基于时序分析的木马控制行为识别方法

陈 利[1] 张 利[2] 姚轶崭[2] 胡卫华[2]
(中国信息安全测评中心 北京 100085)

摘 要 传统指纹识别方法在检测新型未知木马时漏报率较高。为此,提出基于时序分析的无指纹木马控制行为识别方法。该方法先对数据流进行时序分簇处理,再计算分簇数据的加权欧氏距离,通过分簇数据的时序关系来识别木马控制行为。实验表明,该方法无需特征指纹库,且检测准确率高,占用资源少,能实现实时检测和处理。

关键词 时序分析,分簇,木马控制,行为识别,入侵检测
中图法分类号 TP393.09 **文献标识码** A

Trojans Control Behavior Detection Approach Based on Timing Analysis

CHEN Li[1] ZHANG Li[2] YAO Yi-zhan[2] HU Wei-hua[2]
(China Information Technology Security Evaluation Center, Beijing 100085, China)

Abstract Traditional detection approach based on fingerprint has a higher rate of false negatives. To this end, this paper put forward a detection approach of Trojans control behavior based on timing analysis of network sessions. Firstly, it calculats the weighted Euclidean distance between clustering dataflow, then the Trojans control behavior can be detected by ti-ming relationships of clustering data. Experiments show that the approach did not need fingerprint database, and can achieve higher correct detection rate, less consumption of resource real-time detection and processing.

Keywords Timing analysis, Clustering, Trojan control, Behavior recognition, Intrusion detection

1 引言

木马作为黑客的重要攻击手段[1],对网络安全和数据保护构成了严重的威胁。传统上,检测木马的方法之一是特征指纹匹配技术[2],然而由于依赖于指纹库所带来的滞后性,指纹匹配技术在面对新型未知木马或者变种木马时往往无能为力,加之目前在木马网络通信数据流中不易提取特征指纹,传统的木马检测技术具有很大的局限性。基于行为分析的检测技术避免了特征码的问题[3],但由于木马通信行为的多样化,定义准确的行为特征是基于行为分析的检测技术待解决的关键问题。

本文提出一种基于时序分析的木马控制行为识别方法,以木马通过互联网控制服务端的行为特点为研究对象,总结控制型木马的通信机制和控制机制,发现其在网络行为特征方面的时序特性[6],通过镜像木马控制端和服务端之间交互的网络数据流,采用时序聚类[5]、欧氏距离度量[6]以及数理统计方法[7]对木马数据流进行控制行为提取,从而在网络行为层面精确识别木马控制行为。

2 木马通信中控制行为数据流特征

木马通信过程中的重要行为特征即是其控制行为,木马客户端对服务端的每一个控制操作都是通过控制指令下发,服务端接收指令完成相应操作并返回执行结果来实现的[8]。一般来说,为了方便控制,针对控制端下发的每一个控制指令,服务端都会返回指令的返回结果或者执行状态。体现在网络数据流中,由于机器执行的高速性,木马控制数据后面紧跟着木马执行数据,且方向相反,在时序上紧密相依。另外,控制端由于人工操作,需要对每一个返回结果进行查看解读,则控制操作之间存在时间差[9],从木马通信的整个过程中看,所有控制操作数据在时序上构成一个控制操作序,所有返回数据构成一个响应操作序[10],且相应操作序中每一个相应操作在控制操作序中都有一个控制操作与之一一对应。

3 基于时序分析的木马控制行为识别方法

3.1 时序分簇处理

簇是在时序上关系紧密的数据包的集合。若相邻两个数据包的时间间隔不大于时间值 T,则认为这两个数据包属于一个簇;T 根据网络状况计算而来,由网络的往返时延决定,具体计算如下:

$$T=a\times(\sum_{i=0}^{n}t(i+1)-t(i))/n \quad (1)$$

式中,函数 $t(i+1)-t(i)$ 表示当前会话中两个有交互行为的相邻数据包的时间差,代表一定时刻的往返时延。一般取最近20次不同往返时延取的平均值(即 $n=20$),然后经过放大作为 T 的取值,a 为放大倍数,本检测方法默认设置为10。

