文章编号:1008-0570(2007)06-3-0066-03

蠕虫病毒特征码自动提取原理与设计

Toward Automated, The Principle of Extracting Worm Signature

(临沂师范学院)王晓洁
WANG XIAOJIE

摘要:目前网络入侵检测系统(NIDS)主要利用特征码检测法来监测与阻止网络蠕虫,而蠕虫特征码提取仍是效率低的人工过程。为解决这个问题提出了基于陷阱网络的蠕虫特征码自动提取思想,介绍了原型系统的体系结构和主要算法。该系统利用数据包负载中出现频率高的字符串来提取蠕虫特征码。最后通过实验结果分析算法主要参数对系统的影响。

关键词:蠕虫特征码; 陷阱网络; 模式检测;入侵检测
中图分类号:TP39.05　　文献标识码:B

Abstract: Today's Network intrusion detection system (NIDS) monitor edge network to identify and filter malicious flows. NIDS needs worm signatures to contain the spread of a worm. The creation of these signatures is a tedious, manual process that requires detailed knowledge. To solve this problem, this paper describes a system for automated generation of worm signatures for NIDS. Our system generates signatures by analyzing the prevalence of portions of flow payloads. This paper presents the key algorithm and prototype. In the end, the experiment results are provided.

Key words:Worm Signature , Honeynet , Pattern Detection, Network Intrusion Detection

1 引言

近几年爆发过的一系列蠕虫病毒给信息社会造成巨大经济损失,因此对蠕虫病毒的研究仍然是网络安全研究重点。蠕虫病毒是一种利用软件漏洞实现自动传播和破坏的人为恶意程序。由于同构型网络环境中运行的操作系统与服务器软件缺乏多样性,因此蠕虫在 Internet 或 Intranet 中能够迅速蔓延。目前主要有以下两种蠕虫检测技术:流量检测。在网络全局范围内监测可疑随机扫描流量。如果发现可疑 IP 地址,则将其加入 IP 黑名单,列入下一步过滤范围。对利用随机扫描传播的蠕虫该技术十分有效, 但是对于邮件病毒和 P2P 蠕虫检测效果差,由于以上两类病毒不使用 IP 随机扫描;另一种技术是行为检测。因为蠕虫病毒是一种非典型性的 Internet 应用程序, 所以在单机范围内可监测其特异程序行为来发现蠕虫。此项技术缺乏网络层次的检测能力,对蠕虫检测存在很大滞后性。总之,尽量阻断被感染主机试图与潜在受害主机的连接是蠕虫检测的主要策略。

目前商业网络入侵检测系统 NIDS(Network Intrusion Detection System)多数采用比较成熟的误用检测技术,检测网络范围内可疑数据包,如果发现与入侵特征库中相匹配的数据包就向网络管理员发出警报。NIDS 是一种体现主动防御思想的安全工具,其特征规则检测法结合了流量检测和行为检测。NIDS 的特征规则一般表示为一个三元组<协议类型,目标端口,字符序列>, 由于蠕虫通常针对某个端口服务程序的漏洞来实现传播和破坏,因此协议类型、目标端口体现了蠕虫网络层面的特点;另一方面,因为蠕虫行为的非典型性,比如利用溢出实现攻击、自我复制等功能。所以能够提取反映该功能的机器码序列作为检测的依据,这些字符序列即蠕虫的特征码。总之,NIDS 采用基于内容过滤、阻断的策略防御蠕虫。

王晓洁: 硕士　副教授

NIDS 检测的关键是检测规则。首先对检测规则提取时间要求高。Internet 主机对蠕虫的免疫有很高的时间要求。对传播速度慢的蠕虫要求最多 60 分钟作出免疫反应,比如 RedCodeII;高速传播的蠕虫要求 5 分钟甚至 60 秒的免疫时间。另外特征码质量对 NIDS 工作效率影响很大。因此蠕虫特征码的提取工作十分重要, 现在主要由安全专家人工提取蠕虫特征码, 这是个费时、枯燥并且技术水平要求高的工作。人工提取蠕虫特征码需要较长时间,导致了 NIDS 对蠕虫检测存在很大滞后性甚至失败。普通计算机病毒特征码大约有%5 ~ %6 来自陷阱机 Honeypot 捕获的数据,蠕虫与普通病毒不同,在网络中传播速度快,设计精巧的陷阱机能及时有效地捕获到新蠕虫或蠕虫变种。本文提出一个基于陷阱网络 Honeynet 的蠕虫特征码自动提取原型系统 Cuckoo,描述原型系统的体系结构与自动提取原理,最后通过实验分析蠕虫特征码的质量。本文以后各节组织如下: 第二节,介绍蠕虫特征码自动提取的相关研究;第三节,阐述了原型系统 Cuckoo 的体系结构,特征码提取的流程与算法,以及特征码广谱优化策略;第四节,通过实验分析原型系统中重要参数的作用与影响;第五节,总结并提出进一步研究的设想。

