in big data applications. The vulnerabilities and enhancements in secure collection and storage have been presented in [21], [22]. [21] discusses the security issue on NoSQL databases; several database products of four types of database, such as key-value database, column-oriented database, document based database and group database, are studied and their merits and weakness are revealed. Access control models are compared as well. [22] not only considers the privacy and data confidentiality in big data, but data provenance and data trustworthiness are also taken into account.

Data mining is the principal process of discovering knowledge. However, since data mining enables efficiently discover valuable, non-obvious information from large volumes of data, it may result in an extraction of sensitive information. [23] and [19] provide reviews on privacy-preserving data mining techniques and analyse those methods.The transformation methods on the original data in order to preserve privacy is classified as randomization methods, anoymization and distributed methods in [23]. The analytics and comparison of privacy preserving in clustering and association rule mining are given as well. In [19], privacy-preserving techniques in data mining, including privacy-preserving aggregation, operations over encrypted data, and de-identification, are reviewed.

After data mining and analyzing, basic operations include showing the interesting mining results in proper ways. Linkage of private information is the one of the top security risks in this stage. Thus, PPDP is required to publish useful information while preserving data privacy. Fung et al. systematically summarize and evaluate different approaches in their frequently cited survey [25]. Rashid et al. study the PPDM and PPDP, and present the differences and requirements between PPDP and other related problems [26].

IV. Privacy-Preserving Trajectory publishing

In the Section III, the surveys and achievements of security and privacy in big data are reviewed. In this section, we focus on privacy-preserving trajectory publishing techniques in big data.

A. What is Trajectory?

Advancement of wireless communication enables a large number of location based applications and services, along with a massive collection of location information. It is obvious that sharing location information can help improve users' quality of lives, while on the other hand, it may reveal sensitive and private information about individuals.

Compared with single location, a trajectory is an entire set of discrete location samples. There are two major scenarios that we need to protect a user's trajectory data from the privacy leak. One is in on-line location-based services. In this scenario, a user may not want to exactly disclose his or her current location when using a service [27]. The other is the off-line historical trajectories. Based on a collection of trajectories, an adversary may discover an individual's most frequent places, and therefore identify the individual, or even infer sensitive personal information like health condition, religious and sexual preferences [28]. In this paper, we concentrate on the latter.

Trajectory is usually samples of a mobile object's true movements. For the purpose of data analysis, approaches of reducing the uncertainty of a trajectory are studied [29]. On the other hand, to protect a user from the privacy leak caused by the disclosure of the user's trajectories, a trajectory should be even more uncertain.

B. Mechanisms in Privacy-Preserving Trajectory Publishing

Anonymization technique is an efficient method to realize privacy-preserving, and it can be also utilized for trajectory data set. However, spatio-temporal data, different from relational data, have some unique features, including time dependence, location dependence and high dimensionality [18]. Thus, tailored privacy-preserving methods should be considered. In this subsection, three common mechanisms in privacy-preserving trajectory publishing are presented, and existing research under each class are reviewed.

Generalization and Suppression

Generalization and suppression are the most common anonymity operations used to implement fc-anonymity. Generalization means replacing one or multiple specific values with a more general one. Suppression involves deleting values or records of data. A number of solutions provide anonymization protection based on generalization and suppression.

In [30], Terrovitis et al. proposed an anonymization algorithm that iteratively suppresses selected locations from the original trajectories, taking into consider of the benefit in terms of privacy and the deviation from the main direction of the trajectory.

The authors of [31] consider the trajectories as a collection of points, each point represented by intervals on the three dimensions. Then, fc-anonymity model is built.

Yarovoy studies the case that each user has a different set of quasi-identifier(QID)(location,time) pairs for which he or she requires protection [32]. Based on the graph theory, the authors build the fc-anonymity model.

Generalization and suppression operations are feasible and easy. However, the main negative side is that replacing or deleting real values leads to a high possibility of information loss.

Perturbation

While data semantics are retained at a record level by the generalization and suppression mechanisms, the perturbation techniques retain data semantics at an aggregate level [33]. Perturbation techniques usually are based on randomization. Adding noise and swapping data are common means of perturbation [34], [35]. Some studies based on perturbation have been developed.

