Research on Mobile Phishing Detection Association Technology Based on Multi-Source Heterogeneous Security Data Fusion

Huanxiao Xu, Juan Qi, Lanmei Qian, Fang Wu, Jingxiong Yan

School of Computer and Information Engineering, Nantong Institute of Technology, Nantong, Jiangsu 226019, China

1109709083@qq.com

Keywords: Multi-source isomerism, Data fusion, Mobile network, Fishing detection

Abstract: With the increasing scale of network and the diversification and intelligence of network intrusion means, a defense-in-depth system based on a series of security devices such as network intrusion detection, network firewall, anti-virus system and terminal monitoring system has been established, which has exposed some problems and research trends. The extended D-S evidential reasoning method is used to realize the fusion of heterogeneous dynamic security information. According to the different characteristics of heterogeneous security information, a mobile phishing detection association technology based on multi-source heterogeneous security data fusion is proposed. The method can effectively aggregate and correlate the detection results of various security devices, and finally accurately restore the attack scene. Comprehensive use of client and server detection methods, using the flexibility of client detection and the accuracy of server detection, the hybrid wireless LAN phishing AP detection technology is realized, which further improves the detection effect compared with a single detection method.

1. Introduction

With the continuous expansion of computer networks, the information superhighway has been continuously extracted, and the requirements for network security are constantly improving. On the one hand, intrusion detection systems have problems such as high false alarm rate, overlapping alarms, missing alarms and weak alarm semantics, which bring great difficulties and misleading effects to the timely identification, analysis and response of network intrusion. In addition, with the increase of security systems (equipment), all kinds of alarm information and logs increase by orders of magnitude. The original alarm information is basically the bottom information, which is too simple and redundant, and also has the problem of false alarm. Therefore, there is an urgent need for effective heterogeneous security information analysis and processing technology.

Common ways of phishing transmission include sending spam, sending SMS by pseudo base station, sending phishing websites by QQ group, etc. The ultimate purpose is to trick people into opening phishing websites [3]. Commonly used methods include link manipulation, website forgery, free hotspots, hidden redirection vulnerabilities, and so on. The popularity of two-dimensional code makes it a new phishing media. When users scan two-dimensional code through mobile phone devices, they visit URL links in two-dimensional code, and the corresponding phishing method belongs to link manipulation. In this paper, the hybrid wireless LAN phishing AP detection technology is studied, which detects the security of wireless AP on the client and server respectively, detects twin phishing AP by using the hybrid detection method, and according to the output information, prevents users from continuing to visit phishing AP, eliminates the security threat of wireless LAN, and provides a safe wireless LAN environment for users and enterprises.

2. Overall Structure of Network Data Security Protection Mechanism

In this paper, the security protection mechanism is designed based on the visualization of the availability and performance of the network security data service, and the software part of the system is designed by using the client/server mode. Figure 1 shows the structure of the network data

  DOI: 10.25236/mmmce.2021.017

security protection mechanism.

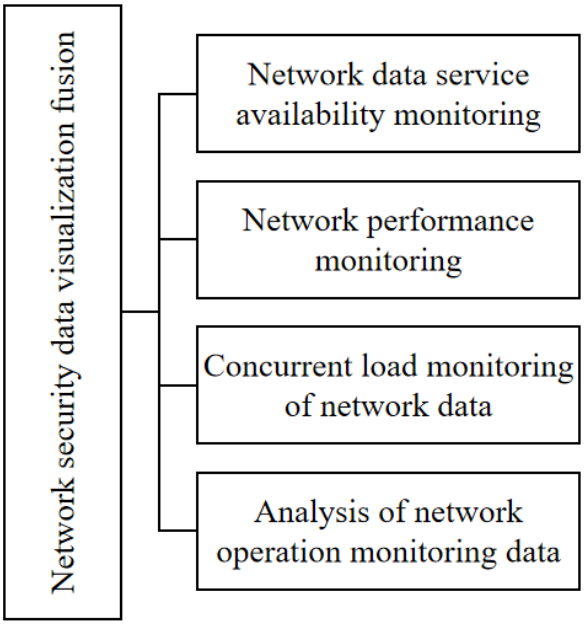

Fig.1 Overall Structure of Network Data Security Protection Mechanism

The functions of the network data security protection mechanism designed in this paper are:

