Design and Implementation of Distributed Intrusion Detection System based on Honeypot

Yun Yang
School of Electrical and Information Engineering
Shaanxi University of Science & Technology
Xi'an, China
yangyunll@163.com

Jia Mi
School of Electrical and Information Engineering
Shaanxi University of Science & Technology
Xi'an, China
lockdog_jia@yahoo.com.cn

***Abstract*—For the shortcoming of traditional intrusion detection system (IDS) in complex and unknown attack detection. A distributed intrusion detection system based on honeypot was proposed. We make use of honeypot to collect the invasion characteristics on the network, and use the method of unsupervised clustering (UC) and genetic clustering to extract the data for analysis. In addition, in order to improve the detection performance of the IDS, it combined protocol analysis with signature detection modules. Experiments result show that this system can better detect intrusion and improve the overall safety performance of large-scale networks.**

Keywords-intrusion detectoin; honeypot; UC; genetic algorithms

I. INTRODUCTION

In recent years, with its rapid development, network has extended to every social corner, people have been led into the era of information technology. In the process of its growing application, network has gradually expanded from a small business to large-scale commercial areas, business management, education, research and government agencies. Everyone enjoys the convenience brought by the Internet, but at the same time has to face with the challenges of information security, especially endless network attacks. How to better defense these attacks, safeguard our network has become an important subject of information technology.

II. DESIGN AND IMPLEMENTATION OF THE SYSTEM

An IDS system mainly refers to the invasion behavior found in the network. According to the method of detection, IDS system is divided into two categories: protocol anomaly detection and signature detection (misuse detection) [1]. Anomaly detection based on protocol can verify the unknown attacks effectively, but can not detect attack violating an agreement. Misuse detection system matched attack action by stored attack signature in intrusion rule databases, the method spent less time and achieves a high detection rate. However, signature detection system is unable to discern new type of attacks or a large number of complicated attacks.

A. System Architecture

In this work, in order to overcome the deficiencies of traditional IDS system, we used the method of distributed signature collection, distributed processing, distributed response, and two detection modules (protocol analysis and signature detection). The system architecture is shown in figure1, it includes the sensors, protocol analysis module, signature detection module, and intrusion alarm module.

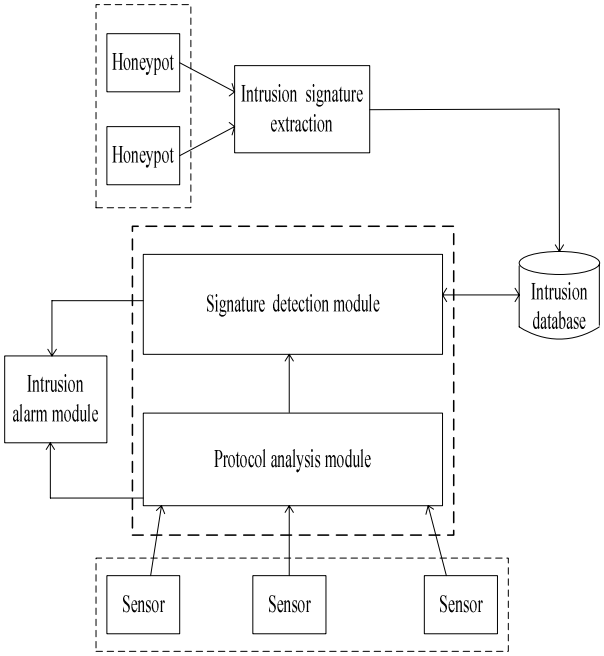

Figure 1. Structure of distributed instrusion detection system.

1) Sensor module: Sensors distributed in different networks of computer systems, its main task is collecting the raw data on the network according to the predefined rules, and organizing into the audit data that is submitted to the protocol analysis module for analysis.The sensor located at the bottom of the whole system, in the large-scale network environment, they are independent among the various sensors, and each detector only receives packets for the scope of their network.

2) Protocol analysis module: Protocol analysis technique has great advantages at present, any data which violates RFC can be considered a protocol anomaly through its technique checking. Such technique has a greatest feature where it can detect unknown buffer overflow vulnerabilities and denial of service attacks effectively. According to the corresponding protocol (TCP, UDP, ICMP) configure into a

protocol tree, and decompose the packet which the sensors submitted , then call the tree structure for the fast match. If there is an exception, then call the intrusion alarm module. Otherwise, classify the packets and anomaly detection module used to continue data analysis. The module can be easily extended, if we discover a new protocol type, the protocol units can be manually increased or modified.