Abul et al. consider the problem of publishing a complete sequence of individuals' trajectory [36], [37]. In [36], a *(fc*, *6)*- anonymity by space translation is proposed to preserve the individuals' privacy. Then, there are *fc* different trajectories coexisting in a cylinder of the radius *6*. It has been later improved by removing some constraints about the input datasets and scales to large datasets at the cost of higher computational requirements [37].

The main idea of [38] is also adding noise, however, instead of spatial distortion, the scheme proposed is built on time distortion. Promesse is designed to smooth the users' speed from original data to a constant speed, and then blur endpoints at the same time Hence, the users' interests spots and endpoint are preserved.

In [39], a *(K,* t)-privacy metric based on the idea of swapping data, is proposed. The algorithm is designed to exchange multiple users' pseudonyms only when they meet the same location, so as to eliminate the linkability of their pseudonyms before and after the exchange. This algorithm can be used in the scenario that many people move through hub locations, such as a train station.

However, schemes based on the partition-based privacy model, including generalization, suppression and perturbation, have been found to be vulnerable to many types of privacy attacks, such as composition attack, deFinetti attack, and foreground knowledge attack [40].

Differential Privacy

Differential privacy is recently introduced to privacy preserving data publishing. The privacy preserving model is designed to ensure an equal probability of any released data among all nearly identical input data sets, and due to this reason, it guarantees that all outputs are insensitive to individuals. Adding random noise to the true output of the function is a common method. The Laplace mechanism [41] and the Exponential mechanism [42] are two major techniques. For real outputs, the Laplace mechanism is used that the noises are generated based on Laplace distribution. When outputs are not real, the Exponential mechanism assigns exponentially greater probability to a output with a higher score, with which it is more likely to be selected. As a consequence, the final output would be close to the optimum with respect to utility function. Some research have been working on differential private publication of trajectory data.

Chen et al. first introduce differential privacy to trajectory publishing [43]. A non-interactive data-dependent sanitization algorithm is proposed. The efficiency of the approach is guaranteed by narrowing down the output domain by constructing a noisy prefix tree under Laplace mechanism. Then Chen et al. develop techniques making use of the inherent Markov assumption in the variable-length n-gram model in order to improve the utility [44].

Considering that there are not many common prefixes or n-grams of raw trajectories, Hua et al. propose location generalization algorithm based on the exponential mechanism for preparation; then design authors design a release algorithm which leverages a noise counting scheme based on Laplace mechanism [40].

V. Key Concepts and Technologies

While it is practical and cost effective to use cloud computing for data-intensive applications, there can be issues with security when using systems that are not provided inhouse. To look into these and find appropriate solutions, there are several key concepts and technologies that are widely used in data-intensive clouds that need to be understood, such as big data infrastructures, virtualization mechanisms, varieties of cloud services, and "container" technologies.

1. Big Data

Computers produce soaring rates of data that is primarily generated by Internet of Things (IoT), Next-Generation Sequencing (NGS) machines, scientific simulations and other sources of data which demand efficient architectures for handling the new datasets. In order to cope with this huge amount of information, "Big Data" solutions such as the Google File System (GFS) [45], Map/Reduce (MR), Apache Hadoop and the Hadoop Distributed File System (HDFS) have been proposed both as commercial or open-source.

During the past few years, NIST formed the big data working group as a community with joint members from industry, academia and government with the aim of developing a consensus definition, taxonomies, secure reference architectures, and technology roadmap. It identifies big data characteristics as extensive datasets that are diverse, including structured, semi-structured, and unstructured data from different domains (variety); large orders of magnitude (volume); arriving with fast rate (velocity); change in other characteristics (variability) [46]. Big data analytics can benefit enterprises and organizations by solving many problems in manufacturing, education, telecommunication, insurance, government, energy, retail, transportation, and healthcare [47].

2. Virtualization Mechanisms

A hypervisor or virtual machine monitor (VMM) is a key component that resides between VMs and hardware to control the virtualized resource [48]. It provides the means to run several isolated virtual machines on the same physical host. Hypervisors can be categorized into two groups [49] as follows.

