Name: Adarsh Mathad Vijayakumar

CWID: A20424847

Email: avijayakumar@hawk.iit.edu

**CSP554 – Research Paper Draft**

***Privacy Protection in Big Data Environment: A Technological Perspective and Review***

**Introduction**

Due to the fast growth of emerging information technologies such as Internet of Things (IoT), cloud computing, Internet services, and social networking an increasing interest in big data security and privacy is aroused. The increasing amount of big data also increases the chance of breaching the privacy of individuals. An entire lifetime of big data can be broadly classified into three phases: big data generation; processing and analytics; storage and management. Since big data require high computational power and large storage, distributed systems are used. As multiple parties are involved in these systems, the risk of privacy violation is increased. The five salient features of big data: volume, variety, velocity, value, and veracity bring great challenges on protecting big data security and privacy during its whole lifetime. Big data normally contains valuable or sensitive information about user behaviors, preferences, interests, mobility, and so on. User privacy is easily leaked if the data cannot be protected well during its lifetime. Therefore, if we want to enjoy the convenience and benefits from big data, guarantee of its security and privacy becomes an essential task.

**Entire Lifetime of Big Data**

To handle different dimensions of big data in terms of volume, velocity and variety we need to design efficient and effective systems to process large amount of data arriving at very high speed from different sources. Big data has to go through multiple phases – like Data Generation, Data Storage and Data Processing - during its life cycle.



*Figure 1: Illustration of the stages of the big data lifetime*

If we want to enjoy the convenience and benefits from big data, guarantee of its security and privacy becomes an essential task. We may get wrong information from big data once some insecure events occur in the above phases. Thus, security and privacy of big data in the above phases are very important. Herein, we regard these three phases as big data lifetime.

**Privacy preserving mechanisms is each stage of the big data lifetime**

**Big data privacy in data generation phase**

Data generation can be classified into **Active Data Generation** and **Passive Data Generation**. By active data generation, we mean that the data owner will give the data to a third party, while passive data generation refers to the circumstances that the data are produced by data owner’s online actions (e.g., browsing) and the data owner may not know about that the data are being gathered by a third party. The major challenge for data owner is that how can he protect his data from any third party who may be willing to collect them. The data owner wants to hide his personal and sensitive information as much as possible and is concerned about how much control he could have over the information. Minimization of the risk of privacy violation amid data generation can be done by either restricting the access or by falsifying data.

A. ACCESS RESTRICTION

If the data owner thinks that the data may reveal sensitive information which is not supposed to be shared, he can simply refuse to provide such data. For that, the data owner has to adopt effective access control methods so that the data can be prevented from being stolen by some third party. If the data owner is providing the data passively, some measures could be taken to ensure privacy, such as anti-tracking extensions, advertisement/script blockers and encryption tools. By using these tools, one can effectively limit the access to sensitive data. For the ease of use, most of these tools are designed as browser extensions.

In addition to these tools, there are some alternative means, such as to use anti-malware and anti-virus software to protect the data stored digitally on their computer or laptop. These tools can help to protect user's personal data by limiting the access. Though there is no guarantee that one's sensitive data are completely protected from untrustworthy sources, making it a habit of clearing online traces/cookies/sessions of one's activity by using security tools can significantly reduce the risk.

B. FALSIFYING DATA

In some circumstances, it may not be possible to protect the personal data/sensitive data. In that case, data can be distorted using certain tools prior to the data gotten by some third party. If the data are distorted, then the true information cannot be easily revealed. The following techniques are utilized by the data owner to falsify the data:

* ***A tool Socketpuppet*** is used to hide online identity of individual by deception. Individual's true activities online are concealed by creating a false identity and pretending to be someone else. By using multiple Socketpuppets, the data belonging to one specific individual will be deemed as belonging to different individuals. In that way the data collector will not have enough knowledge to relate different socketpuppets to one individual. Hence, the user’s true activities are unknown to others and the private information cannot be discovered easily.
* Certain ***security tools*** can be used to mask individual’s identity, such as MaskMe. It allows users to create aliases of their personal information such as email address or credit card number. The data owner can use these masks whenever information is needed. This is especially useful when the data owner needs to provide the credit card details during online shopping.

