RESEARCH ARTICLE OPEN ACCESS

Research on Privacy Protection in Big Data Environment

Gang Zeng

(Police Information Department, Liaoning Police College, Dalian, China)

Abstract:
Now big data has become a hot topic in academia and industry, it is affecting the mode of thinking and working, daily life. But there are many security risks in data collection, storage and use. Privacy leakage caused serious problems to the user, false data will lead to error results of big data analysis. This paper first introduces the security problems faced by big data,analyzes the causes of privacy problems， discussesthe principle to solve the problem. Finally,discusses technical means for privacy protection.

Keywords: big data, privacy protection, causes, principle

I. Introduction

Today, the development of informatization and networking lead to explosive growth of data. According to statistics, 2 million users are using Google’s search engine in every second, Facebook users share 4 billion resources every day, Twitter process 340 milliontweets every day, At the same time,the large amount of data are produced continuously in scientific calculation, medical services, finance, retailing. 8ZB data will be generatedin 2015.

This phenomenon caused the wide attention of people， in the academic circles,Turing award winner Jim Gray proposed fourth scientific research paradigm, research on data intensive science based on big data, the journal Nature published a special issue discussing big data in 2008. The journal Science published a similar issue about data processing. More activities were carried out in IT industry. The reuse of big data was focused oncontinuously, the potential value of big data are mining.

At present, the development of big data still faces many problems, Security and privacy issues is one of the key issues that people recognizedwidely, currently,people’s every word and action on the internet are recorded by businesses,including shopping habits, friends contact situation, reading habits, searching habits, etc.Number of cases show that even after a large number of harmless data is collected, personal privacy will be exposed. In fact, the security implications of big data are more widely,the threat people faced, is not limited to leak of personal privacy, like other information,big data is facing many security risks during storage, processing, transmission, etc. and it needs data security and privacy protection. But data security and privacy protection in big data era is more difficult than in the past(such as data security in cloud computing, etc.). In the cloud computing, the service providers can control the storage and operation of data.

However, users still have some way to protect their own data, such as data storage and security computing through technical means of cryptography, or operational environmental security through trusted computing mode.And in the context of big data,Many businesses not only is a producer of data, but also store managers and users of data, Therefore,only bysimple technical means to limit the businesses to use user’s information,user privacy protection is extremely difficult.

Currently, many organizations have realized the security problems of big data, and take action tofocus on big data security issues.In 2012, the cloud security alliance (CSA) formed a big data working group, aimed at finding solutions for data center security and privacy issues.

In this paper, based on carding the research situation of big data, analyzesthe security challenges to big data, discusses the key technology ofthe current big data security and privacy protection.

II. Whydoes Big Data Threat Personal Privacy

2.1Connectivity of Social Network

In our social activities, there is often the case : Social networking sites recommend some people you may know to you.Why is there such a situation?As our society has connectivity.If you know Tom, Tom know Lucy, then can be speculated that you may know Lucy.Computer has a massive user information，by analyzing any common social networks of two users, or by reading the phone contacts to determine whether acquaintance between two users.Although individual users can set to turn off the reading function of social networking sites, as long as users use social networking sites,he will leave marks - logs, status, messages, and even point praise, connection between user and the whole social network is established, there is the possibility of associated with other users in the network.

Most Internet users do not pay attention to personal privacy, or the concept of privacy is quite weak.CNNIC's"2012 China Research Report on usersof Internet social networking site"found that: For social networking sites use personal information for a commercial purpose,More than half (51.5%) of users said as long as no external leakage of personal information, they can accept this behavior, 22.6% said it does not matter.

For mobile social networking applicationsrecommend friendsby reading phone contacts,about 1 / 3 of the users said he could not accept this behavior, 1/5 of the users said it does not matter,more users said they can accept it only committed to protect personal privacy. This shows that a considerable number of users do not entirely resist the practice of reading the phone contacts.

2.2Commercial Interests

In the browser world, browsing history and Cookies is collectedgenerally. In IE10 browser released by Microsoft,the default setis "Do Not Track (DNT)", this behavior clearly violetes the interests of Internet advertising industry,thus, World Wide Web Consortium affirmed that the setting does not meet their standards, so website can ignore DNT signal sent by IE10 ,continue to track and collect user information.

Many operators record the user's scene and behaviorfor a long time, and labelthe user characteristics，analyze the possible behavior habits and needs, and thenpush advertising information in a range of relatively obscure user groups.

