Network Information Security Privacy Protection System in Big Data Era

Lei Ma(✉), Ying-jian Kang, and Jian-ping Liu

Telecommunication Engineering Institute, Beijing Polytechnic, Beijing, China
malei235@tom.com

Abstract. Traditional information security protection is based on an information data set, at least one information can not be distinguished from its own location information. Therefore, this paper studies the network information security privacy protection system in the era of big data. The hardware of network information security privacy protection system is composed of independent monitoring layer, host layer and mixed layer. It disturbs the original data by adding random numbers and exchanging, shields the original data to unauthorized users, and achieves the purpose of privacy protection and recommendation accurate and non-destructive. The system software encrypts information according to the degree of privacy protection set by users, adopts the key management mode, solves the problem of communication security and node key update, and realizes the network information security privacy protection system.

Keywords: Privacy protection · Information security · Information encryption · Key management

1 Introduction

In the era of big data, the privacy protection of network information security needs to be constantly updated and developed. There are still many problems that can be further studied and discussed. Below is a brief introduction to some classical location privacy protection systems. Literature [1] puts forward that in an information data set, at least one information cannot be distinguished from its own location information, so it can be said that this location information and other location information meet technical requirements. The system can obtain the real location of the attacker, so as to protect information security and privacy. However, this method cannot achieve the privacy protection accuracy in this field. Literature [2] proposed a protection method and corresponding safe data exchange model for protection and control of information transmission risk, and established a simulation platform for this purpose. Simulation results show that the transmission risk of protection control information can optimize the message processing mechanism under normal circumstances and ensure the reliable transmission of key messages under abnormal circumstances. However, the implementation process of this model is complex and cannot be widely applied.

In this paper, a privacy protection system for network information security in the era of big data is proposed. The method of generalized concealment of coordinate

G. Gui and L. Yun (Eds.): ADHIP 2019, LNICST 301, pp. 65–73, 2019.
https://doi.org/10.1007/978-3-030-36402-1_7

positions hides the user's position into a region, where there are many location nodes. After this generalized concealment process, the whole region is regarded as the user's position. In this way, privacy attackers can not query the location information of specific users. Continuous mobile users initiate location queries at different times to form a trajectory. By attacking the trajectory information, we can infer the user's daily life trajectory and habits, so that the privacy information is broken. In order to protect the security of information network, it is necessary to establish corresponding information security protection system. The means of information security protection mainly include identity authentication, information encryption, intrusion detection, boundary integrity detection, etc. Traditional information security protection systems usually take protective measures from a certain aspect, which can not constitute a defense system in depth, leaving an opportunity for attackers to take advantage of. The present information security protection system pays more attention to multi-level comprehensive defense, establishes the defense system in depth, and improves the reliability of information security protection.

2 Hardware Design

Because virtualization technology has changed the computer architecture, it provides a new solution to traditional security problems. As mentioned above, some studies have introduced virtualization technology into the field of security, using the characteristics of virtual machine manager to achieve security functions. However, there are some obvious problems when the Universal Virtual Machine Manager provides security services for clients. There are two main aspects: First, the Trusted Computing Base of Universal Virtual Machine Manager is huge. Trusted Computing Base (TCB) refers to the collection of all security protection mechanisms for the realization of computer system security protection, which can appear in the form of hardware, firmware or software. Once a vulnerability occurs in one part of the trusted computing base, it will seriously endanger the security of the whole system. On the contrary, if there are vulnerabilities in other parts of the trusted computing base, the harm to the whole system is relatively small. Thus, the trusted computing base is very important to the overall security of the system. Virtual Machine Manager (VM) is a layer of software between computer hardware and operating system. It is responsible for managing system resources and providing isolated running environment for multiple VMs in the upper layer.