Type I: Here the hypervisor runs directly on the real system hardware, and there is no operating system (OS) under it. This approach is efficient as it eliminates any intermediary layers. Another benefit with this type of hypervisor is that security levels can be improved by isolating the guest VMs.

Type II: The second type of hypervisor runs on a hosted OS that provides virtualization services, such as input/output (IO) device support and memory management.

3. Cloud Computing Characteristics

When considering cloud computing, we need to be aware of the types of services that are offered, the way those services are delivered to those using the services, and the different types of people and groups that are involved with cloud services.

Cloud computing delivers computing software, platforms and infrastructures as services based on pay-as-you go models. Cloud service models can be deployed for on-demand storage and computing power in various ways: **Software-as-a-Service (SaaS),Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS).** Cloud computing service models have been evolved during the past few years within a variety of domains using the **"as-a-Service"** concept of cloud computing such as **Business Integration-as-a-Service,Cloud-Based Analytics-as-a-Service(CLAaaS),Data-as-a-Service (DaaS)** [50, 51]. This paper refers to the NIST cloud service models features [52] that are summarized in Table 2 that can be delivered to consumers using different models such as a private cloud, community cloud, public cloud, or hybrid cloud.

The NIST cloud computing reference architecture [53], defines five major actors in the cloud arena: cloud consumers, cloud providers, cloud carriers, cloud auditors and cloud brokers.

4. Container Technology

Clouds based on Linux container (LXC) technology are considered to be next-generation clouds, so LXCs has become an important part of the cloud computing infrastructures because of their ability to run several OS-level isolated VMs within a host with a very low overhead. LXCs are built on modern kernel features. An LXC resembles a light-weight execution environment within a host system that runs instructions native to the core CPU while eliminating the need for instruction level emulation or just-in-time compilation [54]. LXCs contain applications, configurations and the required storage dependencies, in a manner similar to the just enough OS (JeOS). Containers are built on the hardware and OS but they make use of kernel features called chroots, cgroups and namespaces to construct a contained environment without the need for a hypervisor. The most recent container technologies are Solaris Zones, OpenVZ and LXC.

VI. The Cloud Security Solutions

This section reviews the research on security solution such as authentication, authorization, and identity management that were identified in Table 3 [56] as being necessary so that the activities of cloud providers are sufficiently secure.

1. Authentication and Authorization

In [57] the authors propose a credential classification and a framework for analyzing and developing solutions for credential management that include strategies to evaluate the complexity of cloud ecosystems. This study identifies a set of categories relevant for authentication and authorization for the cloud focusing on infrastructural organization which include classifications for credentials, and adapt those categories to the cloud context.

2. Identity and Access Management

The important functionalities of identity management systems for the success of clouds in relation to consumer satisfaction is discussed in [58]. The authors also present an authorization system for cloud federation using Shibboleth - an open source implementation of the security assertion markup language (SAML) for single sign-on with different cloud providers. This solution demonstrates how organizations can outsource authentication and authorization to third-party clouds using an identity management system. Stihler et al. [59] also propose an integral federated identity management for cloud computing. A trust relationship between a given user and SaaS domains is required so that SaaS users can access the application and resources that are provided. In a PaaS domain, there is an interceptor that acts as a proxy to accept the user's requests and execute them. The interceptor interacts with the secure token service (STS), and requests the security token using the WS-Trust specification.

3. Confidentiality, Integrity, and Availability (CIA)

Santos et al. [60] extend the Terra [61] design that enables users to verify the integrity of VMs in the cloud. The proposed solution is called the trusted cloud computing platform (TCCP), and the whole IaaS is considered to be a single system instead of granular hosts in Terra. In this approach, all nodes run a trusted virtual machine monitor to isolate and protect virtual machines. Users are given access to cloud services through the cloud manager component. The external trusted entity (ETE) is another component that provides a trust coordinator service in order to keep track of the trusted VMs in a cluster. The ETE can be used to attest the security of the VMs.

4. Security Monitoring and Incident Response

Anand [62] presents a centralized monitoring solution for cloud applications consisting of monitoring the server, monitors, agents, configuration files and notification components. Redundancy, automatic healing, and multi-level notifications are other benefits of the proposed solution which are designed to avoid the typical drawbacks of a centralized monitoring system, such as limited scalability, low performance and single point of failure.

