# Network attack characteristics of automatic data extraction technology

Pu Tian yin[1] Rao zheng Chan[1] QIN Zheng[2]

1 Department of Computer Science, Tong ren College, Tong ren 554300

2. Hunan University of Information Science and Engineering, Chang sha 410006

**Key words**: network security; attack characteristic; Technology

**Abstract**: attack automatic feature extraction technology is an important research of network security technology. From the network present situation research proceed with, to attack the automatic feature extraction technology for definition and classifycation, and for each category of technology are introduced in detail, and presents several attacks of automatic feature extraction technology, finally to present these technical deficiencies and the possible development trend are discussed.

## Introduction

With the rapid development of Internet and wide application, network attack increasing, destruction and threat to network infrastructure. Typical network attack means such as computer virus, spyware, worms and denial of service attacks into the network system, stealing and destroying system sensitive information, and blocking the key network service, seriously affected the normal operation of the network and the use of user. National Computer Network Emergency Response Coordination Center ( CNCERTCC ) research report shows, 2007, computer virus, denial of service attack, worm or Trojan, as well as the webpage of malicious code, network attack occupied our country network security events total 30%[1]. Therefore, timely and accurately detect and prevent network attacks against the network safe and reliable operation is very important.

A large number of network attack analysis showed that the [2], network attack has certain characteristics, according to this feature can detect and prevent network attack. Based on this premise, the academia and industry has been proposed based on the characteristics of the network security system, such as a network intrusion detection system ( NIDS ), network intrusion prevention system ( NIPS ) and firewall etc, and widely implemented in practical environment, in detecting and preventing network attack a respect to obtain certain result. Usually, these systems are deployed in key network position, real-time capture network traffic, carries on the analysis to. From a group of related to flow rate measurement, can choose to attack the metric form characterization of subsets, each subset represents an attack characteristic. Based on the feature matching method, from the network flow feature extraction and predefined attack feature matching, thus identifying and filtering suspicious and aggressive behavior.

In recent years, network attack showing a intelligent and polymorphic characteristics such as [3, 4]. First of all, network attack tools more automated and intelligent, easy to produce novel and variant attack; second, the network hacker continuously produced with polymorphism of attack, to avoid the security defense technology based on features. Based on the characteristics of the network security system should possess from the suspicious network traffic continued to extract and update the attack characteristic ability to defense, emerge in an endless stream of new network attack. Therefore, NIDS, NIPS and firewall based on the characteristics of the network security system is facing a critical question: how the real-time, efficient, accurate extraction of attack characteristics. Therefore, to study how fast, accurately detect the unknown attacks, and automatically extracted features, has become the research hotspot.

Foundation item： Hall of Guizhou province science and technology project LKT201210

Author brief introduction ： Pu tianyin (1974.10), male, Guizhou Sinan, master, associate professor of Tongren University, research direction: project management, information security ( E-mail:trszjsj@163.com )

Published by Atlantis Press, Paris, France.

## Attack of automatic feature extraction and classification

Automatic generation of attack signatures is refers to does not need artificial help, automatically detect new attacks and to extract new attack characteristics of the process, its purpose is to as soon as possible to provide high quality IDS to detect rules as the main form of attack characteristics.

Automatic generation of attack signatures for attack detection and feature extraction of 2 basic steps. Therefore, and intrusion detection system can be divided into network IDS ( NIDS ) and IDS ( HIDS ) similar to the host, according to that the attack position, automatic generation of attack signatures can also be classified based on the network and host based 2 categories, respectively, abbreviated as NSG ( network-based signatures Generation ) and HSG ( host-based signatures generation ).

NSG is mainly through the analysis of the network on suspicious data to extract the character type features. The character of the character is defined by the string ( binary string ) composition, distribution or frequency to describe attack. NSG system through data flow classifier or Honeypot system to discover network data in suspected attack behavior, and may include a sample suspected attack network data; and then it is divided into two parts, part of the NSG system of [4,5] on the suspicious data clustering, made from the same attack data are clustered together, then for each category of extracting attack characteristic, another part of NSG system without the clustering process, but the direct analysis of a mixture of multiple attacks of sample data, extraction can detect multiple attack characteristics. The final NSG system will extract the attack characteristics into detection rules, applied to IDS system detection.