2.1 Independent Monitoring Layer

The VMM of the independent monitoring layer runs directly on the bare machine and has the highest privilege level. It is responsible for the management of the underlying hardware resources. All access to the real hardware of the client operating system must be completed through VMM. When users get personalized recommendation, they can use the system anonymously without requiring their real identity, and they can also get personalized information. The implementation of independent monitoring layer needs to ensure that each tuple can not be distinguished from other tuples, and attackers can

not judge the owner of privacy information, so as to ensure the security of users' personal privacy [3]. Usually the system allows a user to enter with multiple identities, which can protect the user's identity in different activities. It can be divided into two categories: concealment and generalization. Hiding is to protect users' privacy attributes by cutting off the relationship between privacy attributes and non-privacy attributes. Require ordered values in their own data tables not less than the prescribed occurrence rate. Some scholars have proposed an anonymity scheme based on adding, deleting and increasing noise nodes. By adjusting the degree of nodes, the anonymity of node degrees can be realized. By assigning different sensitive attribute values to noise nodes, the number of occurrences of sensitive attributes can be adjusted to achieve attribute anonymity. Generalization refers to dividing local user attributes into several equivalent classes and publishing their generalization attributes for different equivalent classes [4].

2.2 Host Layer

All access to the hardware of the host layer client operating system needs to go through VMM and then through the host operating system, and the association rules must be obeyed. As one of the most important methods in data mining, association rule mining has also achieved some research results in privacy protection, which can be used in personalized services based on Association rules. The basic strategies of privacy protection in association rules are data interference and query restriction. Data jamming strategy is to pre-transform the original data according to certain rules. It disturbs the original data by adding random numbers and exchanging, shields the original data for unauthorized users, and then runs data mining algorithm on the jammed data to get the required patterns and rules. However, the technology needs to ensure that the disturbed data can meet the needs of relevant applications, and ensure that the data is not distorted. Query restriction strategy is to change the support and confidence of specific rules through data hiding, and then use probability and statistics or distributed computing methods to get the desired mining results [5].

2.3 Mixing Layer

Hybrid layer is based on the interest of similar user groups to generate recommendations to target users. It only relies on the user's score matrix for the project. Therefore, it has good adaptability to various specific applications and can improve the scalability and recommendation quality of personalized systems. Mixed layer is mainly divided into two categories: data encryption and data transformation [6]. Among them, data encryption is a common security measure. Based on the cryptography principle, it realizes the invisibility and lossless of the original data, and achieves the purpose of privacy protection and accurate recommendation. The main idea of data transformation is to disguise or slightly change users' real privacy data without affecting the use of the original data. Common data transformation includes random perturbation method and data geometry transformation method. Mixed layer can protect privacy information in data very well, but because of the role of social relations, users with unmarked attributes may also be inferred to have some privacy attributes, which can not be directly

used to protect the privacy security of users in personalized recommendation. In attribute social networks, we must take into account the influence of social structure information on attribute distribution and the characteristics of attribute distribution itself, in order to better achieve the goal of attribute privacy protection [7].

3 Software Design

3.1 Information Encryption

Information encryption can protect network information security. Data can be disrupted by encryption. Only authorized managers can access the data. This is to make the data more confidential. The encryption process is to put the original data together with the key and process it by mathematical formula to get a data that no one can understand. Encrypted data is usually called ciphertext. In order to make the encrypted data readable, when the receiver receives the data, it decrypts the data with the key in the opposite way. However, the above encryption and decryption process will increase the processing time and memory occupancy cost of the computer CPU. Complicated encryption keys are more conducive to improving data security than simple keys. However, a longer key will be more complex to encrypt and decrypt, and it will take a shorter CPU time to encrypt and decrypt, and it will also increase the size of the target data. Generally, there are two types of encryption. One is symmetric encryption. The other is asymmetric encryption. Symmetric encryption is called symmetric encryption when the encryption key used is the same as the key. Symmetric encryption and decryption algorithm is simpler than asymmetric encryption and decryption algorithm. Because symmetric encryption and decryption have the same key and simple algorithm, symmetric encryption is faster than asymmetric encryption in operation speed. Therefore, symmetric encryption is suitable for encrypting and decrypting massive data, while asymmetric encryption is more suitable for small data [8]. The symmetric encryption algorithm is shown below. If the random variable x is the value of the finite set X, then the definition formula of the entropy of the random variable x is as follows:

$$h(x) = -\sum x \in Xp(x) \tag{1}$$

In formula (1), h is the probability of variable value x, and p represents an event set. When an attacker attacks the privacy information, successfully breaking the privacy information into an event set p, and the attacker successfully recognizes that the privacy information of a user is an event in the event set p, the degree of privacy protection can be measured by the attacker successfully attacking the information of the user.

The definition formula of asymmetric encryption is as follows:

$$p = \frac{1}{n+1}, n \geq m \tag{2}$$

In formula (2), p is the hidden area of event set, n is the number of neighbor nodes, and m is the location point of neighbor. According to the privacy protection degree set

by users, the hidden area can be obtained. According to the number n of neighbor nodes in the hidden area, it can be judged whether to insert the location points randomly. If $n > m$, it needs to insert $m - n$ points, and if $n < m$, m neighbor location points will be selected randomly. Because the number of coordinates of the nearest neighbor is m and the real coordinate points are added, the hidden area is composed of $m + 1$ points. When there is no background knowledge, the attacker can successfully obtain the probability of user location information [9].

3.2 Key Management

According to different application environments, the content of key management scheme is designed to ensure the security of data communication and sharing and the validity of key [10]. First, reduce the consumption of key update. In the application of key management to ensure the security of broadcast content and subscription of radio and television programs, most of them adopt group key management mode. In order to alleviate the burden of updating the key on the server side, group members coordinate and generate the group key. The burden of key generation is borne by the group members, and the burden of distributing the group key to users is also saved. By using this method of generating group key, the server is released to a great extent and the resources of group members are utilized more fully. From the point of view of software, a new software structure is proposed. This structure is used to encrypt the generated key and the distributed key, which not only reduces the cost of key update in terms of the number of encryption operations [11]. Considering the execution time of encryption operation, hardware accelerator is used to make digital signature more time-saving. In addition, key generation, management and storage are all in hardware, which improves the security of key usage. The acceleration of encryption operation reduces the time complexity, and the key management by hardware greatly improves the security of the key. However, the limitation of hardware resources should be included in the practical feasibility study.

In view of the frequent changes of node topology in mobile wireless networks, a density function is proposed to dynamically create communication packets for multi-level network nodes based on this function. To a certain extent, it solves the problem of communication security and node key updating, but the time complexity of constructing and selecting responsible nodes is also high. The linear regression formula is used to improve the security of the key. The linear regression formula is as follows:

$$z = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \cdots \\ \delta_n \end{bmatrix}, X = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{bmatrix} \tag{3}$$

In formula (3), z represents independent variables, $\delta_1, \delta_2, \ldots, \delta_n$ represents merchant features, X represents the requirement of each feature vector, and $[X_1, X_2, \ldots, X_n]$ represents system recommendation features. The linear regression formula is used to calculate the weight of the feature in the recommendation system. The size of the

weight value represents the influence of the corresponding feature on the result. Z is the sum of all the eigenvalues of z, Z is $(-\infty, \infty)$, so the value can be compressed to the range of 0 to 1 by linear function formula, and the sample can be further divided into 0 or 1 by discrete method, that is, to buy or not to buy. First, the prediction function is constructed. The formula is as follows:

$$F(X) = \frac{1}{1 + e^{-\theta}} \tag{4}$$

In formula (4), F represents the probability of result 1. When the input value is X, the probability of output results 1 and 0 is higher. Because the input of constructing prediction function must be 0 or 1 that can be recognized by computer, it is required to convert all feature attributes into Boolean values. The concrete method is divided into the following steps: Step 1 is to count all the features and label all the feature attributes so that the feature values can be distinguished. For example, the shop area characteristic value is 3 and the shop score value is 3. They can not be distinguished according to the value, and tags can distinguish the feature attributes. Step 2: Statistics all features, list all features, establish a feature dictionary table, and then sort all features. Step 3: The string is mapped to Boolean variables. According to the dictionary table, the existence of this feature is observed in all dimensions. If there are corresponding features, the value is 1. If there are no corresponding features, the value is 0. All character features are mapped to form Boolean variables. "1" represents that the sample has this attribute, and "0" represents that the sample does not have this feature. After the above steps, the feature can be transformed into model training to achieve key management. So far, complete the network information security privacy protection [12].