2 相关研究

Honeycomb 是利用陷阱网络生成入侵检测规则的系统,该系统以开源代码 Honeyd 为基础, 通过协议分析与内容匹配生成入侵检测系统 IDS 的规则。Honeycomb 采用 LCS 算法生成入侵特征规则,自动生成的检测规则中包括蠕虫检测规则。但是 Honeycomb 运行效率低,生成的蠕虫检测规则质量不高,比如检测特征码较长,这导致 Snort 或 Bro 等入侵检测系统效率低。

Autograph 是 Intel 研究组研发的自动蠕虫特征码提取系统。Autograph 体系合理,运行效率高,可以生成较高质量的蠕虫特征码。Autograph 前端是基于启发式的可疑数据包分类器,而本文 Cuckoo 前端基于陷阱网络。Cuckoo 设计的许多重要思

路来源于 Autograph，但是 Cuckoo 对主要算法进行了一些改进。利用多个休止标志划分字符串因子以及编辑距离计算字符串之间相似程度，最后对蠕虫特征码进行广谱优化。Cuckoo 受实验环境限制，目前没有完成重要参数设定和运行效率分析等实验。

3 特征码自动提取原理与分析

3.1 Cuckoo 的体系结构

由陷阱机 Honeypot 组成的陷阱网络 Honeynet 是一种安全资源，陷阱网络的设计目的是吸引黑客与蠕虫的扫描、攻击、攻陷。所有进入或发出陷阱网络的流量都可能预示着扫描、攻击、攻陷。由于蠕虫病毒只是程序，不具备智能分析能力，因此采用低交互陷阱网络就能完成蠕虫的捕获任务，Cuckoo 系统重点分析流入陷阱网络的数据包。该原型系统是在开源陷阱网络系统 Honeyd 的基础上实现，以软件插件的形式集成到 Honeyd 中，利用钩子技术截获进入 Honeyd 系统的网络数据包，然后自动特征码生成引擎负责提取新特征码并存入数据库。

Honeyd 通过虚拟拓扑组件实现多个 IP 地址的虚拟主机，系统从真实环境接受到的数据经包分发器进行处理，Honeyd 系统利用配置数据库中的模板驱动个性引擎与服务子系统来模拟不同的操作系统和不同的服务。特征码生成引擎利用 Hook 技术截获包分发器接受的网络数据包，经分析后提取蠕虫的特征码，最后输出、发布。

3.2 特征码生成原理

Cuckoo 特征码生成引擎不依靠专家知识工作。而且自动生成蠕虫特征码，并在一定范围内进行特征码优化。下面介绍特征码提取流程。

特征码生成引擎从 Honeyd 包分发器截获进入陷阱网络的数据包。因为在低交互的陷阱环境中，所以只根据流入数据包作为生成特征码的依据；然后进行协议分析，目前仅分析 TCP 与 UDP 两种协议类型；对数据包大小进行判断，由于蠕虫传播需要携带一定长度的恶意代码，该步骤可以过滤掉一般的网络攻击数据包，为以后分析提供质量更好的数据；接着将流入的数据包压入堆栈，堆栈在规定时间充满时将激活特征码提取过程，产生成一条特征码；最后尝试对特征码进行优化处理，处理完毕后继续以上过程。

3.3 特征码生成算法

特征码生成算法的主要思想如下：因为蠕虫感染陷阱网络后，短时间内陷阱机会流入大量携带蠕虫机器码的网络数据包，因此当接收流入数据包的缓冲区在 t 时间段内充满时就会生成一条特征码。首先将每个数据包负载划分为变长字符串因子。然后利用字符串编辑距离计算字符串之间的相似度，相似度小于 d 的字符串因子就编为一个候选因子队列。最后对候选因子按出现频率进行排序，应用贪婪法选择高频候选因子组成特征码。具体描述如下：