通过构造临时数据包列表,对网络数据流中的每一个数据包,判断与临时列表中最后一个数据包的时间差是否大于

本文受国家自然科学基金项目(90818021,60973105)资助。

陈 利(1986—),男,硕士,助理研究员,主要研究领域为信息安全技术、入侵检测;张 利 男,博士,副研究员,主要研究领域为风险评估技术;姚轶崭 男,博士,副研究员,主要研究领域为风险评估技术;胡卫华 女,硕士,副研究员,主要研究领域为信息安全技术。

T，是则临时列表中数据包集合构成一个簇，否则将该数据包添加至临时列表末尾，如此循环，则将网络数据流按时序分解成簇集合列表。以 pk 表示数据包，clu 表示簇，sn 表示会话，则其关系表示如下：

$$clu=\sum_{i=0}^{n} pk_i \tag{2}$$

$$sn=\sum_{i=0}^{m} clu_i \tag{3}$$

运用数理统计思想，计算首包方向 c_{dir}、首包时间 c_{time}、序列数量 c_{num}、序列均值 c_{avg}、序列纯度 c_{pure}、时间跨度 c_{tc}、序列最值 c_{max}、序列速度 c_{sp} 8 个属性。以属性向量作为簇的特征值，记为

$$pro=\langle c_{dir}, c_{time}, c_{num}, c_{avg}, c_{pure}, c_{tc}, c_{max}, c_{sp}\rangle \tag{4}$$

3.2 加权欧氏距离计算

本文利用加权欧氏距离来计算分簇之间的关联度，以反映其在网络行为上的关系，则两个簇 $pro1$、$pro2$ 加权欧氏距离 d 的计算公式如下：

$$d=\sqrt{\sum_{i=1}^{8}(w_i(pro1_{ci}-pro2_{ci})^2)} \tag{5}$$

式中，pro_{ci} 表示簇属性向量中不同属性值，w_i 表示不同属性值的计算权值。本文研究中，对木马控制行为进行时序关联分析，对于木马网络数据流中，数据包之间时序关系越紧密，则数据包之间就越可能存在行为关系，由此首包方向 c_{dir}、首包时间 c_{time} 两个属性计算权值最大，这直接影响分簇距离计算结果。

3.3 控制行为识别

通过计算木马通信数据流和正常应用数据流相邻分簇的加权欧氏距离，以时间为变量建立函数关系 $d=f(t)$，绘制坐标曲线图，其中图 1 表示木马样本数据流分簇距离采样结果，图 2 表示正常网络应用样本数据流分簇距离采样结果。

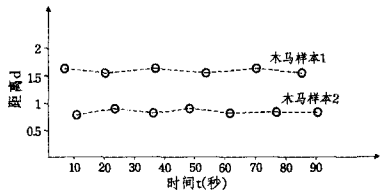

图 1 木马样本数据流分簇距离采样图

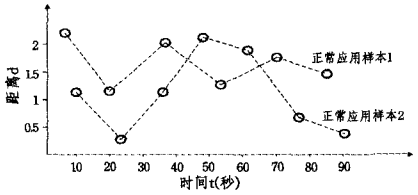

图 2 正常应用样本数据流分簇距离采样图

分析表明，正常网络应用数据流经时序分簇处理后，相邻分簇间无明显控制响应等交互行为，因此相邻分簇加权欧氏距离呈现较大的变化幅度，并且变化过程没有规律，这与正常网络应用通信过程的随机性特征相符；而控制型木马的数据结果则相反，数据流经时序分簇处理后，由于控制会话与响应会话的明显交互行为，相邻分簇加权欧氏距离呈现相对平稳的规律，这与木马的控制行为特性相符，其中簇距离计算值取决于分簇各属性计算权值。因此，通过计算相邻分簇加权欧氏距离的方差 d，判断其是否超过一定阈值 w，来检测木马控制行为。

4 实验及结果

本文实验是对大量控制型木马样本与正常网络应用数据流做控制行为识别。实验环境为 CPU：Pentium 2.49GHz，内存：2GB，开发平台：Python2.6.4，操作系统：Ubuntu10.04 Desktop。