(1)Effectively ensure the security of network traffic data, make statistics on the browsing of network visitors, and manage the number of times users download files, page display, network files, etc.;

(2)Ensure the stability of security data in the process of network users' operation, realize the analysis of users' online type, operating system and browser version, and ensure the stability of the system;

(3)Manage the source of users, ensure the safety of network users in the operation process, and comprehensively monitor the operation of network users [3];

(4)Maintain the security function of network data transmission. The survivability of single IP and IP segment network hosts is detected, and the detection results can show the survivability of hosts.

3. Research and Design of Mobile Phishing Detection

3.1 Alarm Fusion Based on Heterogeneous Dynamic Security Information

D-S evidence theory was put forward by A.P. Dempster in 1967, and then extended by G. Shafer in 1976. The evidence provided by different sensors is equally trusted by D-S inclusion rules. However, the actual situation is not the case. We observed that local sensors should provide more reliable information than remote sensors for the same kind of sensors; Even if the same sensor is installed in different positions of the network, it will have different detection capabilities; Different types of sensors have different detection capabilities and accuracy for attacks of the same kind of plough, and the importance and reliability of the evidence provided for fusion judgment are quite different.

In order to solve the above problems, this paper uses the extended evidence combination rule to give different weights to the evidence of different sensors, that is, to give different trust to different sensors. Literature [8] proposes to improve the evidence combination rules by proportional weighting in the research field of content perception. The combination rules are as follows.

$$m_{12}(A)=\frac{\sum_{A_i\cap A_j=A}\left(w_1 m_1(A_i)\cdot w_2 m_2(A_j)\right)}{\sum_{A_i\cap A_j\neq\varnothing}\left(w_1 m_1(A_i)\cdot w_2 m_2(A_j)\right)} \quad (1)$$

It can be proved that the combined probability distribution function does not satisfy $\sum_{A\subseteq 2} m(A)=1$; It cannot be used for evidence accumulation. Exponential weighting is used in this paper, as shown in formula (2). It can be proved that the combined evidence still satisfies the basic properties of probability distribution function by adopting exponential weighting rule (the proving process is omitted).

$$m_{12}(A)=\frac{\sum_{A_i\cap A_j=A} m_1[(A_i)]^{w_1} m_2[(A_j)]^{w_2}}{\sum_{A_i\cap A_j\neq\varnothing} m_1[(A_i)]^{w_1} m_2[(A_j)]^{w_2}} \quad (2)$$

w_2 is the weight of observed O_i, and when $w_1 = w_2$, formula (2) is simplified as a basic d-s combination rule. The weights can be estimated by training samples based on the maximum entropy criterion or the minimum mean square error criterion. Because of the lack of effective training samples, the empirical weights obtained by many experiments are adopted in this paper.

3.2 Design of Alarm Association Reasoning Rules Based on Ontology

Subordinate relationship between attack and attack scenario. This relationship is represented by their respective attributes. The attack scenario has a HasAttack attribute, and the attack has a BelongTo attribute. The two attributes have limitations and are inverse to each other, and BelongTo has Functional limitations, that is, an attack instance can only belong to one attack scenario. The reliability of an attack refers to the real probability of an attack, which is an integer from 0 to 10, corresponding to the probability of 0% to 100%. This process is implemented using SWRL rules:

Attack(? x) ^ Scenario(? y) ^ AttackReliability(? x,?reli) ^ BelongTo(? x,?Y)

^ swrlb: equal(? reli, 0)BelongTo(? x, null)

Parent-child relationship between attacks. The relationship is expressed by subClassOf(rdfs:subClassOf). It is meaningful to infer the parent-child relationship of attacks in ontology knowledge base, because most attacks have many implementations and variants, that is, the subclasses of attacks, the vulnerabilities utilized by each subclass, the principles of use, the functions realized and the behavioral impacts caused are roughly the same, so the ontology base only stores the relationships between superclasses and ontologies such as vulnerabilities, behaviors and permissions, which can greatly reduce the storage capacity of ontology knowledge base.