P表示数据包负载 p_i 的集合；R 表示字符串因子 r_i 的集合；S 表示候选因子 s_i 的集合；Q_i 表示与候选因子 s_i 相似的字符串因子队列；C 表示特征码 c_i 的集合。

```
AutoGenerate_Signature( P, m, M, d, l )
{
P:    数据包负载集合；
m,M: 字符串因子大小阈值；
d:    编辑距离阈值；
l:    特征码最大长度
S<--{ } // 候选因子集合赋空字符串
Found <-- false
Repeat {
If Size (P) >θ And DelayTime < t then
{
Repeat for every pi in P={pi, i = 1,2,…}
{ ri<-- (pi divided into a small block)
if length(ri) ∈ [m,M] then
R<-- R ∪{ri}
}
Repeat for every ri in R={ri,i = 1,2,…}
{ Repeat for every si in S
{ if Editdistance(ri,si) < d then
{ Add(Qi, ri); Found <-- true }
}until Found = true
if Found = false then S <-- S ∪{ri}
Found <--false
}
Sort(S) //候选因子集按出现频率高低排序
Repeat for each si in SortedS = {si, i = 1,2,…}
ci <-- ci ∪ { si}
Until length (ci) > l
C <-- C ∪ {ci}
}until Size(C) > Maxsize
```

算法首先将数据包负载划分为多个字符串因子，在此改进 Autograph 系统基于内容的 Rabin fingerprint 划分方法。Rabin fingerprint 可以将长字符串转化为等价短字符指纹，并且计算效率高。基于字符内容的划分法适合处理蠕虫的多态性，无论删除、插入字符都能划分出正确的字符串因子。对于每个数据包负载按照一定滑动窗口 k 向后滑动，计算出每 k 个字符串的 Rabin fingerprint 为 r_i (i = 1,2,…)；若滑动到某个位置时 r_j 与预置休止标志 B 匹配，就从前一个休止标志到此休止标志划分为一个字符串因子，继续该过程直到结束。改进之处是使用随机多休止标志。分析已知蠕虫特征码，根据专家经验选取蠕虫机器码中出现频率最高的字符作为待选休止标志，然后在整个划分过程中随机选择多休止标志。

采用编辑距离(Edit Distance)计算字符串相似性，编辑距离源于模式识别中最近邻分类器思想。字符串 x 与字符串 y 的编辑距离描述从 x 变到 y 所需最少基本操作，这些基本操作包括替换、插入、删除，每个基本操作都将增加距离 0.05。例如：串 x = [aabbcc]，串 y = [abbccc]，则 Edit-distance < x , y > = 0.1

3.4 特征码优化算法

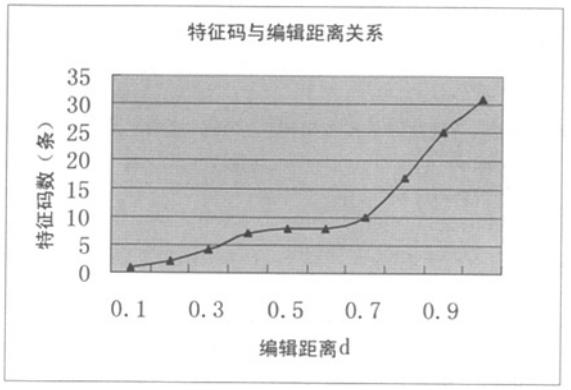
假设特征码 c_m 由候选因子 s_j 与 s_{j+k} 构成，即 $c_m = < s_j , s_{j+k} >$。改变为适应蠕虫多态性与变种的广谱特征码是优化 c_m 的主要思路。s_j 对应的队列 Q_j 中存储了所有与 s_j 因子相似的字符串因子。优化过程从位序 1 开始，统计队列所有串在位序 i 出现的字符，若某字符出现的频率大于 0.5，则优化因子 s_j' 在位序 i 填入该字符；若位序 i 出现的字符呈现均匀分布，则优化因子 s_j' 在位序 i 填入通配符 *。取队列中字符串的平均长度作为优化因子 s_j' 的长度，剩余部分舍弃。s_{j+k} 与 s_j 处理方法一样，最后优化特征码 $c_m' = < s_j', s_{j+k}' >$。