**Big data privacy in data storage phase**

Storing high volume data is not a major challenge due to the advancement in data storage technologies. If the big data storage system is compromised, it can be exceptionally destructive as individual’s personal information can be disclosed. Therefore, we need to ensure that the stored data are protected against such threats. In modern information systems, data centers play an important role of performing complex commutations and retrieving large amount of data. In distributed environment, an application may need several datasets from various data centers and therefore confront the challenge of privacy protection.

The conventional security mechanisms to protect data can be divided into four categories. They are file level data security schemes, database level data security schemes, media level security schemes and application level encryption schemes. Responding to the 3V’s nature of the big data analytics, the storage infrastructure ought to be scalable. It should have the ability to be configured dynamically to accommodate various applications. One promising technology to address these requirements is storage virtualization, empowered by the emerging cloud computing paradigm. Storage virtualization is process in which numerous network storage devices are combined into what gives off an impression of being a single storage device. However, using a cloud service offered by cloud provider means that the organization's data will be outsourced to a third party such as cloud provider. This could affect the privacy of the data.

#### A. Approaches to privacy preservation storage on cloud

When data is stored on cloud, data security predominantly has three dimensions - confidentiality, integrity and availability. The first two are directly related to privacy of the data i.e., if data confidentiality or integrity is breached it will have a direct effect on users’ privacy. Availability of information refers to ensuring that authorized parties are able to access the information when needed. A basic requirement for big data storage system is to protect the privacy of an individual. The approaches to safeguard the privacy of the user when data are stored on the cloud are as follows:

1. *Attribute Based Encryption (ABE)*: This is an encryption technique which ensures end to end big data privacy in cloud storage system. In this, access polices are defined by data owner and data are encrypted under those policies. The data can only be decrypted by the users whose attributes satisfy the access policies defined by the data owner. When dealing with big data one may often need to change data access policies as the data owner may have to share it with different organizations. This can be done in an easy and effective way as described next here. The data owner can send the queries to cloud to update the policy, and the cloud server can update the policy directly without decrypting the data.
2. *Identity Based Encryption (IBE)*: This is an alternative to the conventional Public Key Encryption which is proposed to simplify key management in a certificate-based public key infrastructure (PKI) by using human identities like email address or IP address as public keys. To preserve the anonymity of sender and receiver, the IBE scheme was proposed. By employing these primitives, the source and the destination of data can be protected privately.
3. *Homomorphic Encryption (HE)*: Homomorphic encryption has been recognized as one of the ideal approaches to securing and processing big data in remote servers including the cloud. Homomorphic Encryption (HE) is a method of secure computation which allows for calculations to be made on encrypted data without decrypting it and without giving away information about the operations being done.
4. *Storage Path Encryption:* In the proposed scheme, the big data are first separated into many sequenced parts and then each part is stored on a different storage media owned by different cloud storage providers. To access the data, different parts are first collected together from different data centers and then restored into original form before it is presented to the data owner. A trapdoor function has been incorporated in this scheme. It is a function which is easy to compute in one way and difficult to compute in the opposite direction without some additional information. The owner of the big data will keep the storage index information.

#### B. Integrity verification of big data storage

When cloud computing is used for big data storage, data owner loses control over data. The outsourced data are at risk as cloud server may not be fully trusted. The data owner needs to be strongly convinced that the cloud is storing data properly according to the service level contract. One way to ensure privacy to the cloud user is, to provide the system with the mechanism to let data owner verify that his data stored on the cloud is intact. The integrity of data storage in traditional systems can be verified through number of ways i.e., Reed-Solomon code, checksums, trapdoor hash functions, message authentication code (MAC), and digital signatures etc. Therefore, data integrity verification is of critical importance. It is highly prescribed that the integrity verification should be conducted regularly to provide highest level of data protection.

**References:**

1. <https://www.researchgate.net/publication/301716085_Protection_of_Big_Data_Privacy>
2. <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0059-y>
3. <https://www.researchgate.net/publication/309817879_Security_and_Privacy_in_Big_Data_Lifetime_A_Review>
4. <https://www.ijedr.org/papers/IJEDR1702165.pdf>