Brinkmann et al. [63] present a scalable distributed monitoring system for clouds using a distributed management tree that covers all the protocol-specific parameters for data collection. Data acquisition is done through specific handler implementations for each infrastructure-level data supplier.

Hypervisor-based cloud intrusion detection systems are a new approach (compared to existing host-based and network-based intrusion detection systems) that is discussed in [64]. The idea is to use hypervisor capabilities to improve performance over data residing in a VM. Performance metrics are defined as networking transmitted and received data, read/write over data blocks, and CPU utilization.

5. Security Policy Management

In [65] the authors propose a generic security management framework allowing providers of cloud data management systems to define and enforce complex security policies through a policy management module. The user activities are stored and monitored for each storage system, and are made available to the policy management module. Users' actions are evaluated by a trust management module based on their past activities and are grouped as "fair" or "malicious".

VII. Big Data Security and Privacy

This section outlines several efforts and projects on big data security and privacy including big data infrastructures and programming models. It focuses on the Apache Hadoop that is a widely- used infrastructure for big data projects such as HDFS and Hive, HBase, Flume, Pig, Kafka, and Storm. We also summarize the state-of-the-art for privacy-preserving data-insensitive solutions in cloud computing environments.

VIII. PRELIMINARY OF PRIVACY STUDY

In this section, we present an overview of privacy systems, including different participation roles, anonymization operations, and data status. We also introduce the terms and definitions of the system.

In terms of participants, we can see four different roles in privacy study.

1) Data generator. Individuals or organizations who generate the original raw data (e.g., medical records of patients, bank transactions of customers), and offer the data to others in a way either actively (e.g. posting photos to social networks to public) or passively (leaving records of credit card transactions in commercial systems).
2) Data curator. The persons or organizations who collect, store, hold, and release the data. Of course, the released data sets are usually anonymized before publishing.
3) Data user. The people who access the released data sets for various purposes.
4) Data attacker. The people who try to gain more information from the released data sets with a benign or malicious purpose. We can see that a data attacker is a special kind of data user.

There are three major data operations in a privacy system.

1) Collecting. Data curators collect data from different data sources.
2) Anonymizing. Data curators anonymize the collected data sets in order to release it to public.
3) Communicating. Data users performan information retrieval on the released data sets

Furthermore, a data set of the system possesses one of the following three different statuses.

1) Raw. The original format of data.
2) Collected. The data has been received and processed (such as de-noising, transforming), and stored in the storage space of the data curators.
3) Anonymized. The data has been processed by an anonymization operation.

We can see that an attacker could achieve his goals by attacking any of the roles and the operations.

In general, we can divide a given record into four categories according to its attributes.

1) Explicit identifiers. The unique attributes that clearly identify an individual, such as drive licence numbers.
2) Quasi-identifiers. The attributes that have the potential to re-identify individuals when we gather them to-

gether with the assistance of other information, such as age, career, postcode, and so on.

3) Sensitive information. The expected information interested by an adversary.

4) Other. The information not in the previous three categories

and age form the quasi-identifier, disease belongs to sensitive information.We call the quasi-identifiers of a record as a *qid* group, which is also called *equivalence class* in literature.

IX. THE MILESTONES OF PRIVACY STUDY

To date, the majority work on privacy protection is conducted in the context of databases. There are mainly two categories: data clustering and theoretical frameworks of privacy. The data clustering direction developed from the initial k-anonymity method, then the l-diversity method, and then the t-closeness (interested readers are encouraged to find the detailed information from [66]). The second category mainly includes the framework of differential privacy and its further developments.

X. DISCIPLINES IN PRIVACY STUDY

Based on the content of the previous sections, we can see that privacy research just started, and privacy research in big data is almost untouched. In this section, we try to survey the major disciplines involving in privacy study. Of course, the list of disciplines is not exhaustive.

A. CRYPTOGRAPHY

Based on the current situations in practice, we can conclude that encryption is still the dominant methodology for privacy protection although it is a bit away from the privacy protection theme we talking about here.