4 实验结果与分析

首先，讨论陷阱网络采集时间间隔与特征码生成的关系，

寻找最佳时间间隔。实验环境是校园网,由于校园网存在几种常见的蠕虫病毒,因此可以进行特征码自动生成实验。字符串因子长度限制为;滑动窗口 k = 4;特征码长度限制 l = 100;编辑距离阈值 d = 0.7;采用变休止标志划分字符串因子。当采集时间间隔过小或者过大时,特征码很难生成,而间隔时间为 12~20 分钟时可生成大量特征码。实验结果与蠕虫快速传播特性相吻合,蠕虫在一定时间内迅速感染网络,堆栈在规定时间内充满,将激活特征码生成引擎工作。时间过短或者过长,堆栈不能充满就被清除了,因此几乎不能生成特征码。

特征码自动提取与编辑距离阈值 d 的关系是第二个实验重点,结果如图所示。在其它参数不变的前提下,不断修改编辑距离阈值 d。图中拐点出现在 0.7 附近,当小于 0.7 时,特征码生成速度慢;而当大于 0.9 时,特征码生成速度过快,特征码质量明显下降。所以实验中编辑距离阈值 d 一般选 0.7-0.8 之间。



特征码质量是自动提取系统的关键。设 Cuckoo 能识别的蠕虫数据包数为 n;所有包含蠕虫数据包数为 m,其余不包含蠕虫的数据包数为 v,则特征码敏感度定义为 $S=\frac{n}{m+v}$。因此敏感度越高,蠕虫特征码检测能力越好。选取 10000 数据包作为测试集,其中 200 个数据包中含有 Slammer 蠕虫。使用不同长度的字符串因子,提取 Slammer 特征码并用测试集进行检测,实验发现提取特征码越长,敏感度越高。但是如果特征码过长,检测效率降低。实验发现字符串因子控制在 64-128 之间,效果比较理想。

5 总结

Cuckoo 目前仅是简单的原型系统。由于对 Honeyd 有很强的依赖性,因此功能存在较多局限性。日后研究工作试图建立脱离 Honeyd 而独立存在的蠕虫特征码自动提取系统。尽管如此,该系统证明了利用陷阱网络自动提取蠕虫特征码的可行性。Cuckoo 能够生成比 Honeycomb 更短、更高效的蠕虫特征码。但是特征码质量评价缺少更准确、更权威的实验,特别是 NIDS 使用蠕虫特征码检测效率等问题。同时该系统的思想可以应用于垃圾邮件过滤方面,自动提取垃圾邮件的检测内容。

课题创新点:国内几乎没有涉及蠕虫病毒检测特征的提取问题,本文是在深入了解国外相关研究基础上实现基本原型系统,并证明了陷阱网络工具在自动挖掘蠕虫特征码研究中的可行性。

参考文献

[1]王海峰,段友祥. 针对计算机黑客型病毒的网络防御体系研究[J].2004.6 (4-6).

[2]王振海,王海峰.针对多态病毒的反病毒检测引擎的研究[J]微计算机信息 2006.29

[3]H.Kim,B.Karp, "Autograph:Toward auto-mated,distributed worm signature detection",Proceedings of the 13th Usenix Security S-ymposium"[J],2004

[4]T.Destristan etc, "Polymorphic shellcode engine using spectrum analysis"[J], Pharck Magazine, 11(61),2003

[5]RABIN,M.O. Fingerprinting by Random Polynomials[J]. Tech. Rep.TR -15 -81,Center for Research in Computing Technology, Harvard University,1981.

[6]L.Gao etc, An effective architecture and algorithm for detecting worms with various scan techniques. Proceedings of NDSS[C],2004

[7]Richard O.Duda Peter E.Hart David G.Stork , Pattern Classifi-cation.2th[M],北京,机械工业出版社,2003.9,336

作者简介:王晓洁(1973),女,山东省临沂市人,硕士,副教授。

Biography: Wang Xiao-Jie, borned in 1973. Master of Computer Science.

(276002 山东临沂市 临沂师范学院信息学院)

王晓洁

(Department of Computer, University of LinYi Normal Teacher, Lin yi 276002, China)Wang XiaoJie

通讯地址:(276002 山东临沂市 临沂师范学院信息学院)

王晓洁

(收稿日期:2007.5.03)(修稿日期:2007.6.05)

(上接第 72 页)

