Proceedings of the 6th WSEAS Int. Conf. on NEURAL NETWORKS, Lisbon, Portugal, June 16-18, 2005 (pp158-163)

Worm Detection and Auto-Signature Extraction in Large Scale Network

XinYi,[1],Fangbingxing,Yunxiaochun
Research Center of Computer Network and Information Security Technology,
Harbin Institute of Technology
Harbin, China

Abstract: - Nowadays, worms have been one of the leading threats to information security and service availability. Current operational practices have not been able to manage the threat effectively. So it is very important to make early warning of the burst of worm in large scale network and extract the network signature automatically. Based on the TCP/IP Flows, the paper introduces a novel methodology to analyze the feature attributes of network traffic flow, including real-time data detection and traffic models. Integrated with data preprocessing, we construct an auto-signature extraction algorithm. We deployed them in our campus network (more than 20000 computers with 400M/s).It is shown that the worms are detected with more efficiency and the worm signature is extracted accurately.

Key-Words: - network traffic; statistic; detection; signature extraction

1 Introduction

With the ever fast development of computer networks, internet is exerting an incredible influence on our society and is changing the way in which people work and live. The network-based computer security is attracting increasing attention. From the first widespread worm in 1988, namely Morris worm [1], the number of security incidents reported grows greatly. Up to 2003, the count of incidents reported was 137,529[2]，as in Figture1.

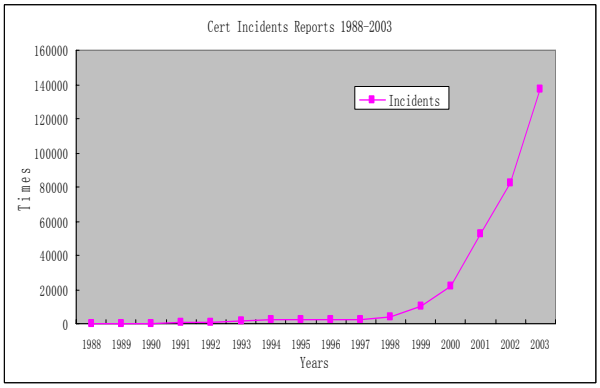

Figure 1: Internet incidents reported from 1988-2003

Our society suffers heavy losses in the incidents. In all of them, worms constitute a significant security and performance threat. Worms self propagate across networks exploiting flaws in operating systems and services, stealing sensitive data, erasing important files and congesting network links. In 2001, the Codered [3] and Nimda[4] infected hundred of thousands computers. In January 2003, slammer [5] infected more than 90 percent of vulnerable hosts within 10 minutes. In August 2003, Blaster [6] and its variation Welchia [7] were spreading widely, as well as Sasser [8] in 2004. Since worms spread so quickly that human response was ineffective, it is crucial to detect worm propagation, extract signature and protect network infrastructure effectively.

2 Related works

There exist several methods by which to analyze worm pattern and detect worm propagation.
GrIDS [9] is a prototype intrusion detection system that was designed to explore the issues involved in doing large scale aggregation of traffic patterns. It puts together reports of incidents and network traffic into graphs and is able to aggregate those graphs into simpler forms at higher levels of the hierarchy. To detect a worm, GrIDS constructs activity graphs which represent hosts and activity in a network, it counts the number of nodes and branches in the graph. When the counts exceed a user-specified threshold, GrIDS reports a worm. The approach only constructs graphs without detail analysis of destination ports and communication content; moreover there is no similarity comparison. It is very hard to detect cross-infection.
By aggregating network error messages resulting from failed attempts at worm scan packet delivery (ICMP-T3), Berk et al [10] present a scalable framework for detecting active Internet worms on

[1] Corresponding author. Xinyi. Computer Network and Information Security Technique Research Center, Harbin Institute of Technology, Harbin 150001, China E-mail: xinyi@hit.edu.cn;xinyi@hlmobile.com

global networks. However, this approach can not detect the worm when the route bans ICMP-T3, for instance ISPs protect infrastructure from Welchia's ICMP scans. Schechter et al [11] present a hybrid approach to detecting scanning worms that integrates two existing techniques: sequential hypothesis testing and connection rate limiting. But there is a limitation to the approach which will consume a lot of system resource, for keeping the state until TCP(SYN|ACK) packet or timeout including normal traffic.

