Personal Data Protection and Anonymization in the Process of Data Commodity Trading

Jing Su*,1, Jiale Gai1, Yaqing Si1, Xinyu Zheng1, Guangkai Li2, Zhixue Liu2

1School of Economics and Management, Beijing University of Posts and Telecommunications, 100876, Haidian District, Beijing, China
2Electric Power Research Institute, China Southern Grid, 510663, Huangpu District, Guangzhou, China
bysj@sina.com

Abstract. Big data brings tremendous commercial value, but at the same time, it also leads to personal information leakage and other problems. This paper analyses the laws and regulations of the personal information and anonymization at home and abroad, and discusses the legal identification criterion of personal information, personal data and data anonymization systematically. And combining with the practice, this paper provides general methods for anonymizing data with personal privacy when they are used or traded by companies.

1. Introduction

In the context of digital economy, companies are all aware of the huge commercial value brought by big data, and various data trading platforms emerge in the market. On the one hand, it responds to the policies of data opening and circulation advocated currently by China, which makes the value of data realized as a new productive factor. On the other hand, the privacy issues in the process of data trading also follow.

When users use all kinds of services on the Internet, network service providers often collect a large number of data containing personal information. If these data with personal information directly used or traded without professionally processing, it will inevitably lead to the leakage of personal information of users. How to find a balance among protecting personal privacy information, ensuring data availability and meeting the needs of data buyers has already become an urgent issue.

Based on the existing legal literature at home and abroad, this paper analyzes the definition of personal data and anonymization, and summarizes the principles and scenarios of anonymization during using and trading process. Then it gives the generally feasible anonymization methods and relevant suggestions.

2. Legal Literature at Home and Abroad

2.1. Personal Information and Personal Data

The EU promulgated the General Data Protection Regulation (GDPR) on May 25, 2018 in order to prevent the abuse of personal information at present, greatly protect the security of personal data in the process of using and trading. In Article.4, it defines that “‘personal data’ means any information relating to an identified or identifiable natural person (data subject)”. And data is the carriers of information. This regulation pushes forward the progress of the personal data protection reformation all around the
world. European countries and America have further strengthened the protection of personal data and established complete data protection frameworks and systems which give citizens more control over their own information. For example, The United Kingdom promulgated *A New Data Protection Bill: Our Planned Reforms*. California has promulgated the *California Consumer Privacy Act*. Developing countries have also begun to take measures. India promulgated *Personal Data Protection Bill* in 2019.

Personal information protection legislation in China is in the process. For the time being, there is no unified personal data protection bill. *The draft of Personal information protection law in 2017* defined the personal information in the Chapter I Article 3. It defines that personal information means any information relating to an identified natural person alone or in combination with other information electronically or on other ways, including but not limited to natural person’s name, date of birth, ID number, personal biometric information, address, telephone number, etc. Personal data is the carriers of recording personal information in electronic form. In terms of national legislation, there is similar definition in the Article 76 of *Cybersecurity Law of the People’s Republic of China* and the Article 1034 of *Civil code of the people’s Republic of China*. *The Personal information security specification* updated in 2020 defines that the personal information also includes the personal information or others after processed by controller, such as user portrait and feature label, which can identify the identity or be related to the activities of a specific natural person independently or in combination with other information.

From domestic and international viewpoint, the key to defining personal data is whether the data can identify or be related to the natural human or not. The most important thing is identification, which means a natural person can be identified according to a certain data object. Identification is divided into direct identification and indirect identification. Direct identification is that the data can identify a specific person by itself, such as ID number, telephone number, name. While indirect identification means that the data cannot identify a specific person by itself, but it can be related to a specific natural person after comprehensive analysis by combining with other data. For example, it can identify a natural person by combining one or more data such as location, gender, age.

Although indirect identification expands the definition of personal data, personal data is much more than that according to existing laws of each country. The scope and extension are much wider than the previous regulations. For example, according to the GDPR, the user’s location data and online activity identification symbols, such as IP address, MAC address, COOKIE on the Internet are clearly included in the scope of personal data.

2.2. Pseudonymisation

Overseas, EU’s *General Data Protection Regulation* calls anonymization as “pseudonymization” in the Articles 4.

> “Pseudonymisation” means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data is not attributed to an identified or identifiable natural person;

And in the Articles 25 and 32, EU *General Data Protection Regulation* request that

> the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation and data minimization.

The related content of anonymization in China first appeared in medical literature. In Articles 11 of the *Electronic Medical Record Management System Functional Specifications* in 2010, it has already involved the anonymization processing and requests that the hospital should provide patients the function of anonymizing their own electronic medical records.