HSG mainly refers to the detection host anomaly and the use of the host on the collection of information to extract the attack characteristic. According to the obtained information of host number, HSG can be further divided into white box HSG method, HSG method and grey box black box HSG 3 kinds of methods. White box is required for the HSG method source code, by monitoring the execution of a program that the attack behaviors occur, then the control source extraction of attack characteristics; gray box HSG method does not need to program source code, but must closely monitor the implementation of the program, when it is discovered that the attack by the process context field analysis to extract the attack characteristic; black box HSG method proposed recently, which does not require source code does not need to monitor the execution of the program, but by the " test data " on the procedures to attack, if the attack is successful, indicated that " effective test data ", and the " data " extraction of attack characteristics.

## Attack automatic feature extraction technology

**machine learning based on the semantic feature extraction method based on Probability:** an effective state automaton attack automatic feature extraction technology, is based on a probabilistic finite state automata algorithm PFSA ( Probabilistic Finite State Automaton ) on the cluster formation of clusters of induction treatment ( Cluster Generalization ), can be converted into finite state automata identified by regular language, and the extracted feature. Finite state automaton system [7] includes data extraction component ( Data Abstraction Component, DAC ) and feature extraction component ( Signature Generation Component, SGC ) two parts. Data extraction component through a series of modules of data preprocessing and normalization, which is suitable for clustering and feature extraction, the final formation of semi-structured conversation tree ( SSTs ). The DAC component SSTs code after the formation of the XML output to the feature extraction component. SGC components on the standardization data using a star cluster on the connection and session clustering, in which the similarity measure using cosine similarity, and then use PFSA algorithm of cluster formation of clusters formed can be summed up, finite state automata identified regular language, namely characteristic.

**based on the COPP ( Content-based Payload Partitioning ) algorithm feature extraction method**

The method is based on the COPP algorithm for data flow dynamic cutting, then adopts the statistical method to extract the characters of attack. Autograph [31] feature extraction is mainly

Published by Atlantis Press, Paris, France.

divided into two processes. First, use COPP algorithm to find the suspicious by heuristic method for data flow data segmentation, method is first to define pause mark, and then to the sliding window concept to compare, every time they meet a pause mark on segmentation. Secondly, the use of statistical methods to identify the most frequently occurring packets as the attack characteristics, calculation of the partition. Each data packet number, if a packet number exceeds a certain threshold value, then generating features.

**based on the Rabin algorithm Fingerprints attack feature extraction method**

The method of data stream is partitioned according to a recursive pattern string with the text string hash function value to attack feature extraction. Rabin Fingerprints algorithm to define a Hashi function, and the function of the pattern string with the text string for pretreatment, and then to be compared with the pattern string with the text string Hashi functionals are numerically equal principle, followed by comparison. The algorithm can use the previous value of Hashi function is recursive the next value of Hashi function, in order to reduce string matching algorithm time complexity and space complexity. The first application of Rabin [8] fingerprints algorithm in automatic generation of attack signatures. The method by comparing the two main attack characteristics -- from the source of infection are common sequence and destination address, identify new attack, then use content filtering algorithm for automatic generation of attack signatures.

**based on the two series common subsequence algorithm attack feature extraction method**

The method is accomplished by finding the longest common substring to attack feature extraction. LCS algorithm to solve the problem is to find the two string is the biggest public substring, and the largest public substring location. The [8] is proposed based on two series LCS algorithm to extract the attack characteristic method, in the reconstituted from intrusion detection system data, using LCS algorithm two two alignment to find the longest common substring, in order to confirm the attack characteristics. Based on the data of Honeypot Honeycomb system, firstly reorganize all the data from the Honeypot, and then in these data, uses the LCS algorithm two two matching, find the longest common substring, so as to find out the attack characteristic.

## Problems and research direction

**problems**

Existing attack automatic feature extraction technique in detecting unknown attacks, and automatically extracted features, made some achievements, but also has some insufficiencies:

( 1) feature extraction algorithm of high time complexity and space overhead. As of [12] using Rabin algorithm need to string pretreatment, a text string length m, string length N, then the algorithm pretreatment time on non ideal condition of time complexity is O ( MN ), also a larger space overhead. And as Autograph[9] COPP algorithm based on the use of statistical methods for feature extraction, program covers a large amount of system resources.

( 2) the extracted attack characteristics of poor quality, high rate of false alarm. The algorithm has its own shortcomings and deficiencies : one is the algorithm is not stable, for example, as a result of Hashi's function of applicable scope is limited, leading sometimes generates error matching; two is the algorithm is too simple, such as the Honeycomb[10] algorithm is too simple, so that for protocols such as NetBIOS of short string types extracted feature often attack attack including a large number of unrelated sequences of characters, led to the extracted attack characteristics of poor quality, high rate of false alarm.