C. Kreibich[12] present a systems, Honeycomb, which monitor network traffic to identify novel worms, and produce signatures for them using pattern based analysis, i.e., by extracting common byte patterns across different suspicious flows. These systems all generate signatures consisting of a single, contiguous substring of a worm's payload, of sufficient length to match only the worm, and not innocuous traffic. The shorter the byte string, the greater the probability it will appear in some flow's payload, regardless of whether the flow is a worm or innocuous.

3 Statistics of Network Traffic

The internet is based on TCP/IP protocols, the IP packets which constitute the net flows are the basic data of our research.

3.1 The types of packets

In order to make early warming of worm propagation and auto-extract signature, we count the IP packets from several aspects:

Classification by protocol. Based on the 8 bit Flag in the IP head, we can identify the protocol of each packet. So we can count the packets amount of every protocol.

Classifiction by port. There are source port and destination port in the TCP or UDP packets head which correspond to different transfer process in net flow. Classify the TCP and UDP packets based on the source port and destination port, then the amount of the TCP and UDP packets at each port can be obtained.

Classification by TCP flag. There is a 6 bit flag in TCP head, indicating the status of the connection. If all the TCP packets on the network are classified by the flags, the amount of the TCP packets with distinct flag will be obtained, especially syn and rst packets.

Packets of special types. Besides all the conditions above, some special packets have also been recorded, such as the amount of broadcasting packets targeted at the networks.

3.2 Two buffers technology

All of that information is the basic information of some network, so we can analyze them in depth in the coming chapters. Due to the high speed and huge amount of packets, we analyze the packets in every minute. If the analyzing process get the data form capture buffer while the statistic thread process the buffer, there will be a crash in the program. To settle this problem, we use 'two buffers' that are presented by P ,Q allocated in the memory. Firstly, the system insert the capture data in buffer P. The system will switch it in one minute inserting data in buffer Q, and the analyzing process deal with the buffer P in the same time. When finished the data-process in P, analyzing process initialize buffer P and wait the coming one minutes to insert the data in Q while deal the buffer Q cyclically. By this scheduling strategy, we avoid the process lock and improve the efficiency, and then fulfill the need of real time analysis, as in Figure 2.

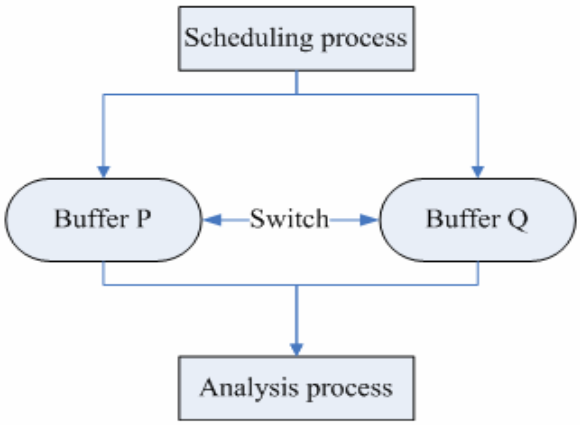

Figure 3: The structure of two buffers

4 Model of Network Traffic

Based on the recorded information, we can abstract the normal network traffic model.For the sake of setting up normal traffic model, it is necessary to accumulate the normal raw network traffic. We use the methods that have described in above chapter to record every packets in longtime as the basic information. The data sample scope is exceeding four hundred million.

4.1 The traffic model.

When setting up the traffic model, we assume the normal traffic change in some cycle ΔT. Based on the accumulation raw information, we abstract the

function relation between the amounts of each classification and time: $f_i(t)$ (i=1…n, represent the different statistic information). According to the properties of function (periodic, consecution, the top and bottom boundary scope, derivative property etc.) and the original data information, we make up the function relation of various types of statistics. We constitute the function description of the traffic curve variety and then set up the traffic curve in idea condition.