Comprehensive utilization of alarm correlation results based on artificial immunity and ontology can form a complete and accurate attack scene. Alarm correlation based on artificial immunity can aggregate alarms belonging to the same attack scenario, and the relationship between these attack alarms is expressed by probability. Alarm correlation based on ontology further determines the probability of attack and effectively corrects the attack scenario. For example, alarm association based on artificial immunity aggregates attack alarms A, B, C, D and E into one attack scenario.

After alarm association based on ontology, the reliability of attack alarms A, B, C, D and E are 8, 5, 0, 7 and 9, respectively, and A is an attack subclass of F, while D and E are part of multi-step attack G. At this time, the attack scenario is corrected, and the attack with reliability of 0 does not participate in the formation of the attack scenario. If there are multiple subsequent attacks, the next step is to select the one with the highest probability. The resulting attack graph is shown in Figure 1.

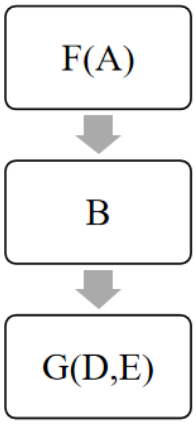

Fig.1 Attack Graph Formed by the Association of Mixed Intrusion Alarms

Risk assessment is aimed at assets, and evaluates the possibility of risks caused by threats and loopholes of assets. Risk assessment can help administrators better understand the current network security situation and existing risks. In this system, the calculation formula of risk assessment is as follows:

$$Risk=Importance \times (AttackPriority \times AttackReliability+VulnerabilityNum) \quad (3)$$

In the formula, $Importance$ represents the importance of assets, $AttackPriority$ represents the priority of attack, that is, the risk of attack, $AttackReliability$ represents the reliability of attack, and $VulnerabilityNum$ represents the vulnerability of vulnerability.

$Importance$ is designated by system administrator, $AttackPriority$ is determined by intrusion detection system and security policy, $AttackReliability$ is obtained by the above association process, and $VulnerabilityNum$ is determined by expert knowledge.

3.3 Rssi-Based Fishing Ap Detection

Learning process: create an authorized AP list by scanning wireless LAN AP. It is very useful for the server administrator to create a wireless AP list. For all authorized AP lists, it contains the MAC address, SSID, RSSI (Received Signal Strength Indication) value, etc. In this learning process, the list of authorized APS updated at any time is used, and the wireless APS we originally trusted are taken as the white list. Figure 2 shows the learning process of server detection.

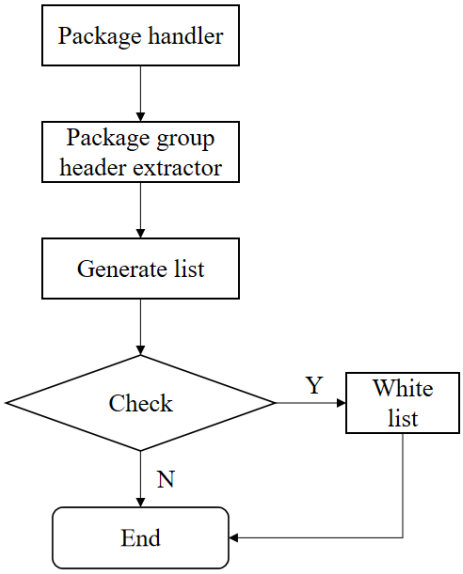

Fig.2 Server-side Learning Process

Detection process: First, detect the SSID of AP. If the system finds the same SSID, then compare the MAC addresses of two APS. If it is found that two APS have the same MAC address, it is necessary to inquire whether the white list information entered in the learning process is repeatedly entered into the same AP. And if not, detecting the RSSI values of the two APS. When the RSSI values are the same, or larger or smaller than the original value between the set thresholds, no warning will be given, otherwise a warning message will be generated to inform the administrator to blacklist the phishing AP (Figure 3).