Cryptography can certainly be used in numerous fashions for privacy protection in the big data age. The public key encryption is obviously not convenient if the number of the authorized persons is sufficiently large due to the key management issue. In this case, Attribute Based Encryption (ABE) is an appropriate tool [67], [68], which was invented in 2004 by Sahai and Waters [69]. In the ABE scheme, a set of descriptive attributes of the related parties, such as hospital ID, doctor ID, and so on, are used to generate a secret key to encrypt messages. The decryption of a ciphertext is possible only if the set of attributes of the user key matches the attributes of the ciphertext.

B. DATA MINING AND MACHINE LEARNING

Data mining and machine learning are the biggest threat to modern privacy protection. The essential purpose of mining and learning is to obtain new knowledge from data sets.

A comprehensive survey in this field was done in 2010 by Fung et al. [66] in terms of privacy preserving data publishing. They surveyed the data publishing issue with privacy protection: given a data set *T*, how to transform it to a publishable data set *T'* under the condition of privacy protection of the data generators in *T*. They classified the attacks in two categories.

- Linkage attack. Attackers combine the publicly released data set *T'* with other data sets they possess to re-identify the data generators at different granularities.
- Probabilistic attack. An attacker gains more new knowledge about a victim based on the released *T'* compared with his original background knowledge of the victim before the releasing.

C. BIOMETRIC PRIVACY

Biometric is a powerful tool for security, which aims to identify individuals based on their physical, behavioral, and physiological attributes, such as face, iris, fingerprint, voice, and gait. Biometrics has been widely used in access control, and the procedure includes two stages: enrollment and release. In the first stage, biometric features, such as fingerprints, are sampled, and the information is stored in a database either as a raw data or in a transformed form. In the second stage, the related biometric characteristics are sampled again on site, and compared with the stored one for authentication

XI. SUMMARY

Big data has become one of the most promising and prevailing technology to predict future trends. In these circumstances, security and privacy should be taken into consideration for applications. In this paper, we have analysed the effects of big data characteristics on security and privacy, which requires conceptual and operational study on infrastructure security, data privacy and data management. Then, having surveyed the research on big data security and privacy in big data, a set of topics and issues have been illustrated and compared. Finally, we have focused on privacy-preserving trajectory data publishing mechanisms, due to sensitivity and widely- usage of trajectory data in telecom operation. Three common mechanisms of privacy-preserving trajectory publishing: generalization and suppression, perturbation and differential privacy have been presented, and relative research under each class have been also reviewed

References

[1] X. Cheng, L. Xu, et al., A Novel Big Data Based Telecom Operation Architecture[C], in Proc. 2015 International Conference on Signal and Information Processing(ICSINC), Beijing, China, Oct. 2015.

[2] L. Xu, Y. Luan, et al., WCDMA Data based LTE Site Selection Scheme in LTE Deployment[C], in Proc. 2015 International Conference on Signal and Information Processing(ICSINC), Beijing, China, Oct. 2015.

[3] M. Viceconti, P. Hunter, et al., Big Data, Big Knowledge: Big Data for Personalized Healthcare[J], IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 4, pp. 1209-1215, July 2015.

[4] M. Herland, T. M. Khoshgoftaar, et al., Survey of Clinical Data Mining Applications on Big Data in Health Informatics[C], in Proc. 2013 12th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2013, pp. 465-472.

[5] Deng Y, Wang L, et al, Artificial-Noise Aided Secure Transmission in Large Scale Spectrum Sharing Networks[J], IEEE Trans. on Communi- cations,vol. 64, no. 5, pp. 2116-2129, May 2016.

[6] A. A. Cardenas, P. K. Manadhata, et al., Big Data Analytics for Security]}], IEEE Security and Privacy, vol. 11, no. 6, pp. 74-76, Nov.-Dec. 2013.

[7] Cloud Security Alliance Big Data Working Group, Expanded Top Ten Big Data Security and Privacy Chanllenges[R], Apr. 2013.

[8] Y. Deng, L. Wang, M. Elkashlan, et al., Physical Layer Security in Three- Tier Wireless Sensor Networks: A Stochastic Geometry Approach[J], in IEEE Trans. on Information Forensics and Security, vol. 11, no. 6, pp. 1128-1138, Jun. 2016.