4.2 The volume of traffic

Definition1. Avg is the average of traffic in a time interval [*T*, *T*+T_0].
If *T* is the origin (*T*=0), then

$$Avg = \frac{\int_{T}^{T+T_0} f(t)\cdot dt}{T_0} = \frac{\int_{0}^{T_0} f(t)\cdot dt}{T_0} \qquad (1)$$

That can be represented in dispersible form:

$$Avg = \frac{\sum_{i=0}^{m-1} f(t_i)\cdot(t_{i+1}-t_i)}{T_0} \qquad (2)$$

In which $t_i \in [T, T+T_0]$ and $t_0 = T, t_m = T+T_0$

$$t_0 < t_1 < \ldots < t_m$$

The *Avg* represents the traffic volume of the network in a cycle. It is not sensitive to partial traffic variety. However, if there is large-scale anomaly traffics, such as burst of worm, the value of *Avg* change greatly. *f(t)* is traffic function.

4.3The diversity of curve shape

We can not identify the partial traffic anomaly from the value of the *Avg*. In order to determine the anomaly of partial traffic in a cycle, we must estimate the traffic function in the cycle is accord with the normal one. Because the volume of absolute traffic can change with some normal event, which differ from abnormal phenomena for example worm, it is necessary to eliminate the influence of the absolute traffic.

Definition2. AVG_0 is the average traffic in ideal condition.
We revise the traffic function for eliminating influence of the absolute traffic.

To set $$f^*(t) = \frac{f(t)\times AVG_0}{Avg} \qquad (3)$$

To set $F^*(t)$ to be the normal traffic function in ideal condition, the curve of $F^*(t)$ is the ideal condition one. Considering of the traffic function $f^*(t)$ in every interval [*T*, *T*+T_0], we define the function $Q^*(x)$ to present the shape different from $f^*(t)$ to $F^*(t)$.

$$Q^*(x) = \sqrt{\frac{\int_{T^*}^{T^*+T_0} (f^*(t) - F^*(t+x))^2 \cdot dt}{T_0}} \qquad (4)$$

In which x is the phase offset， $x \in [0, T_0)$.
When *x* changes, it is equivalent to adjust the phase difference between the $f^*(t)$ and $F^*(t)$. The smaller $Q^*(x)$ is, the more accordant between $f^*(t)$ and $F^*(t)$ is.

To set $$G^*(x) = \min(Q^*(x) \mid x \in [0, T_0)) \qquad (5)$$

Then $G^*(x)$ presents the curve difference between traffic and normal traffic.

Some instance as in Figure3.

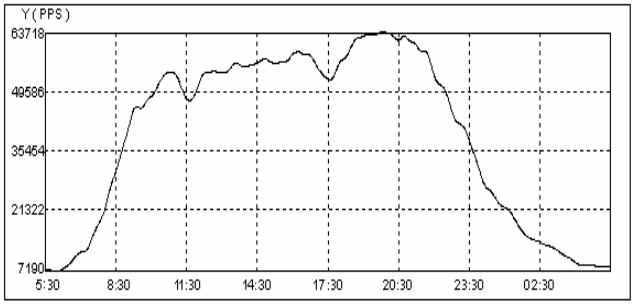

Figure 3:The curve of all flow in ideal conditions

5 The Real Time Worm Detection and Auto-Signature Extraction

After setting up the normal traffic model and calculating the normal traffic, the system compare them with the real time capture date therefore determine the anomaly data flow. We can detect the anomaly by examining the difference between the real traffic and ideal traffic which concerning traffic volume and shape of traffic function. We can set the critical value G^*_0 and Avg_0, when $Avg > Avg_0$ the anomaly takes place. And then $G^*(x) > G^*_0$, the shape of the traffic is abnormal.
And then, there are distinguished features of worm propagation, burst of probe scanning packets, our approach can easily detection them. After detection we can extract the worm signature. In large scale network, the bandwidth is very broad and the client amount is very large. Because the worm propagating

speed is beyond mankinds intervene, the signature extraction must be automatic in real time. We propose a real time feedback approach. When detect anomaly traffic, it is not necessary to be concerned with the total feature of it. The system only extraction parts of the feature by which to pre-filter all the data flowing in the network then feedback of anomaly features.The structure of automatic signature extraction show in fig.4

