Legitimacy of Personal Data Mining and Related Regulatory Approaches: China’s Choices

Shujie Cui
Department of Law, Xi’an Jiaotong University, Xi’an 710049, China
cuishjie1990@163.com

Abstract. Personal data mining is a crucial but controversial topic in the big data era. It is widely accepted in the technical filed as a big data technology. While more and more people begin to question its legitimacy for it causes many social and legal problems. This paper tries to demonstrate the legitimacy of personal data mining from the following reasons: a) personal data has certain public goods attribute and sharing characteristic, which is the internal premise of personal data mining; b) data appreciation theory is the theoretical basis of personal data mining; c) the principle of interests balance is the logical requirement of personal data mining; d) big data sharing economy provides practical soil for personal data mining. Based on the legitimacy analysis and China’s data development practices, personal data mining should be encouraged. Meanwhile, it is particularly important to regulate the approaches of personal data mining in four aspects: a) establishing the purpose of combining personal data mining and protection; b) balancing the interests of all parties in the process of personal data mining; c) improving data classification legislation; d) innovating technical means to achieve privacy protection.

1. Introduction

Human beings have entered the digital era day and night. Data is attracting more and more attention as a new type of resource. The application of data has given birth to the data industry, promoted the development of the data market, and brought about extremely high economic and social values. Personal data mining is an important part of big data technology and the foundation of in-depth data value. In 2015, the Chinese government issued the Outline for Promoting the Development of Big Data of China, emphasizing the promotion of data mining. Subsequently, Year of Big Data Industry Development Plan (2016-2020) pays close attention to improve the ability of big data mining. Meanwhile, personal data mining, as an important part of big data mining, has become a significant theme of China’s data development process.

However, personal data mining is a double-edged sword. In recent years, with the deep mining and using of personal data, personal data leakage, privacy infringement and other incidents occurred frequently. In 2018, 50 million Facebook user data and 1 billion YTO express data were leaked and stolen successively. Personal data security is challenged. In the era of big data, forbidding personal data mining does not meet the practical requirements. This study focuses on the issue of the legitimacy and related regulatory approaches of personal data mining. First of all, this paper defines personal data mining from the legal perspective. Secondly, the legitimacy of personal data mining is presented from theoretical and practical aspects. Thirdly, based on China’s data development practice, some strategies are proposed for regulating China’s personal data mining system.
2. Legal Definition of Personal Data Mining

In the period of industrial economy, personal data belongs to the category of privacy norm and is passively protected by privacy right. While in the big data era, personal data is widely collected and used, and its property value is discovered, which brings about economic and social benefits. Compared with the traditional privacy’s passive protection, people begin to pay attention to the flow and use of personal data. The concept of personal data mining comes into being. According to the definition of data mining[1], personal data mining means the process of searching and analyzing personal data for previously unknown patterns and using these patterns to predict future outcomes. It is a data discovery, knowledge discovery and value discovery procedure[2].

Personal data mining adopted in this paper is a broad sense, including a) data collecting stage, in which personal data is selected from relevant data sources and integrated into data sets for mining; b) mining process, in this period, data processors use big data technology to analyze collected data for efficient analysis; c) result representation procedure. This is the output stage of data mining. In-depth data sets will be used for social and economic requirements. Personal data mining is different from the complete openness of personal data. In the process of mining, we must pay attention to the preservation of personal data, so as to achieve the balance between the mining and protection of personal data. Therefore, it is a kind of protective data application.

3. Legitimacy of Personal Data Mining: Analysis of Legal and Practical Perspective

The Legitimacy of personal data mining is the premise of the technology application, as well as the basis of the regulatory system of personal data mining.

3.1. The common goods attribute and sharing characteristic of personal data are the internal precondition of personal data mining

Personal data has certain public goods attribute and sharing value. Goods can be divided into two categories: private goods and public goods. Public goods is also known as common goods and is non-exclusive and non-competitive. Thus, it can be shared by many people. The cost of the diffusion utility of common goods is zero, which means that others cannot be excluded from sharing. Personal data is generated with the birth of a natural person. Unlike privacy emphasising on the peaceful of physical entities, personal data is intangible. It is reflected in the two-dimensional electronic code recorded in the Internet system. For example, personal identity data is generated with the birth of an individual and recorded in the national identity management system upon household registration. State organs such as public security departments can obtain such kind of data in some circumstances. Therefore, personal data isn’t exclusive to individual, but has natural sharing characteristic[3]. Meanwhile, personal data, such as personal web browsing records, online shopping traces, travel preferences, will be generated continuously with people’s daily behaviours and collected by big data technology, which is not scarce and can be shared by multiple parties.