2.3Need forPublic Power

In order to meet the needs of law enforcement, many countries in the world usually require network or telecom operators to store certain user datain a certain period of time, and provide the raw data and the resultswhen the government need. This requirement is certainly legitimate, and does not pose a great threat to personal privacyin the era of the small data.However, in theera of big data,information communication capacity of the network increases rapidly, the data can reflect the personal background, characteristics, habits, behavior, becomes more and more specific,once this information is abused by public authority in the absence of supervision, it does exist the possibility that personal information has security risk.

III. The Main Principles of Privacy Protection

In era of big data, the focus on privacy issuesshifted to the users.Only to regulate users’behavior，their actions are consistent with theprofessional norms of the big data industry practitioners,the protection of personal privacy is possible.

3.1The Principle of theCertain UsingScope of Data

The goal of handling of personal information must be specific, clear, reasonable, does not expand range of use, not change the purpose of the use of personal information when the owner of the information do not know.This principleis more difficult to do，but we can use "negative list", we stipulate what kind of behavior is not allowed, at the time of collection and use of data, Because these action encroach on personal privacy, as long as not to touch the place, other behavior ofdata usingare acceptable.

3.2The Principles of Quality Assurance

Information managers must ensure that the processing of personal information is confidential, complete, available and up to date. and need to establish internal control mechanisms to protect personal information, and regularly detect security, protection and the implementation ofinformation systems, measuredby themselves or an independent evaluation agency,to develop plans for loss, damage, tampering, improper use and other events during processing; When wefind that personal information has been leaked, lost, after tampering, response at once to prevent the incident further expansion,and promptly notify the affected message body; When a major event occurs, promptly inform the data protection authorities.

3.3 The principle of individual participation

Individuals have the right to decide whether their data is collected, knowing what data is collected, to confirm the data can be collected, modifiedanddeleted.Personal information is divided into two types: general information and sensitive information. Sensitive personal information may include ID number, phone number, race, political views, religious beliefs, genes, fingerprints, etc.By default, general information can be collected, but before the collection and use of sensitive personal information,firstly,the user must obtain the consent.

IV. Key Technologies of Privacy Protection

In big data environment, privacy protection technology is mainly studied from the following perspectives: user privacy protection, data content verifiable, and access control.

4.1 AnonymityData Protection Technology

In the big data environment, anonymity protection is necessary to protect the data. For example, in social networks, anonymity protection can be divided into user identity anonymity, attributes anonymity and relationshipanonymity(known as edgeanonymity).Theinformation of user identification and user attributemust be hidden when published, the relationshipanonymity is to hide the relationship between users when data is released.

At present, the relationship anonymity is a hotspot of research, many scholars have studied multiple methods for the relationship anonymity.Through other public information,an attacker may beinfer anonymous users, especially relationship between the users.

Agglomeration characteristics of social networkhas an important influence on the accuracy ofprediction about the relationship between users, with the connection density growthand agglomeration coefficientenlargement in local social network, the accuracy of the predictive algorithm for connection of the users is further enhanced. Therefore, in the future the anonymity protection technology should beeffective against such speculative attacks.

4.2Data Watermarking Technology

Digital watermarking refers to the identification information is embedded imperceptible within the data carrier and does not affect the method of its use, usuallyused for copyright protection of multimedia data,there is also a watermarking scheme for databases and text files.Due to the characteristics of randomness and dynamic data, watermarking methods are very different on the marked database, document and multimedia files.

The basic premise is that there is redundant information in the data, or can tolerate a certain precision errors.If the fragile watermark embedded in the database table, it can help to detect changes in data items.

There are many types of the watermark generation methodon text, it can be roughly divided into watermark based ondocument structure, watermark based on text-based content. Small changes of the character spacing and line spacingwill cause changes in the structure watermark, add modification ofspaces and punctuation will cause changes in the content watermark.

Robust Watermark can be used to prove the origin of big data. Fragile Watermark can be used to prove the authenticity of big data. One problem is that the current scheme is based on static data sets, but, without taking into account the generation and update in high-speed, which needs to be improved in the future.

4.3Data ProvenceTechnology

Due to the diversification of data sources,it is necessary to record the origin and the process of dissemination, to provide additional support for the latter mining and decision.

Before the emergence of the concept of big data, Data provence technology has been widely studied in database fields.Its purpose is to help people determine the source of the data in the data warehouse.

The method of data provenceis labeled method, through the label, we can know which data in the table is the source, and can easily checking the correctness of the result, or updatethe datawith a minimum price.

In the future data provence technology will play an important role in the field of information security. But Data provence technology for big data security and privacy protection also need to solve the following two questions: 1, The balance between privacy protection and data provence;2,to protect the security of data provence technology itself.

