Big Data Privacy Protection Technology

Jing Yang*
School of Shanghai University, Shanghai, China

*Corresponding author e-mail: yuren@shu.edu.cn

Abstract. With the development of Internet technology, we have ushered in the "big data era". The arrival of big data has made data calculation, acquisition, and statistics more convenient, and has had a significant impact on people's lives, social management, and national governance. However, because the information unearthed by big data analysis technology may exceed people's imagination, its privacy issues have also attracted much attention. This paper first briefly introduces the definition and current status of big data and its privacy issues, describes the privacy protection link model, and then starts from the privacy protection technology of big data, introduces the K-anonymization traditional big data privacy protection technology, advantages and disadvantages, Specific algorithms, etc.

Keywords: Big Data, Privacy Protection, Privacy Leakage

1. Introduction
At present, big data is developing very rapidly at home and abroad, and it has become another growth point of the information industry in the field of information technology after cloud computing. Data business has become the "new normal" of industrial development, but the security and privacy protection issues of big data that follow are not uncommon. With the rapid increase in data dissemination and growth capabilities, data privacy has gradually become transparent, and our privacy may be maliciously stolen during the various processes of data dissemination. The privacy protection of big data is not only for the privacy of users, but also for better use of big data.

2. Big Data and its Privacy
Big data has many formats and rapid growth. Because of its high-speed, large-scale, and diverse characteristics, it is difficult to achieve effective processing of big data with a single database management technology and tool. From the perspective of basic technology, the basic technology on which big data depends is NoSQL (Non-Relational Database). SQL (relational database) is now widely used, and its technology has been improved and perfected for a long time. SQL has strict standards, access control, and privacy management tools for maintaining data security. But in NoSQL technology, there are no such requirements. The security challenges of big data are that it is difficult to effectively apply for permissions from users, to implement role presets, to predict and monitor the access behavior of developers, and to prevent excessive utilization of big data, a lot of data collection often takes place without knowing the purpose, and value creation is often secondary development. Companies cannot foretell users of unexpected uses, and individuals cannot predict the unknown uses.
In addition, some data do not represent the truth and can be deceptive. First, for their own benefit, some companies create false data, create illusions, and deceive consumers; second, data lag can also lead to data duplication or inaccuracy.

For the term privacy, the commonly accepted definition in scientific research is "certain attributes of a single user." As long as it meets this definition, it can be regarded as privacy[1]. When we talk about "privacy", we put more emphasis on "single user". Then, certain attributes of a group of users can be regarded as not privacy. Therefore, from the perspective of privacy protection, privacy is the concept of individual users. Disclosing the information of group users is not a privacy leak[2]. We often use group information when we conduct statistical activities, but if we can get accurate data from the data need to infer the individual's information, then even if it is a privacy leak[3].

3. Privacy Protection Link Model
A privacy protection system includes various participant roles (participation role), anonymization operation (anonymization operation) and data status (data status). The relationship between them is shown in Figure 1.

![Figure 1. Privacy protection system diagram](image)

In the study of privacy protection, there are 4 data participant roles.

1. Data generator: refers to those individuals or organizations that generate raw data, such as patients’ medical records and customers’ banking transactions. They actively provide data in some way (such as posting photos to social networking platforms) or passively providing data to others (such as leaving personal credit card transaction records in e-commerce, electronic payment systems, etc.).

2. Data curator: refers to those individuals or organizations that collect, store, control, and release data.

3. Data user: refers to users who access the published data set for various purposes.

4. Data attacker: Refers to those who attempt to obtain more information from the published data set for benign or malicious purposes. A data attacker is a special type of data user.

First, the generation and owner of big data. Big data owners acquire data/information in active or passive forms, such as transaction information of bank users and user name, telephone number, address, occupation, bank deposit, economic status, consumption habits and other information filled in by users when they conduct transactions or open an account in the bank through big data. Once these information records complete the user transaction process, they will be out of the control of the user who is the creator of the data, and become the information that the bank holds and uses, posing a huge threat to the user's privacy protection.

Second, data collection and user privacy leakage caused by regulatory agencies. These units or individuals with big data information are usually data managers or data analysts, who use various
technologies and methods to analyze and mine large amounts of data, to find valuable information, and to mine business value from it. In this process, if the user information is not anonymized, there is a risk of revealing the user's privacy.

The third aspect is data users, that is, data or query information about data obtained through paid or unpaid methods from data collectors. Although these data have been processed by desensitization, they may also be restored through some techniques, resulting in leakage of user privacy.

Finally, there are data attackers. These data attackers gain access to relevant data information through legal purchases and illegal attacks, access to sensitive data, including name, age, consumption habits, etc., so as to perform certain tasks to achieve a certain purpose. Active behaviors and data attackers are the most likely factors to cause user privacy leakage.

In short, there is a lack of effective supervision and supervision technology in the generation, storage, use, and supervision of big data, and users cannot ensure whether their relevant information is used for reasonable research or illegal trading.

In addition, it can be seen from Figure 1 that there are three main data operations in the privacy protection system, including data collection, anonymization, and data transmission.

1. Data collection (collecting): refers to data managers collecting data from different data sources.
2. Anonymizing: refers to the anonymization of collected data by data managers in order to publish data.
3. Data transmission (communicating): Refers to data users performing information retrieval on published data sets.