[9] ISO/IEC JTC 1, Big Data [R], Preliminary Report, 2014.

[10] D. S. Terzi, R. Terzi, et al., A Survey on Security and Privacy Issues in Big Data[C], in Proc. 2015 IEEE International Conference on Internet Technology and Secured Transactions(ICITST' 2015), 2015.

[11] C. Lynch, “Big data: How do your data grow?,” Nature, vol. 455, pp. 28-29, Sept. 2008

[12] A. Szalay and J. Gray, “2020 Computing: Science in an exponential world,” Nature, vol. 440, pp. 413-414, Mar. 2006.

[13] B. Matturdi, X. Zhou, et al., Big Data Security and Privacy: A Review[J], China Communications Magazine, 2014.

[14] D. Mittal, D. Kaur, et al., Secure Data Mining in Cloud Using Homomorphic Encryption [C], in Proc. 2014 IEEE International Conference on Cloud Computing in Emerging Markets(CCEM' 2014), Oct. 2014.

[15] NIST, NIST Big Data Interoperability Framework: Volume 4, Security and Privacy[R], National Institute for Standards and Technology, 2015, http://dx.doi.org/10.6028/NIST.SP.1500-4.

[16] M.Yang, X.Zhou, et al., Challenges and Solutions of Information Security Issues in the Age of Big Data[J]. China Communications Magazine, Mar. 2016.

[17] NIST, NIST Big Data Interoperability Framework: Volume 4, Security and Privacy[R], National Institute for Standards and Technology, 2015, http://dx.doi.org/10.6028/NIST.SP.1500-4.

[18] K. Hu, D. Liu, et al., Research on Security Connotation and Response Strategies for Big Data[J], Telecommunications Science, vol.2, pp.112117, Feb. 2014.

[19] R. Lu, H. Zhu, et al., Toward Efficient and Privacy-Preserving Computing in Big Data Era[J], IEEE Network, Aug. 2014.

[20] L. Xu, C. Jiang, et al., Information Security in Big Data: Privacy and Data MiningJ]. IEEE Access, vol.2, pp. 1149-1176, Oct. 2014.

[21] E. Sahafizadeh, and M.A.Nematbakhsh, A Survey on Security Issues in Big Data and NoSQL[J]. Advances in Computer Science: an International Journal(ACSIJ), vol.4, no.16, pp.68-72, Jul. 2015.

[22] E. Bertino, Big Data-Security and Privacy[C]. in Proc. 2015 IEEE International Congress on Big Data, Jun. 2015.

[23] K. Saranya, K. Premalatha, et al., A Survey on Privacy Preserving Data Mining[C], in Proc. 2015 IEEE Sponsored 2nd International Conference on Electronics and Communication System(ICECS'15), 2015.

[24] J. Abawajy, M. I. Ninggal, et al., Privacy Preserving Social Network Data Publication[J], IEEE Comm. Surveys Tutorials, vol. pp, no.x, Mar. 2016.

[25] B. C. Fung, K. Wang, et al, Privacy-Preserving Data Publishing: A Survey of Recent developments[J], ACM Comput.Surv., vol. 42, no.4, Jun. 2010.

[26] A. H. Rashid, and N. Yasin Privacy Preserving Data Publishing: Review[J], International Journal of Physical Sciences, vol.10 pp. 239-247, Apr. 2015.

[27] A.G.Divanis, P. Kalnis, et al., Providing K-Anonymity in Location Based Services[J]. SIGKDD Explorations, vol.12, no.1, pp.3-10, 2010.

[28] F. Bonchi, L. V. S. Lakshmanan, et al.,Trajectory Anonymity in Publishing Personal Mobility Data[J],

SIGKDD Explorations, vol.13, no.1, pp.30-42, 2011.

[29] Y. Zheng, Trajectory Data Mining: An Overview[J]. ACM Transactions on Intelligent Systems and Technology, vol.6, no.3, Article 29, May 2015.

[30] M. Terrovitis and N. Mamoulis, Privacy Preservation in the Publication of Trajectories[C], in Proc 9th International Conference on Mobile Data Management(MDM 2008), pp.65-72, Beijing, Apr. 2008.