4.4 Access ControlTechnology

4.4.1Role Mining

Role-based access control (RBAC) is an access control model used widely.By assigning roles to users, roles related to permissions set, to achieve user authorization,to simplify rights management, in order to achieve privacy protection.In the early, RBAC rights management applied "top-down" mode: According to the enterprise's position to establish roles.

When applied to big data scene,the researchers began to focus on "bottom-up" mode, that is based on the existing "Users - Object" authorization, design algorithms automatically extract and optimization of roles, called role mining.

In the big data scene, using role mining techniques， roles can be automatically generated based on the user's access records，efficiently provide personalized data services for mass users. It can also be used to detect potentially dangerous that user’sbehavior deviates from the daily behavior.

But role mining technology are based on the exact, closed data set, when applied to big data scene, we need to solvethe special problems: the dynamic changes and the quality of the data set is not higher.

4.4.2Risk Adaptive Access Control

In the big data scene, the security administrator may lack sufficient expertise,Unable to accurately specify the data which users can access, risk adaptive access control is anaccess control method for this scenario.By using statistical methods and information theory, define Quantization algorithm, to achieve a risk-based access control.At the same time, in the big data environment, to define and quantify the risk are more difficult.

CONCLUTION

This paper first introduces the security problems faced by big data, discusses the reasons of privacy problems,then, discusses the principles to address privacy issues,finally, from four aspects discusses the technology to solve the problem of privacy protection.At present, although there have been some methods to solve the problem of privacy protection, but research is not enough, only combination of the technical and legal means can solve the problembetter.

REFERENCES

[1] Feng Deng-Guo, Zhang Min, Li Hao. *Big Data Security and Privacy Protection[J]*. Chinese Journal of Computers,2014,14(1):246-258.

[2] Liu Yahui, Zhang Tieying, JinXiaolong,ChengXueqi.*Personal Privacy*

Protection in the Era of Big Data[J]. Journal of Computer Research and Development, 2015,15(1):229-247.

[3] Chen ChangFen, yuxin. *Privacy Protection in the Era of Big Data[J]*, News and Writing, 2014,6:44-46.

[4] Viktor Mayer-Schonberger, Kenneth Cukier. *Big Data: A Revolution that Will Transform How We Live* , Work and Think. Boston: Houghton Mifflin Harcourt, 2013.

[5] Bu Ying-Yi,Fu Ada WaiChee,Wong Raymond Chi Wing,et al. Privacy preserving serial data publishing by role com-position//Proceedings of the 34th International Conference on Very Large Data Bases(VLDB'2008).Auckland, New Zealand, 2008:845-856

[6] Ying X, Wu X. Randomizing social networks: A spectrum preserving approach//Proceedings of the SIAM International Conference on Data Mining (SDM'08).Georgia, USA, 2008:739-750

[7] Zou Lei,Chen Lei, zsu M T.k-automorphism:A general framework for privacy preserving network publication//Pro-ceedings of the 35th International Conference on Very Large Data Bases(VLDB'2009).Lyon,France,2009:946-957

[8] Hay Michael, Miklau Gerome, Jensen David, et al. Resisting structural re-identification in anonymized social networks//Proceedings of the 34th International Conference on Very Large Data Bases(VLDB'2008).Auckland, New Zealand,2008:102-114

[9] Zhang Li-Jie, Zhang Wei-Ning. Efficient edge anonymization of large social graphs. http://venom.cs.utsa.edu/dmz/techrep/2011/CS-TR-2011-004.pdf.2013-06-10

[10] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks.Nature,2008,453(7191):98-101

[11] Cui Y, Widom J. Practical lineage tracing in data warehouses//Proceedings of the 16th International Conference on Data Engineering(ICDE'2000).San Diego,USA,2000:367-378

[12] Chen M Y,Yang C C,Hwang M S.Privacy protection data access control [J].International Journal of Network Security,2013,15(6):391-399

[13] Ayenson M, ambach D, Soltani A, et al. Flash cookies and privacy II: Now with HTML5 and etagrespawning [OL].(2011-07-29)[2013-02-13].http:??ssrn.com?abstract=1898390

[14] Banisar D, Davies S. Global trends in privacy protection: An international survey of privacy, data protection, and surveillance laws and developments[J]. Journal of Computer &Information Law,1999,18(1):3-111

[15] Warren S D, Brandeis L D. The right to privacy [J].Harvard Law Review,1890,4(5):193-220