Research on Virus Detection Technology Based on Heuristic Model

Yuanyuan Wang[1,a] ,Shouzheng Li[2,b] ,Ye Yuan[3,c] , Chao Liu[4,d] , Weimiao Feng[5,e] ,and Min Yu[6,f]

[1]Northeast Forestry University, Harbin , China, 150001

[2]Harbin Engineering University, Harbin , China, 150001

[3]Harbin Engineering University, Harbin , China, 150001

[4]Institute of Information Engineering,Chinese Academy of Sciences,Beijing,China,100093

[5]Institute of Information Engineering,Chinese Academy of Sciences,Beijing,China,100093

[6]Institute of Information Engineering,Chinese Academy of Sciences,Beijing,China,100093

[a] stdong0451@163.com, [b] 26407166@qq.com, [c]569093011@qq.com, [d]2443532866@qq.com, [e]330227346@qq.com, [f]272523869@qq.com

Keywords: heuristic feature, virtual execution, PE file virus.

Abstract. First, the article introduces the definition and characteristics of computer virus, and the classification of computer virus. Then, it introduces the mainstream of virus detection technology, such as signature detection technology, checksum method, behavior detection technology and artificial immune technology, etc. Finally, the experiment verification proves the feasibility of heuristic virus detection model .Experimental results proves that the method has high detection efficiency, and in terms of detecting unknown viruses, this method has a relatively low rate of false positives and non-response rates. Under the premise of heuristic feature library not being updated, static heuristic detection model can detect unknown viruses, winning the precious time for subsequent extracting signature.

Introduction

With the progress of science and technology in com-puter and Internet on all areas of actuation, triggered by a computer security problem, more and more people is arouse strong concern. Computer virus gradually in-creases and becomes the main threat to the computer security. Windows operating system has the most users, in order to benefit,Windows operating system becomes the primary target of the virus. Most common file format in Windows is PE, and there are many different kinds of PE infectious virus and it affects the worst, So it is urgent to curb PE infectious virus.

The serious threat to computer posed by virus greatly accelerates the development of anti-virus technology. Signature detection technology is one of the traditional virus detection technology, adopted by most virus detec-tion software, and its characteristic is easily to be realized and can identify the virus types. Signature detection technology needs to use the virus signature database. With the virus species increasing, it is need to constantly update signature database, declining the detection effi-ciency. Virus signature extraction mainly relies on man-ually extraction of experienced experts, but this method is time-consuming, and its extraction rate significantly lagged behind the virus update speed, causing the useless in the signature detection technology of the new un-known virus. In order to deal with the large number of unknown viruses, it is imperative to minimize the dam-age of unknown viruses, and study the new virus detec-tion means. This article gradually deep into virus re-search in PE file type started by the principle of comput-er.

Current research

Mainstream divided virus detection technology Ref.[1]: signature detection, calibration and testing, behavioral surveillance law. Advantages signature detection methods are known for the high efficiency of virus detection, to identify the type of virus, make the appropriate treatment based on the test results; disadvantage is the high cost of extracting signatures need to constantly update the signature database Ref.[2]. Calibration and testing method uses the history of the original file checksum value and the current value is detected file checksum comparison to determine whether the contents of the file is modified, and then infer whether the virus tampered with. This method can find known viruses, but also to discover unknown virus, but can not get the virus name, can not deal with virus-infected files, the early outbreak, no signature extracted from the emergency measure Ref.[3]. Behavior Surveillance Act, also known as host-based intrusion prevention technology, is a proactive defense technology. This technique is to conduct an executable program as a feature to detect viruses, these acts include: file read and write, port operations, network access, modify system services, registry and other sensitive operations. The advantage of this method is to known and unknown viruses have a more accurate test results. Its disadvantage is the higher rate of false positives, can not distinguish between viral species, likely to cause instability in the system Ref.[4].

Architecture heuristic detection model

Heuristic detection model can quickly determine whether a sample to be tested using a modified encryption (self-modifying) technology, if it is an ordinary PE files directly to the static heuristic detection model; if it is deformed to encrypt dynamic heuristic virus detection model dynamic heuristic model based on hardware-assisted virtualization technology Xen has high transparency, can effectively prevent virus detection for virtual environments. Produce new variants of the virus expressly PE virus in a virtualized environment, will be submitted to the nascent PE static heuristic virus detection model. Static heuristic detection model mainly uses PE file heuristic information as a feature, feature streamlined through heuristic signatures deposited using KNN classifier trained to classify, to achieve the detection of unknown viruses.

Design and implementation of a static heuristic detection model

Diomidis Spinellis proposed and proved "accurate detection of viruses of limited size is an NP-complete problem." Ref.[15]. Literature Ref.[16] proposed a program to accurately determine whether the code is malicious algorithm does not exist, it needs human intervention. Data mining algorithms can be applied to previous knowledge and experience in the approximate virus detection, to improve the detection of unknown viruses efficiency, reduce false positives and false negatives are very helpful.