4 Experimental Analysis

Through the processing of user privacy by the network information security privacy protection system proposed in this paper, the experiment will measure the feasibility of the system from function test and performance test. Functional test verifies the function of the system. Firstly, test the function of the lightweight virtual machine manager, then test the password protection function of the system by keyboard recording attack, and then test the function of the system by common web Trojan Horse attack test system. In order to analyze the advantages and disadvantages of the system, the performance loss of the system is evaluated by performance testing, and the micro-performance loss and macro-performance loss of the system are tested respectively.

4.1 System Password Protection Function Test

According to the introduction above, there are many types of keyboard record attack methods and password protection products. In order to verify the password protection function of the system, the system and the existing password protection products are tested and compared. Anti-KeyLogger Tester is one of the keylogger attack tools,

because Anti-KeyLogger Tester basically contains a variety of common Keylogger attack methods. The experimental results are as follows:

From Table 1, we can see that the network information security privacy protection system can provide general password protection similar to 360 safe box and Jinshan secret guard, and the strength of password protection is similar to that of Internet banking plug-in and QQ password box special password protection.

Table 1. Comparing test results of system password protection function

SPEC CPU2006	Native OS		Network information security privacy protection system		Running time ratio
400.perlbench	6.32 s	5.32 s	7.32 s	6.32 s	102.3%
401.bizp2	11.23 s	5.36 s	7.36 s	6.36 s	103.2%
403.gcc	16.23 s	5.31 s	7.36 s	6.31 s	100.6%
429.mcf	6.32 s	3.21 s	16.36 s	6.54 s	1006%
445. gobmk	8.36 s	11.23 s	19.36 s	16.32 s	100.5%
456.hmmer	3.61 s	13.32 s	24.00 s	25.00 s	100. 6%
464.h264ref	3.63 s	0.25 s	26.32 s	23.11 s	100.6%

Based on the above experimental results, the efficiency of information security and privacy protection (%) of different methods is mainly compared. Literature [1] method and literature [2] method are selected as comparison methods for simulation experiment. The specific comparison results are shown in Fig. 1:

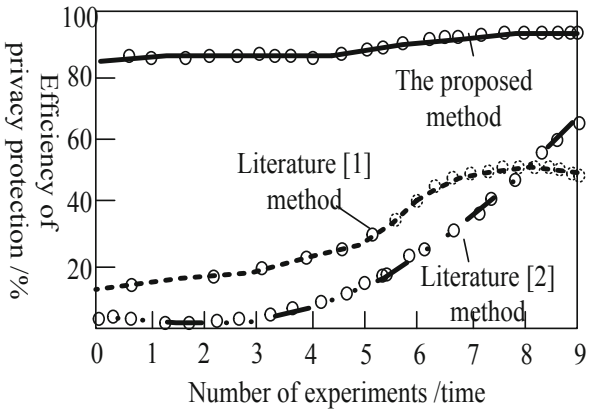

Fig. 1. Comparison results of information security and privacy protection efficiency of different methods

It can be seen from the above analysis that the efficiency of information security privacy protection of different information security privacy protection methods varies with the number of experiments. At the early stage of the experiment, the efficiency of privacy protection of each method presents a straight upward trend. When the number of experiments is 5, the efficiency of information security and privacy protection of literature [1] begins to decline, but the efficiency of information security and privacy protection of the other two methods presents a stable trend. Through the analysis of specific experimental data, it can be seen that the efficiency of information security and privacy protection of the proposed method has been significantly improved compared with the traditional method, which fully verifies the superiority of the proposed method.