( 3 ) for large deformation characteristics of attack to produce inaccurate, false negative rate is high, only the fairly limited attack type automatic feature extraction. Earlybird[9] and COPP based on the Autograph[11] are not considered protocol semantics, so only in a limited attack types by feature extraction. Autograph features are single continuous substrings, the attacker through the insertion or deletion of a few bytes can escape detection. Based on the probability of valid state automaton attack automatic feature extraction technology, it is difficult to extract substantial deformation of attack characteristics.

**research directions**

For the NSG method, the study shows that in recent years: as long as the attack sample high quality and be able to correct the clustering, we can extract high quality characteristics. Therefore, the current NSG method to study the key lies in: one is how to get high quality attack samples? Through the improvement of Honeypot system as well as an increase in noise filtering mechanism is one of feasible direction. The two is how to correctly clustering attack samples? Existing clustering methods have lower accuracy and anti malware attacks the problem such as ability difference, a new clustering method can be used for reference in related fields, such as bioinformatics research. Three is able to effectively resist the malicious attacks and NSG system design must consider the important issues. HSG can rapidly and accurately find attack and positioning of attack area, extraction of compound features (i.e., in addition to byte features also joined with the process context related information) will have a stronger description ability, the rate of false positives is generally lower, however, they are difficult to be used by the existing IDS. Therefore, the extraction accuracy can be as simple as possible the existing IDS application byte type feature is the HSG development direction.

The current HSG system output byte type characteristics compared to the NSG system also is not accurate enough, how to make full use of the host acquisition information makes the extracted Boolean byte character more precise and more complete is the direction in the future. In addition, reduces to the platform and environment dependent and reduce the influence on the system performance is also HSG system needs to emphasize improvements. The HSG system deployment in Honeypot there is generally no performance problems, so HSG technology and one high-alternative combining Honeypo system is a valuable research direction.

## Reference

[1]CNCERT/CC. 2007 [DB/OL]. http://www.cert.org.cn/articles /docs/ network security work report of common/ 2007 082123431.shtml. 2008-02-10

[2]Anderson J P. Computer Security Threat Monitoring and Surveillance[R]. James P Anderson Co, Fort Washington, Pennsylvania, 1980.

[3]CNNIC. twenty-first China Internet development statistical report [DB/OL]. http://tech.qq.com/zt /2008/cnnic21. 2008-02-18.

[4]S Staniford, V Paxson, N Weaver. How to own the Internet in Your Spare Time. In: Proc. of the 11th USENIX Security Symposium. CA: USENIX Association Berkeley Publisher, 2002, 149-167.

[5]YEGNESWARAN V,GIFFIN J T,BARFORD P,et al.An architec-ture for generating semantics-aware signatures[A].Proceedings of the14th USENIX Security Symposium[C].Baltimore,MD,USA,2005.97-112.

[6]NEWSOME J,KARP B,SONG D.Polygraph:automatically generat-ing signatures for polymorphic worms[A].Proceedings of IEEE Sym-posium on Security and Privacy[C].Washington,DC,USA,IEEEComputer Society,2005.226-241.

[7]Kreibich C，Crowcroft J．Honeycomb creating intrusion detection signatures using honeypot．In: Proc. of the 2nd Workshop on Hot Topics in Networks，Boston: HotNets-II, 2003, 51-56.

[8]Hyang-Ah Kim, Brad Karp. Autograph: Toward automated, distributed worm signature detection. In: Proc. of the 13th USENIX Security Symposium. San Diego: USENIX Press, 2004, 271-286.

[9]Piotr Kijewski. Automated Extraction of Threat Signatures from Network Flows. In: Proc. of the 18th Annual FIRST Conference. Baltimore: CERT, 2006, 261–271.

[10]Urjita Thakar. Honey analyzer - analysis and extraction of intrusion detection patterns & signatures using honeypot．In: Proc. of the 2nd International Conference on Innovations in Information Technology. Shri GS: Institute of Technology and Science, 2005, 88-104.

[11]Singh S, Estan C, Varghese G, Savage S. Automated worm fingerprinting. In: Proc. of the 6th Symposium on Operating Systems Design and Implementation. San Francisco: USENIX Association , 2004, 45-60.

Published by Atlantis Press, Paris, France.