Static heuristic detection model is mainly workflow: First, the sample set containing normal files and virus files for training, extract relevant features heuristic heuristic information based on PE, using information entropy heuristic features streamlined heuristic streamlined save feature to the feature library for virus detection based on SVM using distance weighted KNN classification algorithm, due to heuristic virus is characterized by common features, it can detect unknown viruses to some extent.

This module has a PE legitimacy detection module, heuristic feature detection modules. Main functions: to traverse folders to be scanned, the simple signature detection file, if the detection is successful explanation is known viruses. Failure to detect improved KNN classifier to detect unknown viruses; then if the test results are unsure, indicating that the virus uses a shell or encryption technology, the class file to a separate virtual dynamic heuristic detection model run, the virus in order to run will generate a new file, locate the newly created file sent plaintext heuristic virus detection model again classified.

PE legitimacy detection module will be detected PE file to the memory map and read into RAM, and PE file format compared to normal and then determine whether a file is detected as a legitimate PE file. E_magic structure by comparing the DOS file header and 'MZ' is equal to determine whether the legitimate head of DOS; then compare the PE header structures Signature and 'PE' is equal to determine whether the legitimate head of PE, legality testing process as shown above.

Design and implementation of dynamic heuristic detection model

Hardware-assisted virtualization Xen

Xen developed by the University of Cambridge a key open source projects. Xen supports paravirtualization (PV) and full virtualization (HVM). The former requires the upper virtual machine operating system be modified to adapt to Xen virtual environment provided; the latter does not need to modify the virtual machine operating system but requires CPU supports virtualization.

Xen virtual environments is mainly composed of Domain 0 and Domain U. Domain 0 is a modified Linux kernel, Xen VMM runs on a unique virtual machine, as the expansion of secondary HVM VMM to create, close other virtual machines. Domain 0 is always paravirtualization, it can directly access some of the physical hardware, and run simultaneously with other interacting (Domain U HVM Guests and Domain U PV Guests) in a virtual machine on the platform. Domain U HVM Guests unprivileged entire virtual machine has a special daemon Qemu-dm, the process is carried out to support HVM Guests Internet access and disk access requests. Virtual firmware HVM Guests must be provided by the Xen simulation BIOS initializes initialized HVM Guests of the operating system to work properly.

Hardware-assisted virtualization Xen

Dynamic heuristic detection model is based on hardware-assisted virtualization on Xen. The model effectively use self-modifying techniques to detect variants of the virus. The main features of the model shown in Fig. Because the model is running in Domain 0 and VMM, a priority over the Domain U Guest OS, and therefore priority over virus program, the virus program can not detect the presence of the monitoring program, the virus self-protection measures fail. From this model, a new generation of virtual machines external positioning location of the program, call the static heuristic model for testing.

Domain 0 in the Control module to control the Monitor by libxc VMM library, and receive event notifications from the VMM. Control module after notification by analyzing suspicious process to obtain information including process context and the new generation of location-virus program, to prepare for the next virus scan.

Working in VMM Monitor module can detect the specified process HVM special behavior, such as CR3 update these actions will trigger VMExit. According to these behaviors determine whether there is a new virus program generation, if there is to save the PC counter current process, stack and registers and other information to the Share Page in Domain 0. After the information is copied, Monitor module calls sent to Domain 0 Event Channel Control Module Message in a suspended Domain 0 HVM waiting for treatment.

Monitoring the specified process:

- Monitor section to determine the name of the executable program mirroring process, because the process of switching Guest OS need to change the contents of the register CR3, and if you set the first 16 VMCS in VMX-EXECUTE FIELDS structure is a content CR3 changes will trigger VMExit into Xen.
- Xen in vmx_vmexit_handler functions can catch and handle the exception, while access to the Guest OS FS register information.
- Ring 0 is stored in the register under FS KPCR (processor status block) information, the data structure has an important role in the operating system kernel, the structure can be obtained by the name of the currently running processes.
- Process Name and procedures will be monitored Mirror Monitor name matches a successful match is called virus positioning module, register CR3 preservation process page directory base address, the contents are saved to improve the efficiency of the next match.

File micro filter driver to achieve positioning nascent virus

Variants of the virus to use self-modifying code technology, the machine code to the operating system is encrypted. Such viruses are executed in a virtual environment to produce new viruses plaintext file, the file system filter driver is the best way to locate these new files. File-based micro-filter driver nascent virus positioning module is mainly responsible for monitoring the nascent virus production and self-destruction.

Design Objectives: The main objective of monitoring the progress of the file designated driver layer is created, the file write operations, file delete operation to locate the newly generated virus file in plain text. Driver layer also need to establish a communication port and application layer, to establish communication mechanisms. Part of the application layer is responsible for initializing the drive, monitor the process-related information will be sent to the driver layer.

Typically the process access to the file through the file related API function, the kernel function is implemented as: I / O Manager get the file access request, the request is generated based on the corresponding IRP file access, IRP contains the common file access requests, e.g. create, read, write, and finally the IRP will be sent to the file system driver, to the file system driver to complete a specific request. File system filter driver micro-encapsulation of the above file operations to the corresponding pre / post operation, complete the appropriate filtering.

The main works: a kernel driver file to get the type of operation to obtain contextual information of the target file, and then get the file object to be operated concrete path and make progress information of the operation, if the process of monitoring the path of the file list will be passed to the application layer .