以控制型木马灰鸽子为例，采集其网络数据流进行会话重组，并进行控制行为识别，绘制数据流时序关系图，如图 3、图 4 所示。图中横轴代表时间，纵轴柱形代表灰鸽子数据包，柱形长度代表包大小，顶端数字标识具体值，正负代表方向，正数代表该柱形是控制端发向服务端的数据包，反之，负数代表服务端发向控制端的数据包。等长虚线代表时间分隔线，一条虚线代表 1 秒间隔，两分隔线之间的数据包在时间上属于同一秒。

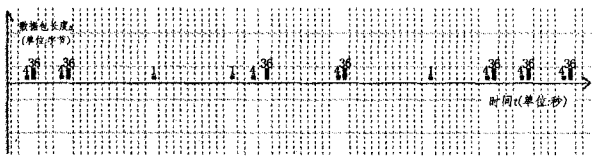

图 3 灰鸽子木马控制会话部分时序图

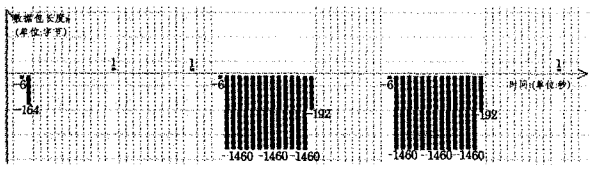

图 4 灰鸽子木马响应会话部分时序图

检测过程中，首先获取网络数据流，计算出簇间最小时间间隔 T 为 1s，经过时序分簇处理，得到 21 个分簇，通过计算得到加权欧氏距离均值为 0.45，距离方差为 0.001，该值小于设定方差阈值 w，从而识别出该灰鸽子木马的控制行为。大量实验证明，方差阈值 w 设定为 0.05 对所有木马样本具有最好的识别效果。灰鸽子数据流分簇距离采样如图 5 所示。

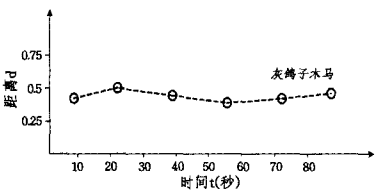

图 5 灰鸽子木马数据流分簇距离采样图

本文选取了 96 个有控制功能的木马样本和 322 个正常网络应用对木马和正常应用在真实网络环境中所产生的网络数据流进行控制行为识别。实验结果如表 1 所列。

表 1 识别结果统计

实验环境	网络状况	96 个控制型木马样本			322 个正常网络应用		
		识别出数(个)	正确识别率(%)	平均识别时间(s)	识别出数(个)	错误识别率(%)	平均识别时间(s)
实验网	良好	96	100	0.95	0	0	1.01
公网	拥塞	93	96.87	2.19	4	1.24	2.26

经过大量样本实验，从正确识别率、错误识别率、平均识别时间3个方面证明了该识别方法的可用性。由表1中统计数据可以看出，该检测方法达到了比较高的正确识别率，将错误识别率降低在5%以下，同时耗费了很少的CPU运算时间，证明了该识别方法的可用性，对木马数据流特征的分析具有重要的意义。

结束语 本文在对木马会话数据流进行时序分析的基础上，提出基于时序分析的无指纹木马控制行为识别方法。该方法首先对木马网络数据流进行时序分簇处理，然后计算分簇数据的加权欧氏距离，利用木马数据流分簇距离的平稳分布特性进行判别，从而识别出木马控制行为。实验表明，该识别方法能够有效地识别出一般木马数据流中的控制行为，具有较高的正确识别率；同时该方法高效地利用了处理器资源，在高速的网络环境中也能做到实时处理。此外，该方法已经被实际应用于网络入侵检测和木马行为监控等实时网络数据流分析当中，并取得了较好的效果，表明该方法具有很强的实用性。但是方法仍存在一些不足，即在时延较大的拥塞网络状况下，准确识别率略有降低，如何在拥塞网络环境下保证较高的准确识别率还有待进一步研究。

参 考 文 献

[1] Zhang Li-ke, White G B. An Approach to Detect Executable Content for Anomaly Based Network Intrusion Detection[C]// Proc. of Parallel and Distributed Processing Sysmposium. Long Beach, USA: [s. n], 2007: 1-8

[2] 井小沛，汪厚详，聂凯，等. 面向入侵检测的基于IMGA和MKS-VM的特征选择算法[J]. 计算机科学，2012，39(7)：262-264

[3] Nie Fei-ping, Xiang Shi-ming, Jia Yang-qing, et al. Trace Ratio Criterion for Feature Selection[C]// Proceedings of National Conference on Artificial Intelligence. Chicago, USA: [s. n], 2008: 672-675