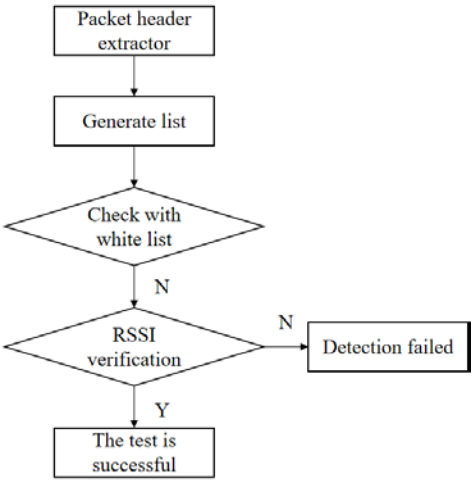

Fig.3 Server Detection Process

As for the server-side detection program, it can't connect to the wireless AP autonomously at any time and place like the client-side program. The server-side detection requires the server to conduct uninterrupted wireless monitoring 24 hours a day in a fixed range and place, scan and detect the surrounding wireless signals, add the phishing AP to the blacklist according to the detection process, and notify the network administrator to handle the phishing AP.

The phishing AP that modifies the MAC address of the wireless network card will still be bypassed only by the MAC address detection. When we detect the wireless AP with the same SSID and the same MAC address in the wireless scan, it is very necessary to use RSSI value to detect the two wireless APS.

3.4 Implementation of Association Reasoning System

The input of the association reasoning program is the normalized data of various heterogeneous data, which are converted into corresponding OWL instances by the program, in which the attack alarm of intrusion detection will generate new instances and add them to the knowledge base, while the behavior state data, system state data and system vulnerability data will modify the instances in the original knowledge base to reflect the current state. Then, according to SWRL rules, the attack is associated with user behavior, system state, system vulnerability, etc., and the reliability of each attack alarm instance is judged. Finally, the corresponding reliability value is given, and finally the attack scenario diagram and risk assessment value are output. The main part of the alarm association reasoning program is represented by pseudo code as follows:

```
PROCEDURE AlertCorrelation
INPUT:alert,ontology
OUTPUT:ontology,attackSeranio,risk
BEGIN
Load(ontology);
Bind(Pellet);
SetSystemStatus();
SetEventListener(alert);
WHILE flag!=stop DO
BEGIN
```

```
Read(alert);
OntoInstance=Convert(OntoInstance);
ConsistencyCheck(OntoInstance);
attackSeranio=OntoInstance.Seranio.TopoSort();
risk=OntoInstance.Risk;
END
END
```

When the alarm data is sent into the program, listening state awakens the processing program, first reads the alarm information, converts the alarm into the corresponding ontology instance, and then starts Pellet for consistency check, that is, reasoning. After reasoning is completed, the original empty attack scenarios and risk assessment examples in ontology library are generated, and the program results are obtained by outputting these examples. Because alarm correlation is a real-time system and may accept new reasoning tasks at any time, the program will run until the stop flag is set artificially.

The output attack scenario contains several attack examples in sequence, and the attack scenario can be represented as an attack graph more intuitively. The attack scenario can be transformed into an attack topological graph, in which nodes represent attack instances and directed edges between nodes represent the context of attacks.

4. Experimental Analysis

4.1 Effectiveness of Intrusion Identification in the Whole Process

In the course of the experiment, 16 kinds of attacks including Ftp_overflow, POP3_STAT_overflow, Sadmin_Ams lverify_Overflow, UDP_flood, Port_Scan, Large_ICMP_Packet were launched against different servers and hosts. Generate a large number of basic alarm information including IDS alarm and terminal anomaly detection alarm. For the convenience of comparison and analysis, we only use the output of the original IDS as the comparison basis for the change of alarm number. Table 1 shows the performance statistical results of alarm verification and alarm fusion process.