3.2. Data appreciation theory is the theoretical basis of personal data mining

Data appreciation theory is an important theory in economics, including the increment about data quality, quantity and value[4]. It is a purposeful, non-linear and efficient increased process[5]. To be specific, the increase of data quantity means the increment of data measurements. And the qualitative increment refers to the improvement of data utilization efficiency. When it comes to the increment of data value, it turns to be the satisfaction degree of data application. Moreover, data appreciation also brings about the knowledge expansion, economic growth and social progress.

Personal data is a kind of important data resource, having in-depth mining appreciation value. Through big data mining technology, such as the cloud computing, distributed database, cloud storage, derived data and derived data relations are generated, which can be used to detect business trends, prevent the spread of diseases, fight crimes and determine real-time traffic conditions etc. It has great economic and social value. Data mining technology develops the original individual electronic code into deep utilizable data. It generates series of digital industries and promotes the development of
sharing economy driven by data. Meanwhile, data resources are widely used in social and public affairs, which promotes the construction of digital government and smart cities, as well as produces great social benefits. Both the development of data economy and the realization of intelligent society need to rely on the mining of personal data. As a result, personal data is no longer limited to individual identification and privacy protection. If personal data mining is legally forbidden, the huge benefits will be reduced, even vanish.

3.3. The principle of interests balance is the logical requirement of personal data mining

Mostly, we talk about interest in philosophy category, it means the satisfactory and value that the object to the subject. According to the different objects of interests, personal data interests can be divided into material interests and spiritual interests, physical interests and property interests. However, depending on the subjects of interests, personal data interests include individual interests, public interests, and third-party interests. Most of the time, interests can’t be simply corresponded. One object may correspond to multiple subjects of interests, and a certain interest subject may also involve different kinds of interests, thus forming a complex intersection relationship. Commonly, there are some contradictions between different interests subjects on interests division, which need to be clarified and divided, which means the interests should be sorted and balanced. Law is often referred to as the “instrument of interests”, which identifies and divides interests according to the value and the social production relations they represent.

As a new resource in the digital age, personal data is a collection of interests. First of all, personal data has identification value and the individuals can be identified through personal data. At this time, the interests behind are the dignity and freedom of human, such as privacy and information self-determination, which are kinds of self-interested and introverted personal and spiritual interests[6]. Moreover, personal data interests include material and economic interests for it is widely used in economic improvement. Lots of different interests parties are included, first of all, the government and related departments, they reflect common interests’ parties. We mention about such common interests in some public affairs, such as the construction of intelligent city, the protection of human health. In addition, personal data interests is closely related to some third-party interests, mainly about economic interests.

3.4. Sharing economy provides practical soil for personal data mining

In recent years, the global sharing economy has grown unprecedentedly, with an annual growth rate of 30% to 40%, accounting for an increasing proportion of global GDP. Since 2014, China’s sharing economy has seen a blowout development. In 2017, the transaction volume of China’s sharing economy is about 4920.5 billion Yuan, an increase of 47.2% over the previous year. It is expected that the transaction scale will reaching more than 10% of GDP by 2020. The core of sharing economy is cooperation and benefits sharing, which provides the economic foundation of personal data mining. It emphasizes the ability of control rather than property, which helps to optimize competition and promote economic creativity. Sharing economy is based on the big data technology, especially the personal data mining process. The idle resources, shared platform and shared participants are the three basic elements of the sharing economy. The idle resources are reused by the sharing participants through third-party sharing platforms. In this process, personal data mining technology helps people of different nationalities, races and requirements to communicate and connect conveniently. The supply and demand of resources can be satisfied based on massive data and data analysis. Didi, the famous sharing travelling company, integrates users’ personal data though its registration platform to connect users’ real-time positioning with the private car data, so as to realize intelligent travel. What’s more, trust economy is one of the core elements that cannot be ignored in the sharing economy. It realizes the transformation from acquaintance transaction to stranger transaction. In addition, it has solved the problem of information asymmetry in the transaction process, which helps to form relationship thinking. Relationship thinking requires buyers and sellers to establish relationships and realize point-to-point transactions through the shared platform, so as to make the communication efficiently and
smoothly. Both sides can understand each other’s situation through data, so as to realize trustable relationship and resources sharing between strangers. Sharing economy, to some extent, has changed the original property system to the use and control style. Therefore, the mining and using of personal data get much attention than before.