The data set of the privacy protection system also has the following three different states.

1. Raw status: refers to the original format of the data.
2. Collection state: the data has been received and processed (such as denoising, data conversion, etc.) and stored in the data manager.
3. Anonymized status: refers to the status after the data has been processed through anonymization operations.

As shown in Figure 1, the goal of a data attacker can be achieved by attacking any data role and data operation.

4. The Significance of Big Data Privacy Protection

We must realize that big data security and privacy protection can bring about national stability and economic prosperity[4]. The security of big data is not only related to ordinary life and production, but also related to the country’s long-term peace and stability and national security. Security and privacy protection can safeguard national cyberspace sovereignty, protect national information security, enhance national information data competitiveness, and maintain national stability and sustainable economic development. At the same time, big data has made breakthroughs in technologies such as security and privacy protection, which may generate new industries or industries. The resulting data services and data materials have entered a new stage of informatization and digitization, and finally realized man-machine -The organic integration of the three materials has realized the upgrading of industrial technology. In short, big data security and privacy protection are of profound significance, and we should attach great importance to them.

5. Challenges and Opportunities of Big Data Privacy Protection

Here are the challenges and opportunities of big data privacy protection:

1. Privacy measurement issues[5]. Privacy is a subjective concept that changes from person to person over time, making it difficult to measure. Privacy measurement is a difficult subject, which needs not only technical research, but also sociological and psychological research.

2. The theoretical framework of privacy protection[6]. At present, the privacy protection methods mainly include data clustering method and the theoretical framework of differential privacy protection. However, data clustering privacy protection methods, such as K-anonymity, have some limitations. At
present, differential privacy protection is mainly used in privacy protection. In the age of big data, it is a new challenge to research new and groundbreaking privacy protection theories.

3. Extensibility of privacy protection algorithm. Some existing mechanisms and strategies mainly adopt divide-and-conquer method when dealing with large databases. However, compared with large databases, the scale of big data is larger, and how to design scalable algorithms to protect privacy is also a difficulty. [7].

4. Heterogeneity of data sources. Current privacy protection algorithms are mainly oriented to homogenous data, similar to records in a database. However, the data source of big data is not homogenous data, but heterogeneous data. How to efficiently process heterogeneous big data is also a challenge for the future.

5. Efficiency of privacy protection algorithm. A large amount of big data needs to be calculated, and its computational efficiency cannot be underestimated. How to improve the efficiency of the algorithm will be a challenge in the process of big data computation. It is foreseeable that in the future, people need to improve the current privacy protection methods to meet the unprecedented demand in big data computing[8]. Furthermore, they expect the emergence of new privacy protection frameworks and mechanisms. This article believes that the following research directions have certain prospects, and it is worthwhile for privacy protection researchers to invest time and energy in research.

6. Conclusion
In terms of data encryption, big data can be obtained from multiple channels. Data is usually stored on a cloud platform[9]. Therefore, the key to privacy protection is how to ensure data security during storage. Contrary to our assumptions, in the real world, insecure cloud platforms actually exist. On such platforms, user data and privacy may be disclosed. Based on this, an important research topic in the field of data encryption is how to protect data privacy and three-party interaction privacy through technologies such as public key encryption, functional encryption, and homomorphic encryption.

In terms of differential privacy, in the differential privacy protection technology, the privacy parameter $\varepsilon$ is very important. This parameter directly determines the efficiency of differential privacy, and it determines the balance of data privacy and availability[10]. Therefore, designing a reasonable privacy parameter is the direction that differential privacy technology is always worth studying.

In general, as big data is more and more widely used, big data protection technology is gradually paid attention to. The combination of the use of big data and the protection technology of big data will standardize the use of big data and give full play to the value of big data.

References
[1] Yao Peng. Big data security and privacy protection. Technological wind. October 2019: 106.
[2] Ran Ran, Li Feng, Wang Xiniu, Yang Lichun, Ding Hongfa. A privacy protection-oriented power big data desensitization scheme and application research. Cyberspace Security, 2018(1): 105.
[3] Zhang Xiaofen, Research on Key Technologies of Big Data Security and Privacy Protection. Policy and Commercial Law Research. 2019, Issue 32: 146-147.
[4] Fang Xianjin, Xiao Yafei, Yang Gaoming. Big data and its privacy protection. May 2017: 45-56.
[5] Liu Yi, Deng Qing, Peng Yusu. Security Challenges of Data Sovereignty and Gold Wire Protection in the Era of Big Data. Risk Management. January 2019: 104-107.
[6] Feng Dengguo. Zhang Min. Li Hao. Big data security and privacy protection. Computer Science News, 2014, (37).
[7] Zeng Qin. Research on privacy protection and its key technologies in a big data environment. Digital Technology and Application, 2018, (07).
[8] Liu Liang. Technical Essentials of Big Data Security and Privacy Protection. Information and Computer (Theory Edition), 2016.
[9] Yu Shi. Research on Big Data Security and Privacy Protection Technology. Digital Technology and Application, 2017.
[10] Peng Changgen, Ding Hongfa, Zhu Yijie, etc. Information Entropy Model and Measurement Method for Privacy Protection. Journal of Software, 2016, 27(8): 1891-1903.