In recent years, China begins to pay more attention to the anonymization of personal information data. The Articles 42 of *Cybersecurity Law of the People’s Republic of China* requests that Network operators shall not disclose, tamper with or damage the personal information which they collect, and
personal information they collected shall not be provided to others without permission of the owner, except the data unable to identify specific individuals and recover after processing. There is similar regulation in the Articles 1038 of *Civil code of the people's Republic of China*. And it requests information processors should take technical measures and other necessary measures to ensure the security of the personal information which they collected and stored and prevent the information from leaking, tampering and losing. Personal information security specification updated in 2020 also request that the personal information controller needs to de-identify immediately after collecting personal information, and take technical and management measures to store the information separately from the de-identification information.

### 2.3. Data Anonymization

According to the relevant regulations of data anonymization in domestic and foreign legal literature, the data that cannot be identified and recovered after processing is not belonged to personal data and protected by law, so the definition of data anonymization can be summarized. Data anonymization, known as data de-identification or data pseudonymisation, refers to the process of technical transformation of personal data according to certain rules, which makes the data deformed, fuzzy and unable to identify the data subject, and cannot be restored to the original data. Although the data after anonymization cannot identify specific individuals, these data still have considerable commercial value, companies can make them into a service data as a product for using or trading.

Generally, data anonymization includes the following three objectives. First, data processors need to anonymize data under strict and efficient standards to ensure that the data after anonymizing cannot identify specific individual directly or indirectly by combining with other data. In addition, after data anonymization it is necessary to establish a user personal data leakage risk model and conduct real-time monitoring which can analyze the risk that users may disclose data, so that the data will not lose actual effect with the emergence of new data, nor will it be increase the risk of being attacked due to the development of technology. Second, the data processor should record the data processing process and establish an access mechanism to give users an access to their own data. The processing and use of data should be traceable on the basis of records. When a data leak occurs, responsibility of the leakage link through the source tracking mechanism. Third, companies should establish a standardized system according to relevant laws and regulations to maximize the business value of data on the premise of protecting privacy as far as possible.

The anonymization of personal data by the data processor shall comply with the relevant laws and regulations, standards and specifications, as well as the relevant provisions on personal information security protection in the data supply and demand contracts. Data processors should continue to improve the methods, techniques and tools for data anonymization, and regularly re-evaluate the risk of data identification to ensure that the data is continuously controllable, safe and effective during use.

In addition, data processors should implement anonymizing measures by using both technical and management methods comprehensively to ensure the availability of data after desensitization, and maximize the value on the basis of fully protecting the security of personal data[2].

### 3. Anonymization Method in the Process of Data Commodity Trading

Network service platforms collect data during users use the network. After collection, the platform needs to anonymize data immediately to convert personal privacy data into usable data resources. After anonymization, the data can be placed in the database of the companies to be retained, used or traded[3]. What follows in the paper summarizes and analyze the anonymization methods in the process of using data based on the *General Data Protection Regulation*, *The draft of Personal information protection law* in 2017, and *Cybersecurity Law of the People's Republic of China*.

Common methods of data anonymization include but are not limited to this table:

| Technology | Method | Explanation | Applicable data types | Example |
|------------|--------|-------------|-----------------------|---------|