3) Signature detection modules: Its work is the process of downloading and evaluating the data classified by protocol analysis module. We make use of misuse detection technology to ensure detection accuracy and real-time. In order to enhance the matching speed, an efficient algorithm for string matching (Boyer-Moore) is used [2]. Furthermore, taking into account the system open characters, invasion rules can be added dynamically by the honeypot to enhance the performance of pattern matching method.

4) Honeypot: Honeypot is distributed in the network designed deliberately to confuse the invaders and lure intruders to attack [3] [4]. Honeypot can collect intrusion information and grasp the current new attack methods timely. The system uses the distributed honeypot architecture to ensure data control and capture in large-scale networks.

5) Intrusion alarm module: Alert message can be sent to terminal display to us and classified statistics in terms of warning log. Besides, as system has a large quantity of statistical data, we compress the data through statistics compression technology.

B. Implementation of Key Technologies:

1) sensor technology: Sensor module located in the bottom of IDS system, it is a foundation of the whole system. In the system, according to a predefined strategy, sensor is responsible for monitoring the network and IP recombination. As we know that Ethernet data transmit by the broadcast. Generally, a program only receives Mac address packets which belong to them. To capture all packets, the network interface is set to promiscuous mode [5]. We using winpcap to capture data which is an open source library for packet capture and network analysis for the Win32 platforms, it can receive and send the packets independently from the host protocols. Data capture code is as follows:

```
// Create multi-thread to handle capture packets
LPDWORD IThread=NULL;
m_ThreadHandle=CreateThread(NULL,0,MyCaptureThrea
d,this,0, IThread);
// Capture packets in this thread
DWORD WINAPI MyCaptureThread(LPVOID
lpParameter)
{
CSnifferDlg* pthis=(CSnifferDlg*)lpParameter;
int res; // Use global variables
struct pcap_pkthdr *header;
const u_char *pkt_data; // Pointer to const
while((res=pcap_next_ex(pthis-> adhandle , & header , &
pkt_data ))>=0)
{
// Dealing with capture packets
pthis->SavePacket(header,pkt_data); //Save data
pthis->UpdateList();//Informed view list of updates
}
}
```

2) Protocol analysis technology: Protocol analysis technique parses the captured packets. In accordance with TCP/IP model, all TCP, UDP, ICMP packets transmiss by IP packet format in the network [6]. In Ethernet, packets with protocol signs can be matched with the protocol tree, its work principle is as follows: from the Ethernet frame access to the Ethernet header (14 bytes), which is consist of destination MAC address (6 bytes), source MAC Address (6 bytes) and the frame type (2 bytes). The frame type contained some protocol type, such as ARP, IP.Corresponding to the protocol number are 0x0806 and 0x0800. IP header include as follows: source IP address, destination IP addresses, flags, and IP protocol type. We can identify a protocol type by the identification number, such as TCP (6), UDP (17) or ICMP (1), and the specific application-layer protocol type can also be recognized by the port number, such as SSH corresponding port number is 22, Mysql port number is 3306. For all protocols, the system uses these methods to identify, if the packets conducted are abnormal, the system will call the alarm module. What's more, to make the protocol analysis module work better, the protocol structure needs to be defined. We defined the ARP protocol as follows:

```
typedef struct arp
{
unsigned short    arp_hdr; //format of hardware address
unsigned short    arp_proto; //format of protocol address
unsigned char     arp_lha; //length of hardware address
unsigned char     arp_lpa; //length of protocol address
unsigned short    arp_opt; // operation code(ARP or RARP)
unsigned char     arp_ha[6]; //sender's hardware address
unsigned long     arp_pa; //sender's protocol address
unsigned char     arp_tha[6]; //target's hardware address
unsigned long     arp_tpa; //target's protocol address
}ARP;
```

3) Honeypot System: In this work, Honeyd has been deployed in the network, it is the low-interaction honeypot system. In order to make the honeypot have more attractive to attackers, a lot of deceptive techniques need to be used. Information deceptive technology can be divided as follows:

a) Tips deceive: Some applications or services will leak some sensitive information, we can modify these tips to deceive the attacker.