[4] Wang Sui-yu, Baird H S. Feature Selection Focused Within Error Clusters[C]// Proceedings of the 19th IEEE ICPR'08. [s. l]: IEEE Press, 2008: 1-4

[5] 易军凯，陈利，孙建伟. 网络心跳包序列的数据流分簇检测方法[J]. 计算机工程，2011，37(24)：201-524

[6] Nehinbe J O. Automated technique for debugging network intrusion detection systems[A]// IEEE 2010 International Conference on Intelligent Systems, Modelling and Simulation (ISMS)[C]. Liverpool, 2010: 363-367

[7] Wuu L C, Hung C H, CHEN S F. Building intrusion pattern miner for Snort network intrusion detection system[J]. Journal of Systems and Software, 2007, 80(10): 1701-1714

[8] 郭文忠，陈国龙，陈庆良，等. 基于粒子群优化算法和相关性分析的特征子集选择[J]. 计算机科学，2008，35(2)：113-147

[9] 陈友，沈华伟，李洋. 一种高效的面向入侵检测系统的特征选择算法[J]. 计算机学报，2007，30(8)：1395-1407

[10] 陈友，程学旗，李洋，等. 基于特征选择的轻量级入侵检测系统[J]. 软件学报，2007，18(7)：1639-1650

(上接第333页)

[2] Sundaram A. An Introduction to Intrusion Detection [J]. Crossroads, 1996, 2(4): 3-7

[3] Forrest S, Hofmeyr S A, Somayaji A, et al. Sense of self for Unix processes [C]// Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy. Oakland, CA, USA: IEEE Computer Society Press, 1996: 120-128

[4] 吴瀛，江建慧，张蕊. 基于系统调用的入侵检测研究进展[J]. 计算机科学，2011，38(1)：20-25

[5] Forrest S, Hofmeyr S A, Somayaji A. Intrusion Detection Using Sequences of System Calls [J]. Journal of Computer Security, 1998, 6(3): 151-180

[6] Liao Yi-hua, Vemuri V R. Use of K-nearest Neighbor Classifier for Intrusion Detection [J]. Networks and Security, 2002, 21(5): 438-448

[7] Rawat S, Gulati V P, Arun K P, et al. Intrusion Detection Using Text Processing Techniques with a Binary-Weighted Cosine Metric [J]. Journal of Information Assurance and Security, 2006, 1(1): 43-50

[8] Jecheva V, Nikolova E. An adaptive KNN algorithm for anomaly intrusion detection [C]// Interaction of theory and practice: key problems and solutions. Burgas Bulgaria: Burgas Free University, 2011: 198-204

[9] 吕锋，刘泉永. 利用KNN算法实现基于系统调用的入侵检测技术[J]. 微计算机信息，2006，22(93)：76-78

[10] Forrest S, Warrender C, Pearlmutter B. Detecting Intrusions Using System Calls: Alternate Data Models[C]// Proceedings of the 1999 IEEE ISRSP. IEEE Computer Society, Washington, DC, USA, 1999: 133-145

[11] Tax D M J, Duin R P W. Support Vector Data Description[J]. Machine Learning, 2004, 54(1): 45-66

[12] University of New Mexico. Computer Immune Systems Project [OL]. http://www.cs.unm.edu/~immsec/systemcalls.htm

[13] Budalakoti S, Srivastava A, Otey M. Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety [J]. IEEE Transactions on Systems, Man and Cybernetics (Part C: Applications and Reviews), 2009, 39(1): 101-113

[14] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data set[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, TX, United states: IEEE Computer Society Press, 2000: 427-438

[15] Thomas C, Sharma V, Balakrishnan N. Usefulness of DARPA dataset for intrusion detection system evaluation[C]// Proceedings of SPIE-The International Society for Optical Engineering. Orlando, FL, United States: IEEE Computer Society Press, 2000: 220-237

[16] Cerioli A, Farcomeni A. Error rates for multivariate outlier detection [J]. Computational Statistics and Data Analysis, 2011, 55(1): 544-553

[17] Fawcett T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874