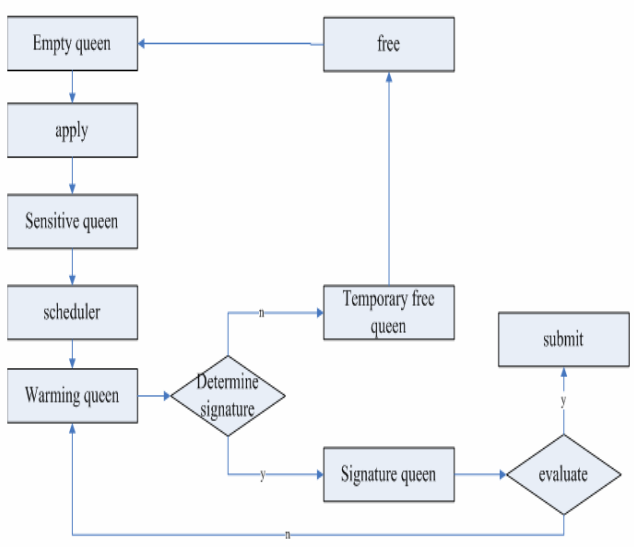

Figure 4: The structure of automatic signature abstraction.

The anomaly signature extraction algorithm is mainly based the features of different protocols to make depth analysis, which is varied from different protocols. For example, the system is just concerned with length, source port, destination port, the flag and part content of the TCP/UDP packets. For ICMP, just the length, type and part content.

5.1 The length priority queen

The length priority queen reflects the anomaly signature set of certain length packets. When the certain length packets increase greatly unexpectedly, the system will add the length as a signature into sensitive queen. Once the signature reaches the setting threshold, the system will add this length signature into warming queen. Then the system mainly analyzes the anomaly packets, show as below.

The protocol type of the packets. In a analysis cycle, the system records above certain length packets. If p_i (i=1…255) represents the percent of protocol i packets. Set protocol I to be $p_I=max(p_i)$.

The source and destination. From the above description**,** the system continues analyzing the packets of certain length and protocol I. If $s_i(i=1…65535)$ and d_j $(j=1…65536)$ represent the percent of source port and destination ports in the whole packets respectively. If there exit *I, J,* to set $s_I=max(s_i),d_J=max(d_j).P$ is the threshold, if $s_I>P$ then the anomaly source port is *I* and if $d_j>P$, the anomaly destination port is *J,* otherwise there is not certain port in the anomaly event.

The signature of packets

After the above To set the length of the anomaly is *l*, $data_0$, $data_1…data_{i-1}$ represent the value of highest frequency in every bit. $p_0, p_1....p_{l-1}$ represent the frequency of the value in certain bits.

Calculate the average and the variance respectively.

$$E(p)=\sum_{i=0}^{l-1} p_i$$
$$D(p)=\frac{1}{l}\sum_{i=0}^{l-1}(p_i-\overline{p})^2 \qquad (6)$$

Based on the $E(p)$ and $D(p)$, we can set propert threshold to determine whether there is anomaly in the data of packets.

5.2 The port priority queen

The port priority queen represents the anomaly signature set of certain port. When the packets of certain port increase greatly, the system will add the port as signature into the sensitive queen. After that, when they reach the set threshold, the analysis process will add the port into warning queen. The port priority queen contains two part, source port priority queen and destination port priority queen.

To extraction the signature of the port priority queen, the system analyze the length of packets, protocol of the packets and the anomaly data in the packets same as the length priority.

5.3 The flag priority queen

The flag priotity queen represents the anomaly of certain flag in TCP packets.When worms propagate, they will scan the network for victims. They are usually SYN packets which always congest the network, but the Worm.Welchia and Worm.Slammer are exceptions whose probe packet are UDP and ICMP respectively.

Because the scanning strategy, the worms scan the hosts which do not exit or the ports do not open. Consequently, the TCP scanning worms propagatting in large scale network always produce more ACK&RST packets. And this is the important feature of scanning worm, and the system can determine the scanning IP of the hosts and the attacked ports.

6 Evaluations in Real Network

When the Worm.Welchia propagate, it can produce a lots of ICMP request packet to find the live targets in the network. So there are a large number of ICMP packets in the network, below is the curve of ICMP traffic as Figure 5.