b) Port deception: As we know that each program has its port for communication in the network, for example, Telnet server use TCP 23 port, PC Anywhere use TCP 5631-5632. Honeyd can be configured to run arbitrary port so that it appears to be running certain services.

c) Operating system(OS) deception: At present, there are some security scanning tools like: Nmap. These tools use TCP/ IP protocol stack for remote OS identification [7]. It uses the technology of different OS making the different

handle to TCP/IP protocol. However, these can be applied in turn, we can use the Honeyd to intercept the probe packets, and send packets to respond such requirements, so that we can deceive opponents successfully.

4) Instrution signature exstraction: Intrusion database is one of the important components in the misuse detection module, it also affects the overall system performance. Each rule should have the source IP address, target IP address, protocol type, source port, destination port, trigger the rule action, attack signature. In this work, the system has already contained a custom rule-base intrusion, and in order to ensure that the latest attack features can be detected in time, the system uses multiple honeypot placed in the different network to collect the invasion characteristics, and UC and genetic algorithms are used for mining invasion feature from the honeypot's audit records.

The clustering analysis method used an idea that the normal behavior is greater than the abnormal to divide datasets into various classes. The method based on unsupervised clustering dividing datasets into different groups. In the environment of the group, various objects have a higher similarity. On the contrary, among the groups there will be a lower one. The distance between two objects is determined by using the Euclidean Distance equation (1):

$$d(x_i,x_j)=\left(\sum_{k=1}^{n}\left|x_{ik}-x_{jk}\right|^2\right)^{\frac{1}{2}}=\sqrt{\left|x_{i1}-x_{j1}\right|^2+\left|x_{i1}-x_{j1}\right|^2+\cdots+\left|x_{i1}-x_{j1}\right|^2} \quad (1)$$

a) $d_{ij} \le d_{ik+}d_{kj}$, for any three objects i,j,k(triangle inequality)

b) $d_{ij}>0$, for any objects i,j, $X_i \neq X_j$;

c) $d_{ij}=0$, for any objects i,j, $X_i = X_j$;

d) $d_{ij}=d_{ji}$, for any objects i,j;

Initial clustering use the following method:

Step1: Make choice of an arbitrary object, and use it as a cluster center to construct a new class.

Step2: Continue to read a new object, and use Euclidean Distance formula to calculate the minimum distance. If it does not exceed the threshold, it will be placed in minimum distance class and form a new cluster center. Otherwise, a new class will be constructed with ones and repeat step 2.

Step3: Sort by the number of objects that the class contains, filter out noise and isolate points.

In the genetic algorithm, the choice of the fitness function is very important for us. In order to achieve a better result, we should try to reduce the distance within clusters while increasing the distance between clusters. Fitness function is as follows:

$$f=\frac{\sum_{i\neq b}\left|a_i-q_b\right|^2}{\sum_{a_i\subset d_b}\left|a_i-q_b\right|^2} \quad (2)$$

$$\text{Where } d_b=\frac{\sum_{a_i\subset d_b}a_i}{n} \quad (3)$$

In equation (2),(3), a_i indicated the center of cluster in the initial stages, q_b show a_i's subordinate cluster d_b. If the bigger f, the better clustering.

$$p_c=\begin{cases}k_1\dfrac{f_{\max}-f'}{f_{\max}-f_{avg}}, & f'\ge f_{avg}\\ k_2, & f'<f_{avg}\end{cases} \quad (4)$$

$$p_m=\begin{cases}k_3\dfrac{f_{\max}-f}{f_{\max}-f_{avg}}, & f\ge f_{avg}\\ k_4, & f<f_{avg}\end{cases} \quad (5)$$

In equation (4), (5), $k_i \in [0,1](i=1,2,3,4)$, $f_{\max}$ indicated the largest fitness of a group, f_{avg} indicated the average fitness, f' indicated crossing one of the largest fitness. According to the above operation, the algorithm terminated in the case that evolutional generation is less than the default value.

III. EXPERIMENTAL ANALYSIS

The system use VC++ 6.0 sp6 and the last Platform SDK to design. At the same time, KDDCUP 99 intrusion dataset is used for evaluation [8], we extract the 9500 records (500 records of abnormal behavior, 9000 records of normal behavior). In the experiments, tests use the different thresholds.