Table 1 Experimental Result

Contrastive item	Alarm quantity	Correct alarm	Attacks detected	Detection rate	False alarm rate	Average redundancy of alarm
Original IDS output results	102533	8214	38	81.23%	91.25%	
Results of alarm verification and aggregation	125	66	35	79.1.2%	41.36%	90.12%
Alarm fusion judgment result	68	61	36	80.03%	5.56%	92.53%

It can be seen from Table 1 that the alarm preprocessing and alarm fusion method proposed in this paper can effectively realize the accurate detection and judgment of intrusion, greatly reduce the false alarm rate of intrusion detection system, and make the final output intrusion result authentic; At the same time, the number of alarms is greatly reduced. In the process of alarm verification and aggregation, the false alarms caused by attacks unrelated to the system are eliminated through mutual confirmation of heterogeneous static information. From the experimental results, it can be seen that this kind of false alarms account for the vast majority of the total false alarms, and the simplification of the number of alarms mainly depends on the alarm verification and aggregation mechanism.

4.2 Hybrid Fishing AP Detection

Changing the built wireless AP, changing its IP address, MAC address and other information

respectively, according to the detection analysis, its wireless AP routing transmission path cannot be changed, so our client can detect according to this obvious characteristic information. The propagation time of each test is recorded in the experiment, which is also a very important detection factor for the test time factor. The total route tracking time of each test is obtained in the experiment. We carried out the above time for 15 times, and got the fastest test time of 15 groups of AP1 and AP2 respectively. Now we analyze the test time of two APs. Figure 4 shows the fastest time for wireless route tracking.

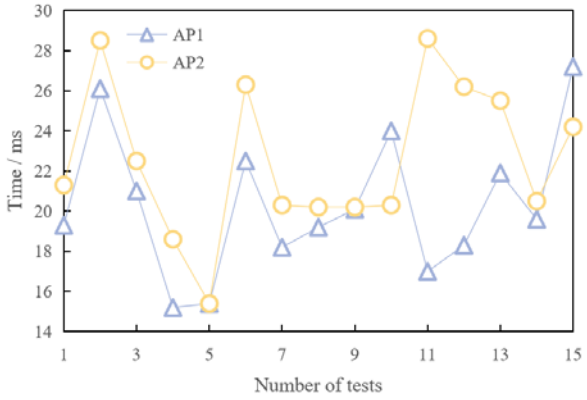

Fig.4 Propagation Time of Wireless Route between Ap1 and Ap2

From the figure, we can see that the fastest time between AP1 and AP2 is only a few milliseconds, which is because the platform built in the above experiment belongs to the wireless man-in-the-middle attack, and in the wireless man-in-the-middle attack, the routing access time is longer than the normal access time due to the addition of propagation nodes.

In order to determine which one is the fishing AP, it is necessary to measure the threshold. In the experiment, the fixed server detects the location and the normal wireless AP, changes the location of the fishing AP, measures the $RSSI_i$ of the normal wireless AP and $RSSI_j$ of the fishing AP, and makes a difference between them to get $\left|RSSI_i - RSSI_j\right|$, whose threshold $a = \left|RSSI_i - RSSI_j\right|$. Change the fishing AP position to get the threshold a and detection rate data, and measure it 50 times when a is 1dB to 6dB. As shown in fig. 5.