When the Worm.Welchia propagated the volume of the ICMP traffic was grew more than 50 times that affected the network greatly and the deference of the curve exceeded the normal scope. So, there were lots of computer in the network had been infected by the worm.

After the pre-filter, in order to extract more features, it combine the new extracting feature with the pre-filter one as a new feature which will be submitted to the filter system for more accurate extraction. Repeat that process until find out the final signature. For example, the worm wechia bursted in august 2003. Firstly ,system detect the amount of ICMP packets grow greatly, then the pre-filter system analyzed all the ICMP packets. After that it found that packet length is 92, the pre-fliter system feedback all the ICMP packet with 92 byte and found the type of those ICMP is Request, then feedback again. Until found the final signature:

Protocol:ICMP;Length:92;Flag:Request;Content:A
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAA

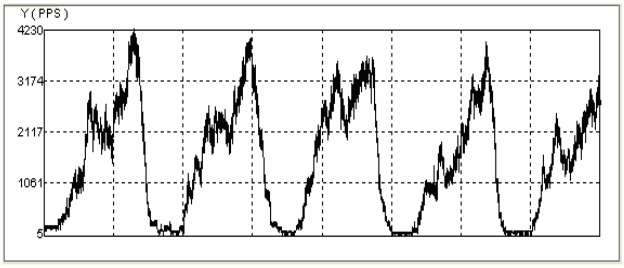

Figure 5.The curve of ICMP traffic when Worm.Welchia propagate

7 Conclusions

In this paper, we have described an approach to make early warning of the burst of worm in large scale network and extract the network signature automatically. Since Internet is based on TCP/IP protocols, we statistic all the IP packers in real large scale network. By means of real-time data flow detecting and setting up of network traffics model, the early warning and signature abstraction were realized

Acknowledgements

This paper is supported by the Chinese National Fund of Science. (Granted No. 60403033)

References:

[1] Cert,CERT/CC Statistics 1988-2003， Available at http://www.cert.org/stats/cert_stats.html

[2]Mark Eichin and Jon Rochlis. With Microscope and Tweezers: An Analysis of the Internet Virus of November 1988. In IEEE Computer Society Symposium on Security and Privacy, 1989.

[3]D. Moore, C. Shannon, and J. Brown. Code-Red: a case study on the spread and victims of an Internet Worm. In Proc. ACM/USENIX Internet Measurement Workshop, France, November, 2002.

[4]CAIDA. Dynamic Graphs of the Nimda worm.http://www.caida.org/dynamic/analysis/security/nimda/

[5]D. Moore, V. Paxson, S. Savage, C. Shannon, S.Staniford, and N. Weaver. Inside the Slammer Worm. IEEE Security and Privacy, 1(4):33-39, July 2003.

[6]Eeye digital Security. Published Advisories, ANALYSIS: Blaster Worm. Available at: http://www.eeye.com/html/Research/Advisories/AL20030811.html

[7]Symantec Crop. Security Response W32.Welchia. Worm. Available at:http://securityresponse.symantec.com/avcenter/venc/data/w32.welchia.worm.html

[8]Eeye digital Security. Published Advisories, ANALYSIS: Sasser Worm. Available at: http://www.eeye.com/html/Research/Advisories/AD20040501.html

[9]S. Cheung, R. Crawford, M. Dilger, J. Frank, J. Hoagland, K. Levitt, J. Rowe, S. Staniford-Chen, R. Yip, and D. Zerkle. The design of GrIDS: A graph-based intrusion detection system. Technical Report CSE-99-2, Department of Computer Science at UC Davis, 1999.

[10]V. Berk, G. Bakos, and R. Morris. Designing a framework for active Worm detection on global networks. In Proceedings of the IEEE International Workshop on Information Assurance, pages 13-23, Darmstadt, Germany, March 2003. IEEE.

[11]Schechter S E, Jung J, Berger A W. Fast Detection of Scanning Worm Infections. Seventh International Symposium on Recent Advances in Intrusion Detection, Sophia Antipolis, France, September 2004

[12]C. Kreibich and J. Crowcroft. Honeycomb - creating intrusion detection signatures using honeypots. In Proceedings of the Second Workshop

on Hot Topics in Networks (HotNets-II), November 2003.