4.2 System Performance Test

Hardware virtualization technology is used in network information security privacy protection system. In this paper, the performance of hardware virtualization of network information security privacy protection system is tested using standard test set SPEC CPU 2006. The performance of hardware virtualization of network information security privacy protection system is still considered by using native operating system as a reference standard. The running time of SPEC CPU 2006 in native operating system is normalized to 100%. The ratio of running time between network information security privacy protection system and native operating system is obtained as follows:

As can be seen from Table 2, the hardware virtualization cost of network information security privacy protection system is very small, and the average additional performance loss of each test is only about 1.3%. It can be inferred that, compared with the original operating system, the performance loss caused by each test of network information security privacy protection system is smaller, the average performance loss is only about 0.52%. The test results show that the overall performance of network information security privacy protection system is very good.

Table 2. Performance test results.

Keyboard recording attack method	Network information security privacy protection system	360 safety box, Jinshan secret protection	Internet banking plug-in
GetKey State	✓	×	✓
GetAsyncKeyState	✓	×	✓
GetKeyboardstate	✓	×	✓
GetRawInputData	✓	×	✓
WH KEYBOARD LL	✓	×	✓
WH JOURNALRECORO	✓	×	✓
Screenshots	✓	×	×

5 Conclusion

The exposure of personal data security issues has aroused unprecedented concern. Data security protection systems are generally designed and developed for enterprises, units, government organs and other large organizations. Personal data security systems are still implemented by encryption software alone. This method not only affects the convenience of users to operate data, but also proves that encrypted documents are not safe. This system adopts a low-cost, high-security solution to control the process of file operation and monitor the process of file operation. Considering the external storage of files and the key renewal and protection of files, it is a relatively perfect solution.

References

1. Deng, W.: Research on information security and privacy protection in the big data era. China New Commun. **19**(3), 1226–1227 (2017)
2. Wang, S., Du, W.: Progress of New Zealand's privacy protection in the big data era and its enlightenment to China. E-government **23**(11), 2165–2171 (2017)
3. Maliwei, Meng, W., Zhang, Y.: Research on personal information security in big data era. Netw. Secur. Technol. Appl. **22**(4), 1364–1365 (2018)
4. Gao, Y., Dai, G., Yan, S.: Research on information security in network environment in big data era. Inf. Syst. Eng. **22**(2), 2189 (2017)
5. Liu, Y.: Banking information security protection strategy in the age of big data network. Electron. Technol. Softw. Eng. **32**(7), 1211 (2017)
6. Liu, W.: Opportunities and challenges of network information security in the era of big data. Netw. Secur. Technol. Appl. **32**(11), 2176–2177 (2017)
7. Chen, H.: Computer information security and privacy protection strategy in the background of big data. Netw. Secur. Technol. Appl. **32**(11), 1168 (2017)
8. Zhang, Y., Wang, X.: Research on information security and privacy protection in big data environment. Digit. Technol. Appl. **32**(7), 3190–3191 (2017)
9. Zhu, X., Zhang, H., Ma, J.: Android platform privacy protection system based on hook technology. J. Netw. Inf. Secur. **32**(4), 1621–1693 (2018)
10. Yan, W., Yao, Y., Zhang, W., et al.: Logistics system privacy protection scheme based on two-dimensional code and information hiding. J. Netw. Inf. Secur. **3**(11), 2222–2228 (2017)
11. Li, C., Shi, Z., Gao, H., et al.: Development and design of personal privacy protection system for mobile intelligent terminal. Comput. Appl. Softw. **34**(6), 217–220 (2017)
12. Li, C., Zhang, Z., Zhang, C., et al.: Data fusion algorithm for privacy protection in wireless sensor networks. Inf. Secur. Res. **3**(6), 523–527 (2017)