Table 1. Common methods of data anonymization.
| Statistical technology | Data Sampling | Selecting a representative subset from a large amount of data in the original data set for analysis. | Data sets with a large amount of data. | Selecting 10000 citizens information from 10 million randomly for anonymizing. |
|------------------------|-------------|-----------------------------------------------------------------|---------------------------------|--------------------------------------------------------------------------------|
| Data aggregation       | Calculating the data by summing or averaging to produce representative records instead of the original data. | In 2019, the average height of men aged 18 and above in China is 167.1cm. In the height attribute of a certain male data set, 167.1cm can be used to record. |
| Shielding              | Complete Shielding | Deleting all or part of the data that can directly identify personal information in the original data set. | Name, age, ID number, height, weight, or other numerical data. | Name, ID number, height, weight, salary, etc. can be replaced by “***”. |
|                       | Local Shielding | Deleting specifically or partially data that indirectly identifies an individual from the original data set. | Postal code, bank card number, telephone number, date, time or other numerical data. | Replace a part of the mobile phone number with “***”. For example, “13836125648” can be replaced by “13****85648”. |
| Record shielding       | Deleting several unique data records in the original data set separately. | The data with obvious differences from others in original data set. | In the data record of a certain examination, only one student failed, which means that he can be identified through the data of failing, so the record can be deleted. |
| Pseudonymization       | Pseudonymization | Using fake data to replace the data that can directly identify personal information in the original data. | Name, net name, user name and so on. | Selecting fake data randomly in the data dictionary with thousands of fake data instead of the original data, such as using “Zhang Wensheng” in the data dictionary instead of “Sun Zhengbin” in the original data set. |
| Randomization       | Noise addition                                                                 | ID number, phone number, weight, height, age or other numerical data. | Generating a value randomly and added to the original data value to replace themselves. For example, if the height value is added 0.05m, the height of 1.68m will be 1.73m. |
|---------------------|--------------------------------------------------------------------------------|---------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Transposition       | Rearranging the attribute values of a string of data records in the original data set. | Age, weight, height, salary, date or other numerical data.          | The distribution position of height data in the original data set can be disturbed and placed in the original data set according to the order defined by the rules. |
| Classification      | Grouping the original data set is by a certain rule, and then use a certain calculation method of different groups. | Data set with large amount of data and part of data can be grouped. | For a certain urban resident data set, it can be grouped by different urban areas, in which there are at least x records in each group, and then the attribute values in different urban areas groups can be anonymized by other methods. |
| Complete Randomization | Changing certain original data directly through randomization technology. | Personal age, weight, height, salary or other numerical data. | Change the data according to completely random rules, such as 758 becomes 364. |
| Generalization      | Rounding the values in the selected data according to specific rules.            | Personal age, weight, height, salary and other numerical data.      | Use "Mr. Zhang" instead of "Zhang Wensheng" for the name; For ages over 60 years old, the specific age will always be “over 60 years old”. Wages can be generalized to 5k-10k, 10k-15k, 15k-20k. The working years are generalized to 0-3 years, 4-6 years. |
| Maximization or Minimization | Replacing all data values in a column with the maximum or minimum value of them. |                                                                                   |                                                                                                                                                                                                |
| Interval method     | Expanding the selected value to the range of unified length.                     |                                                                                   |                                                                                                                                                                                                |
| Fuzzification       | Using generalization or abstraction for the selected values.                     |                                                                                   |                                                                                                                                                                                                |
| Encryption          | Using cryptography                                                               | Name, ID                                                                           | Use "ASDFAKASD" instead of the original name.                                                                                                                                                    |
In the process of data anonymization, different methods of anonymization can be adopted for different types of data according to the data’s ability to identify individuals. Personal data which can be able to directly identify require strict anonymization methods, while those data that are more difficult to identify a specific individual can be anonymized in a loose way. Take the data generated by users in the transaction process of e-commerce platforms as an example, the data that can be fully disclosed without anonymization include the information which describe the commodity or service, such as commodity specifications, commodity prices, and the public information of the seller such as store name, address, contact information. The data that needs to be anonymized partly includes personal information such as home address, personal birthday, ethnicity, gender and other information generated during website browsing records and transactions. Data that requires complete anonymization include identity information such as name, mobile number, user ID, email address, ID card, passport, sensitive information such as religious belief, sexual orientation, physical disease, and criminal record, biometric information such as fingerprint, voiceprint, palm print, iris, face recognition, network identity information such as login account, IP address and password, and device information such as Login main hardware serial number, device MAC address, unique device ID.

4. Realistic Confusion and Prospects
At present, there is no unified standard of data anonymization around the world, and there is no clear boundary between the content and extent of personal data used by companies. There is no specific statutory requirement of anonymization rules and methods, and there are not enough relevant laws to punish the companies which neglect their duty. Many companies do not comply with the law in terms of data acquisition, use and sale. Most of the data anonymization methods adopted by some companies are recoverable. Companies still retain the key algorithm to recover the original data. Although it increases the difficulty of other companies to restore it, there are still reversible risks in personal data. In addition, some data owners are afraid to invest data in the field of data circulation because of the unclear legal boundary and the uncertainty of the effectiveness of anonymization, which greatly limits the effective supply of data trading market[4].

The establishment of uniform rules and standards for data anonymity is still an urgent task to protect personal information. In terms of legislation, laws and regulations stay at the level of principles and basic norms, and lag behind in the protection of personal data and the standards of laws and regulations of anonymity. Besides, data collectors should consciously abide by existing laws and regulations, and reasonably formulate service agreements, and use data in compliance. Data processors should continuously strengthen the research and development of key technologies such as data anonymization, anti-recovery, anti-identification.

Acknowledgments
This work was supported by National Key R&D Program of China (2017YFB1401100).

References
[1] Han XZ. Legal Construction of Anonymization Information in the Big Data Era[J]. Journal of Dalian University of Technology, 2018,39(04):64-75(in Chinese).
[2] Xie AM, Jin Tao, Zhou Tao. Personal information de-identification architecture and standardization[J]. Big Data Research, 3(5):20-29(in Chinese).
[3] Jiang L. Research on service data resource rights and desc description standard based on ODRL [D]. Beijing: Beijing University of Posts and Telecommunications,2019:22-25(in Chinese).
[4] Wang R. The legal Requirements for Data Anonymization[J]. Information and Communications Technologies, 2016, 10(4):38-44 (in Chinese).