4. The Regulatory Approaches of Personal Data Mining: Some Strategies for Promoting China’s Personal Data Mining System

4.1. Establishing the purpose of combining personal data mining and protection
The purpose of mining determines the legislation directions and regime of personal data mining. In the small data era, data resources are limited and centralized. Data owners tend to keep data privately. People are unwilling to make their behaviours and intentions publicly. While in the big data age, the opening and sharing of data become inevitable. Big data technology has brought about huge changes to the whole society, as well as human’s thinking style and digitization has become a kind of social movement. People’s daily behaviours are recorded and personal data is generated every now and then, such as searching traces collected by Baidu, Google and other search engines, shopping and browsing preferences of Taobao, Jingdong and other e-commerce platforms, travel records gathered by Didi, Uber and other smart transportations[7]. People enjoy all kinds of intelligent and data-oriented services, and they are also willing to share personal data and become contributors, diggers and enjoyers of data resources. Only by insisting mining of personal data can we better promote the improvement of data era.

As all new technologies, personal data mining causes certain risks, such as privacy infringement, data leakage, distribution problems of data interests, etc. In the small data era, the amount of data is limited and digital relationship is weak. Therefore, it is easy to monitor data application and solve data security problems. But in the age of big data, it has changed sharply. We should insist encouraging personal data mining as well as emphasizing the protection of personal data. Certainly, it is a systematic project which can’t be completed by a single individual[8]. It relies on related parties to work together. Moreover, interests of personal data should be reasonably distributed and this is a vital topic in personal data mining process. In addition, improving personal data legislation and continuously innovating data technology are also important.

4.2. Balancing the public interests, personal interests and third-party interests of personal data mining
As mentioned above, the personal data itself has certain public goods attribute as well as closely related with public interests. Meanwhile personal data represents personal interests. According to Hegel’s theoretical analysis on the classification of the internal and external world, personal data can be used to identify individual and is inseparable with individual, so we should attribute it to the spiritual category[6]37. Moreover, personal data has the in-depth mining value which is closely related to third-party interests. The mining of personal data is bound to involve conflicts and contradictions between public interests and personal interests, as well as third-party interests. It is necessary to coordinate and balance different interests, so as to resolve conflicts of interests and social contradictions.

Public interests reflect the value of the whole society, representing the common sense. Its legitimacy comes from universal rights. All around the world, the majority of countries in different periods have given priority to the protection of public interests. In some extent, only by maintaining public interests as the basis can we achieve the maximum value. Therefore, we should encourage personal data mining for public purpose[9]. Meanwhile personal interests are the basis and final destination of public interests, so we can’t ignore the personal interests contained in personal data mining[10]. In addition, personal data based on third-party interests should be selectively mined. The legitimate and reasonable third-party interests should be mined without conflict with public interests and personal interests. However, it is necessary to strengthen the self-discipline and supervision, so as
to prevent the third party from harming public interests and personal interests. In a word, we should
give full consideration to the public interests, personal interests and legitimate third-party interests, so
as to achieve the unity and balance of the above three.

4.3. Improving data classification legislation
In order to promote the mining and protection of personal data, it is necessary to classify personal data
and improve data classification legislation. From the perspective of content, personal data can be
classified to personal identity data, personal social data and personal privacy data. Personal identity
data, including personal name, phone number, home address and other direct external identification
data, is the basic outward sign of people. Personal social data is also an important component of
personal data, which includes personal occupation, resume, social status and so on. It reflects the
social evaluation of human beings. Personal privacy data, including intelligence level, gender
orientation, medical history, genes and other potential expressions, which is closely related to personal
privacy, is difficult to identify and speculate from external representations. Comparatively speaking,
some personal identity data is automatically recorded at birth in official systems, such as the national
household registration information system. Although this kind of data is private, it is not absolutely
unknowable. The widely use of big data technology increases the exposure of such data, for example,
persons need to provide their phone numbers when registering for e-commerce platforms. It is easy to
access, but closely related to the privacy of the entity. Therefore, it should be mined carefully and
lawfully. Apart from personal identity data, personal social data can also be mined, it has extroversion
and sociality meaning, and is one of the approaches for people to conduct external communication.
Moreover, it has a positive influence in urging judicial execution and should be mined except
malicious mining goal. Personal privacy data, which involves a large amount of sensitive personal data,
should be mined with restrictions. It must be explicitly agreed or authorized by the parties concerned,
and the data should be processed anonymously. Personal identity data, personal social data and
personal privacy data together constitute personal data, which comprehensively and deeply express
individuals from different aspects. Personal data should be mined in terms of content, and all of them
need to be processed anonymously.

Personal data is easily to be collected and recorded in the digital age. Therefore, according to the
disclosure condition, personal data can be divided into disclosed and non-disclosed data. The former is
already known to the public or belongs to the public data category. While the latter is still in un-
obtained state. To former, collecting and using are legitimate. For the latter, it is necessary to classify,
identify and obtain according to the content of personal data.

4.4. Innovating technical means to achieve privacy protection: “de-identification” and strict control
of “re-identification”
Personal data mining is not only a legal theme, but also a technical one. Data has natural tool
neutrality and technology neutrality. It is important to adopt some technical principles to protect the
mining of personal data.