[31] M. E. Nergiz, M. Atzori, et al., Towards Trajectory Anonymization: a Generalization-Based Approach[C], in Proc. ACM GIS Workshop on Security and Privacy in GlS and LBS, 2008.

[32] R. Yarovoy, F. Bonchi, et al,Anonymizing moving objects (how to hide a MOB in a crowd?[J], Extending Database Technology, 2009.

[33] Junqiang Liu, Privacy-Preserving Data Publishing: Current Status and New Directions [J], Information Technology Journal, vol. 11, no. 1, pp.1-9, 2012.

[34] Y. Xu, T. Ma, et al., A Survey of Privacy Preserving Data Publishing using Generalization and Suppression[J]. Appl. Math. Inf. Sci. vol. 8, no. 3, pp. 1103-1116, 2014.

[35] A. Al-Talabani, Y. Deng, et al., Enhancing Secrecy Rate in Cognitive Radio Networks via Multilevel Stackelberg Game[J], IEEE Commun. Letters, vol. 20, no. 6, pp. 1112-1115, Jun. 2016.

[36] O. Abul, F. Bonchi, et al., Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases[C], in Proc. 2008 IEEE International Conference on Data Engineering(ICDE'08), 2008.

[37] O. Abul, F. Bonchi, et al., Anonymization of moving objects databases by clustering and perturbation[J], Information Systems, vol.35, no. 8, pp. 884-910, Dec. 2010.

[38] V. Primault, S. B. Mokhtar, et al. Time Distortion Anonymization for the Publication of Mobility Data with High Utility[C], in Proc. 2015 IEEE Trustcom/BigDataSe/ISPA, Helsinki, Aug. 2015.

[39] K. Mano, K. Minami, et al.,, Pseudonym Exchange for PrivacyPreserving Publishing of Trajectory Data Set[C]. in Proc. 2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE), Tokyo, 2014.

[40] J. Hua, Y. Gao, et al.,Differentially Private Publication of General Time- Serial Trajectory Data[C], in Proc. 2015 IEEE Conference on Computer Communications(INFOCOM), pp.549-557, Kowloon, Apr. 2015.

[41] C. Dwork, F. Mcsherry, et al., Calibrating Noise to Sensitivity in Private Data Analysis[C], in Proc. Theroy of Cryptography Conference, Jan. 2006.

[42] F. Mcsherry, and K. Talwar, Mechanism Design via Differential Priva- cy[C], in Proc. Foundations of Computer Science, 2007.

[43] R. Chen, B. C. M Fung, et al., Differentially Private Trajectory Data Publication[J]. Arxiv e-prints, Dec. 2011.

[44] R. Chen, G. Acs, et al., Differentially Private Sequential Data Publication via Variable-Length N-Grams[C], in Proc. ACM Computer and Communication Security (CCS), Oct 2012, Raleigh, United States. 2012.

[45] S. Ghemawat, H. Gobioff and S.-T. Leung , "The Google File System" , SOSP , 2003.

[46] NIST Special Publication 15001-291 version 1, Definitions and Taxonomies Subgroup,September 2015,Available at http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf.

[47] S. Rusitschka and A. Ramirez, "Big Data Technologies and Infrastructures." http://byte- project.eu/research/, Deliverable D1.4, Version 1.1, Sept. 2014.

[48] "Hypervisors, virtualization, and the cloud: Learn about hypervisors, system virtualization, and how it works in a cloud environment." Retrieved June 2015.

[49] M. Portnoy, Virtualization Essentials. 1st ed., 2012.Alameda, CA, USA: SYBEX Inc.,

[50] S. Sharma, "Evolution of as-a-service era in cloud," CoRR, vol. abs/1507.00939, 2015.

[51] S. Sharma, U. S. Tim, J. Wong, S. Gadia, "Proliferating Cloud Density through Big Data Ecosystem, Novel XCLOUDX Classification and Emergence of as-a-Service Era," 2015

[52] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," tech. rep., July 2009.