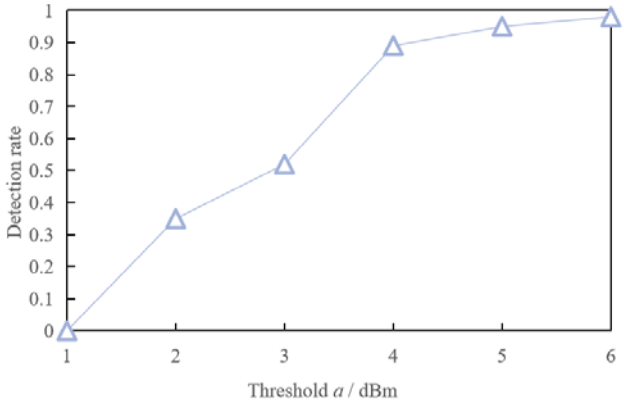

Fig.5 Detection Threshold

From 5, it can be concluded that the detection rate increases with the increase of threshold.

5. Conclusion

In the process of network development, network security is a particularly important limiting factor. With the continuous development of network security visualization technology, network security problems can be alleviated. This paper designs a heterogeneous security information processing framework based on alarm preprocessing, alarm fusion and event association. The extended DS evidential reasoning method is used to realize the fusion of heterogeneous dynamic

security information, and the trusted wireless AP is added to the white list, and the detection method is realized for the server detection program newly added to the wireless LAN. The mobile phishing detection association technology based on multi-source heterogeneous security data fusion is realized. Experimental results show that the hybrid phishing AP detection technology in WLAN is superior to the unilateral detection in terms of detection flexibility and accuracy.

Acknowledgement

Nantong Science and Technology (guidance) project “Research on Application of multi source heterogeneous data fusion analysis technology in mobile phishing attack detection” (No.: JCZ20141)

References

[1] Xiao Hongjiang, Zheng Guanwen, Jia Hongjun, et al. Exploration of Multi-source Heterogeneous Audience Big Data Platform Architecture. Radio and television technology, vol. 045, no. 007, pp. 33-37, 2018.

[2] Wu J, Wu Y, NNiu, et al. MHCPDP: multi-source heterogeneous cross-project defect prediction via multi-source transfer learning and autoencoder. Software Quality Journal, no. 3, pp. 1-26, 2021.

[3] Ovando D, Poon S, Costello C. Opportunities and precautions for integrating cooperation and individual transferable quotas with territorial use rights in fisheries. Bulletin of Marine Science - Miami-, vol. 93, no. 1, pp. 101-115, 2017.

[4] Xing, Long, Yin, et al. Heterogeneous cross-project defect prediction with multiple source projects based on transfer learning. Mathematical biosciences and engineering : MBE, vol. 17, no. 2, pp. 1020-1040, 2019.

[5] Huo G, Zhang Q, Zhang Y, et al. Multi-Source Heterogeneous Iris Recognition Using Stacked Convolutional Deep Belief Networks-Deep Belief Network Model. Pattern Recognition and Image Analysis, vol. 31, no. 1, pp. 81-90, 2021.

[6] Zhou Yangjun, Liang Shuo, Yu Xiaoyong, et al. Research and Implementation of Distribution Network Operating Analysis Platform Based on Multi-Source Heterogeneous Data. Southern Power Grid Technology, vol. 012, no. 008, pp. 59-64, 2018.

[7] Cheng J, Zhang Y, Feng Y, et al. Structural Optimization of a High-Speed Press Considering Multi-Source Uncertainties Based on a New Heterogeneous TOPSIS. Applied Sciences, vol. 8, no. 1, pp. 126, 2018.

[8] Wang W. Government ecological governance management based on heterogeneous multi-processors and dynamic image sampling - ScienceDirect. Microprocessors and Microsystems, vol. 82, no. 1, pp. 103909, 2021.

[9] Nagel L, Galeazzi R, Wisniewski R, et al. Fault Detection and Isolation in Linear Heterogeneous Multi-Agent Networks. IFAC-PapersOnLine, vol. 51, no. 24, pp. 784-789, 2018.

[10] Jung H, Koo K, Yang H. Measurement-Based Power Optimization Technique for OpenCV on Heterogeneous Multicore Processor. Symmetry, vol. 11, no. 12, pp. 1488, 2019.