The National Institute of Standards and Technology has proposed the concept of “Deidentification”,
which means the process used to prevent a person’s identity from being connected with data using. It
asks for data controllers and users to avoid the identification of individuals by deleting or changing
their identifiable data in the data sets. This technology, known as “Anonymous Information” in the EU,
aims to permanently and completely remove personal identification elements from the data sets, thus
make data users unable to associate with personal data in any way and remove all potential
identification threats. In a broad sense, personal identity data includes direct identity data and indirect
identity data. Direct identity data refers to the fact that an individual’s data can be identified without
any contact with other data, such as name, ID number, etc., which must be eliminated obviously.
Indirect identity data itself cannot be identified individually, such as personal zip code, birthday,
gender. This kind of data is not recommended to be removed directly, but can be combined with
specific scenarios and data application purposes for judgment before removal or retention.
In some extent, personal data de-identification technology can prevent the privacy risk in data utilization, which needs to be supported and refined at the technical level and legalized. Of course, personal data that is removed from identification is not absolutely unrecognizable, and re-identification at the technical level is common. The re-identification of data needs to be regulated, and the legal responsibility of the re-identification subjects should be strictly regulated. Of course, it is necessary to balance the conflicts of interests of involved subjects and make choices based on specific data application procedure.

5. Conclusion
Personal data mining is legitimate in the big data era. Firstly, it has public goods attributes and sharing value, composing the internal precondition of legitimacy. Secondly, the theory of data appreciation and interests balance constitutes the theoretical basis and logical requirement of personal data mining. Thirdly, the big data sharing economy provides practical circumstances for personal data mining. Meanwhile, it is very important to choose rational approaches and adopt some strategies for regulating China’s personal data mining system. First of all, the purpose of combining personal data mining and protection should be established. Moreover, the principle of balancing public interests, personal interests and interests of a third party should be followed. In addition, we should pay attention to improve data classification legislation. Last but not least, innovating and improving technical means to achieve privacy protection need to be taken seriously. Certainly, the ethical and legal risks of personal data mining are extremely sharp, such as the challenges to privacy protection, the leakage of personal data, the relief of data infringement, etc., which are important parts in the regulation of personal data mining. Due to time and length limitation, these parts are not included in this paper but will be studied later. Appropriate regulatory approaches are the basis of the technical innovation and will promote the development of the technology. Personal data mining is an important theme both in technical and legal realms, which deserves our attention and continuing study.

Acknowledgments
The author is most grateful to the Law School of Xi’an Jiaotong University. The research is funded by the projects of “Promoting the Construction of the Silk Road Economic Belt” (14JZD022) and “Research on Governance Innovation in Network Society” (15ZDA047).

References
[1] George Gutierrez. The inbalance of security&privacy: what the snowden revelations contribute to the datamining debate[J]. Intellectual Property Law Bulletin, 2015(19):161-162.
[2] Xiang Liling, Yang Hua. The revolution and evaluation on the theory of personal data mining and protection[J]. Information Studies: Theory & Application, 2017(11):90-95(in Chinese).
[3] Wu Weiguang. Criticism on personal data privacy protection under big data technology[J]. Political Science and Law, 2016(7):119(in Chinese).
[4] Xiang Liling, Yang Hua. The multi-disciplinary theory of personal data development and protection[J]. Information Studies: Theory & Application, 2016,2(39):24(in Chinese).
[5] Zhu Jianwei. The analysis of information appreciation theory: the prospective of information economics[J]. Library Theory and Practice, 2008(3):52(in Chinese).
[6] Qi Aimin, Liyi. The regulation of personal information protection: the balance of personal interests and the freedom of information spreading[J]. Legal Forum, 2011(22):42(in Chinese).
[7] Huang Xinrong. Ethical issues of big data technology[J]. Journal of Xinjiang Normal University(Philosophy and Social Sciences), 2015(5):47(in Chinese).
[8] Cao Yujia, Huang Weiqun. Construction of Government Data Open Ecosystem: a perspective of data security[J]. Library Theory and Practice, 2016(10):20(in Chinese).
[9] Zhang Xiaojuan, Wang Wenqiang, et al. Research on policies and regulations of open government data and personal privacy protection in the U. S. and China[J]. Information
Studies: Theory & Application, 2016(1):40-42(in Chinese).

[10] Wang Jingbo. Measurement of government information opening: the prospective of public interests[J]. Social science in China Press, 2014(9):110(in Chinese).

[11] Simson. L. Garfinkel. De-Identification of Personal Information[EB/OL]. Information Technology Laboratory National Institute of Standards and Technology(Oct. 2016), 1-27. http://dx.doi.org/10.6028/NIST.IR.8053.