[53] F. Liu, J. Tong, J. Mao, R. Bohn, J. Messina, L. Badger, and D. Leaf, NIST Cloud Computing Reference Architecture: Recommendations of the National Institute of Standards and Technology (Special Publication 500-292). USA: CreateSpace Independent Publishing Platform, 2012.

[54] R. Dua, A. Raja, and D. Kakadia, "Virtualization vs containerization to support paas," in Cloud Engineering (IC2E), 2014 IEEE International Conference on, pp. 610-614, March 2014.

[55] B. Russell, "Realizing Linux Containers (LXC)." http://www.slideshare.net/BodenRussell/linux- containers-next-gen- virtualization-for-cloud-atl-summit-ar4-3-copy. Retrieved October 2015.

[56] NIST Special Publication 500-291 version 2,

NIST Cloud Computing Standards Roadmap, July 2013, Available at http://www.nist.gov/itl/cloud/publications.cfm.

[57] N. Mimura Gonzalez, M. Torrez Rojas, M. Maciel da Silva, F. Redigolo, T. Melo de Brito Carvalho, C. Miers, M. Naslund, and A. Ahmed, “A framework for authentication and authorization credentials in cloud computing,” in Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on, pp. 509-516, July 2013.

[58] M. A. Leandro, T. J. Nascimento, D. R. dos Santos, C. M. Westphall, and C. B. Westphall, “Multi - tenancy authorization system with federated identity for cloud-based environments using shibboleth,” in Proceedings of the 11th International Conference on Networks, ICN 2012, pp. 88-93, 2012.

[59] M. Stihler, A. Santin, A. Marcon, and J. Fraga, “Integral federated identity management for cloud computing,” in New Technologies, Mobility and Security (NTMS), 2012 5th International Conference on, pp. 1-5, May 2012.

[60] N. Santos, K. P. Gummadi, and R. Rodrigues, “Towards trusted cloud computing,” in Proceedings of the 2009 Conference on Hot Topics in Cloud Computing, HotCloud’09, (Berkeley, CA, USA), USENIX Association, 2009.

[61] T. Garfinkel, B. Pfaff, J. Chow, M. Rosenblum, and D. Boneh, “Terra: A virtual machine-based platform for trusted computing,” in Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, SOSP ’03, (New York, NY, USA), pp. 193-206, ACM, 2003.

[62] M. Anand, “Cloud monitor: Monitoring applications in cloud,” in Cloud Computing in Emerging Markets (CCEM), 2012 IEEE International Conference on, pp. 1-4, Oct 2012.

[63] A. Brinkmann, C. Fiehe, A. Litvina, I. Luck, L. Nagel, K. Narayanan, F. Ostermair, and W. Thronicke, “Scalable monitoring system for clouds,” in Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing, UCC ’13, (Washington, DC, USA), pp. 351-356, IEEE Computer Society, 2013.

[64] J. Nikolai and Y. Wang, “Hypervisor-based cloud intrusion detection system,” in Computing, Networking and Communications (ICNC), 2014 International Conference on, pp. 989-993, Feb 2014.

[65] C. Basescu, A. Carpen-Amarie, C. Leordeanu, A. Costan, and G. Antoniu, “Managing data access on clouds: A generic framework for enforcing security policies,” in Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference on, pp. 459-466, March 2011.

[66] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, ‘‘Privacy-preserving data publishing: A survey of recent developments,’’ ACM Comput. Surv., vol. 42, no. 4, 2010, Art. no. 14.

[67] V. Goyal, O. Pandey, A. Sahai, and B. Waters, ‘‘Attribute-based encryption for fine-grained access control of encrypted data,’’ in Proc. 13th ACM Conf. Comput. Commun. Secur. (CCS), Alexandria, VA, USA, Oct./Nov. 2006, pp. 89-98.

[68] A. B. Lewko and B. Waters, ‘‘Decentralizing attribute-based encryption,’’ in Proc. 30th Annu. Int. Conf. Theory Appl. Cryptogr. Techn., Tallinn, Estonia, May 2011, pp. 568-588.

[69] A. Sahai and B. Waters, ‘‘Fuzzy identity-based encryption,” in Proc. IACR Cryptol. ePrint Arch., , p2004. 86.