Conclusions

Due to advances in computer and internet science and technology, in all areas of the actuator is great, by the problems caused by computer security, has attracted intense attention. Computer security-related issues are the most difficult threats from viruses. Windows family of operating systems has the most customers, more vulnerable to hackers of all ages. And because Windows PE files are the most widely used file formats, so the maximum impact of the PE file viruses variety. In response to PE file viruses, many methods have been proposed to detect the virus. These viruses have their own means of detection for the case. This paper presents a heuristic detection model can effectively detect unknown viruses, virus variants.The main work:

The relevant principles of the virus, classification and propagation mechanisms of computer viruses. Details of the relevant technical anti-virus software used to analyze the advantages and disadvantages of these technologies.

Starting from the PE file structure, study PE file viruses, including general technical and advanced technologies used in the PE file virus. Virus detection means of deformation virtualization technology. Summarizes the series and PE file structure associated heuristic features. In order to increase the degree of heuristics which distinguish the characteristics of the use of information entropy contribute little to streamline the classification features. The use of distance-based SVN weighted KNN classifiers, classification is better, avoid the curse of dimensionality. After selecting the appropriate test to verify the value of K, effectively reducing the false positive rate and false negative rate. Static heuristic model designed to detect, through experimental verification can detect unknown viruses

A hardware-assisted virtualization, dynamic detection model based on the model with the file system filter driver effectively detect micro self-modifying code technology virus. Dynamic heuristic model provides a virtual environment, virtual execution encrypted viruses to produce new viruses plaintext files, files are filtered capture files created driven behavior, effectively locate the position of the new file generated with a static heuristic detection model effectively detect the virus files.

Follow-up work and outlook, heuristic features used in this paper is an abstract general characteristics of the virus, although able to detect unknown viruses, but can not classify the virus can not know the specific information the virus, they can be promptly submitted to the unknown virus specifically pick extracted signature to the true purpose of detection. Dynamic heuristic model uses a file system filter driver technology, have some impact on the performance of the system. For some take-virus kernel level rootkit technology, the paper has not made a corresponding detection methods.

Acknowledgements

The research is supported by the National Science Foundation of China under Grant No. 61202455; the Doctoral Program of Higher Education Foundation of China under Grant No. 20112304120025; the Special-ized Fundation for the Basic Research Operating ex-penses Program of Central College No. HEUCF100612, HEUCFT1202.

References

[1] Peter Szor computer virus prevention Art (1st edition) new paragraph sea, Bo, WANG De-qiang Beijing: China Machine Press 2007.6- 8, 31-42, 79-118, 281-325.

[2] Zeng Ming, Zhao Cai, Yai Jingsong such as feature extraction based on binary comparison of Technology Computer En-gineering and Applications, 2006: 711.

[3] Fred Cohen.A Cryptographic Checksum for Integrity Pro-tection in Untrusted Computer Systems. Computers and Security,1987:357-360.

[4] Xu, Chen Chun, should call the classification of grain-based anomaly detection system of Software, 2004,15 (3):.. 391 ~ 403 || Wang Haifeng, segment Friends Xiang, Liu Renning improvement based on behavioral analysis of computer virus detection engine. applications, 2004,24 (22): 109 - 10.

[5] Kang M G,Poosankam P,Yin H.Renovo:A hidden code extractor for packed executables[C]//Proceedings of ACM Workshop on Recurring Malcode,October 2007..

[6] Han Xiaoqing, Wang Jianfeng, Zhong Wei computer virus analysis and prevention Daquan Beijing: Electronic Industry Press, 2006: 40-44.

[7] H. Cheng, X, Yan, J. Han, and C. Hsu. Discriminative fre-quent pattern analysis for effective classification. In ICDE-07, 2007: 101-107.

[8] M. Fan and C. Li. Mining frequent patterns in an fp-tree without conditional fp-tree generation. Journal of Computer Research and Development, 2003, 40: 1216-1222.

[9] J.H. Wang, P.S. Deng. Virus Detection using Data Mining Techniques. IEEE International Carnahan. Conference on Security Technology, pp. 2003: 71-76.

[10] H. Witten, E. Frank. Data mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, 2005: 25-31.

[11] D. Brumley, C. Hartwing, M. G. Kang, et al. Bitscope: Au-tomatically dissecting malicious binaries. In CMU-CS-07-133, 2007:56-65.

[12] C. Willems. CWSandbox: Automatic Behaviour Analysis of Malware. http://www.cwsandbox.org, 2006.

[13] Christopher Kruegel. Increase Dynamic Coverage. Secure Systems Lab Technical University Vienna, Sep. 2007.

[14] Norman. Normal Sandbox. http://sandbox.norman.no/, 2006.

[15] Zhong Luo Pan Hao, Feng Yun pattern recognition [M] • Wuhan: Wuhan University Press, 2006..

[16] SCHULTZ M, ESKIN E, ZADOKE, et al. Data mining methods for detection of new malicious executables. Pro-ceedings of the IEEE Symposium on Security and Privacy. Los Alamitos, CA, 2001: 38-49.