When the threshold is 40, the detection rate of Probing and U2R is higher and mission rate is relatively lower. When the threshold is lower, the normal recording is far away from the normal clusters, which led to a higher mission rate. When the threshold is 50, the result is lower. The simulation results are given in table1.

TABLE I. SIMULATION RESULTS OF DIFFERENT THRESHOLDS

Threshold	Intrusion Type							
	DOS		*Probing*		*U2R*		*R2L*	
	Detected	*Missed*	*Detected*	*Missed*	*Detected*	*Missed*	*Detected*	*Missed*
10	71.21%	4.31%	74.24%	2.52%	71.31%	3.22%	62.22%	7.11%

30	69.12%	6.62%	88.32%	3.12%	87.51%	2.76%	68.66%	5.43%
40	74.88%	3.12%	86.56%	1.43%	88.25%	2.61%	87.38%	4.61%
50	65.44%	4.79%	81.44%	2.61%	79.54%	3.76%	61.12%	5.45%

IV. CONCLUSION

The abnormal activity detected timely and effectively in the network is the research emphasis of IDS system at present. In this paper, we presented an IDS system based on honeypot. We use the honeypot to set network trap for the attacker's attention. In addition, using honeypot to collect more record of intruders, existing rules databases are updated in time. In order to effectively analyze datasets from ones, the system uses the UC and genetic algorithms to improve detection results. The paper used a distributed system architecture to expand the detection range, effectively solving the problem of current centralized NIDS in large-scale detection, it also improved the overall safety performance of large-scale networks.

ACKNOWLEDGMENT

This project was supported by the Graduate Innovation Fund of Shaanxi University of Science and Technology.

This work was supported by a grant from Scientific Research Foundation of Shaanxi University of Science and Technology (BJ10-01).

REFERENCES

[1] Ozgur Depren, Murat Topallar, Emin Anarim, "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks," Expert Systems with Applications, vol. 29, Issue 4, pp. 713-722, November 2005.

[2] Shmuel T. Klein, Miri Kopel Ben-Nissan, "Accelerating Boyer–Moore searches on binary texts," Theoretical Computer Science, vol. 410, Issue 37, pp. 3563-3571, September 2009.

[3] Anton Chuvakin, "Honeynets: High Value Security Data": Analysis of real attacks launched at a honeypot, Network Security, vol. 2003, Issue 8, pp. 11-15, August 2003.

[4] Babak Khosravifa, JamalBentaha, "An experience improving intrusion detection systems false alarm ratio by using Honeypot," 22nd International Conference on Advanced Information Networking and Applications, 2008.

[5] Mohan Krishnamurthy, Eric S. Seagren, "Network Analysis, Troubleshooting, and Packet Sniffing," How to Cheat at Securing Linux, pp. 203-247, 2008.

[6] Sándor Molnár, Balázs Sonkoly, Tuan Anh Trinh, "A comprehensive TCP fairness analysis in high speed networks," Computer Communications, vol. 32, Issues 13-14, pp. 1460-1484, August 2009.

[7] Angela Orebaugh, Becky Pinkard, "Nmap OS Fingerprinting," Nmap in the Enterprise, pp. 161-183, 2008.

[8] Chi-Ho Tsang, Sam Kwong, Hanli Wang, "Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection," Pattern Recognition, vol. 40, Issue 9, pp. 2373-2391, September 2007.

[9] Cheng Xiang, Png Chin Yong, Lim Swee Meng, "Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees," Pattern Recognition Letters, vol. 29, Issue 7, pp. 918-924, 2008.

[10] Chia-Mei Chen, Ya-Lin Chen, Hsiao-Chung Lin, "An efficient network intrusion detection," Computer Communications, vol. 33, Issue 4, pp. 477-484, March 2010.

[11] Benjamin Morin, Ludovic Mé, Hervé Debar, Mireille Ducassé, "A logic-based model to support alert correlation in intrusion detection," Information Fusion, vol. 10, Issue 4, pp. 285-299, October 2009.

[12] Xiaojun Tong, Zhu Wang, Haining Yu, "A research using hybrid RBF/Elman neural networks for intrusion detection system secure model," Computer Physics Communications, vol. 180, Issue 10, pp. 1795-1801, October 2009.