参考文献

[1]南湘浩、陈钟,网络安全技术概念[M],北京:国防工业版社,2003。

[2]卿实汉 安全协议[M] 清华大学 2005 。

[3]安全协议的建模与分析:CSP 方式[M] 机械工业出版社 2005

[4]石曙东,李芝棠 一种安全协议的逻辑分析与改进[J]华中科技大学学报(自然科学版),2004,07。

[5]边培泉. 基于逻辑的电子商务协议属性的分析与研究[D]兰州理工大学 , 2004。

[6]肖德琴,周权,张焕国,刘才兴. 基于时序逻辑的加密协议分析[J]计算机学报 , 2002,(10) 。

[7]张玉清. 计算机通信网安全协议的分析研究[D]. 西安电子科技大学 , 2000。

[8]安靖,王亚弟,韩继红, 安全协议的 CSP 描述技术[J]微计算机信息, 2006, 10-3: P52-55。

作者简介:李新中(1968.6-),男,汉族,河南沁阳人,河南省焦作市焦作大学电大教学部讲师,在职研究生。研究方向:计算机网络。

Biography: Li Xinzhong(1968.6-), Male, Han Nationality, Born in Henan Qinyang, Lectuer in Department of Broadcasting and Television Teaching, Jiaozuo University, Postgraduate Student, Research Fields: Computer Networks

(454003 河南省焦作市 焦作大学电大部)李新中 周小燕

(Department of Broadcasting and Television Teaching, Jiaozuo University, Jiaozuo Henan, 454003, China)

Li XinZhong Zhou XiaoYan

通讯地址:(454003 河南省 焦作市焦作大学电大教学部)

李新中

(收稿日期:2007.5.03)(修稿日期:2007.6.05)

蠕虫病毒特征码自动提取原理与设计

作者：王晓洁， WANG XIAOJIE
作者单位：276002,山东临沂市,临沂师范学院信息学院
刊名：微计算机信息
英文刊名：MICROCOMPUTER INFORMATION
年，卷(期)：2007,23(18)

参考文献(7条)

1. 王海峰;段友祥 针对计算机黑客型病毒的网络防御体系研究[期刊论文]-微型机与应用 2004(4-6)

2. 王振海;王海峰 针对多态病毒的反病毒检测引擎的研究[期刊论文]-微计算机信息 2006(27)

3. H Kim;B Karp Autograph:Toward auto-mated,distributed worm signature detection 2004

4. T Destristan Polymorphic shellcode engine using spectrum analysis 2003(61)

5. RABIN M O Fingerprinting by Random Polynomials[Tech.Rep.TR-15-81,Center for Research in Computing Technology Harvard University] 1981

6. L Gao An effective architecture and algorithm for detecting worms with various scan techniques 2004

7. Richard O Duda;Peter E Hart;David G Stork Pattern Classification 2003

本文读者也读过(10条)

1. 涂浩.李之棠.柳斌.TU Hao.LI Zhi-tang.LIU Bin 一种基于特征提取的高效蠕虫自动防御系统[期刊论文]-小型微型计算机系统2009,30(6)

2. 冯朝辉.王东亮.Feng Zhaohui.Wang Dongliang 基于TRAP SERVER变形病毒特征码的分析与定位技术[期刊论文]-网络安全技术与应用2006(7)

3. 李茜.Li Qian 基于特征码扫描的挂马监控技术研究[期刊论文]-科技广场2010(7)

4. 李志东.云晓春.杨武.辛毅.LI Zhi-dong.YUN Xiao-Chun.YANG Wu.XIN Yi 基于公共特征集合的网络蠕虫特征码自动提取[期刊论文]-计算机应用2005,25(7)

5. 崔翔.季振洲.袁权 "中国黑客"病毒三线程结构分析[期刊论文]-计算机工程与应用2003,39(4)

6. 谢丰.孟庆发.XIE Feng.MENG Qing-fa 蠕虫预警技术研究与进展[期刊论文]-计算机应用研究2006,23(10)

7. 李镇伟.LI zhen-wei 基于复杂网络的校园网络蠕虫病毒抑制研究[期刊论文]-常熟理工学院学报2008,22(10)

8. 王少华.刘忠强 变形蠕虫特征码自动提取算法研究[期刊论文]-山东电大学报2008(1)

9. 刘杰.迟利华.胡庆丰.LIU Jie.CHI Li-hua.HU Qing-feng 优化并行计算的性能评价[期刊论文]-计算机工程与设计2000,21(6)

10. 王平.方滨兴.云晓春.WANG Ping.FANG Bin-xing.YUN Xiao-chun 基于分割的蠕虫传播抑制方法[期刊论文]-北京邮电大学学报2006,29(5)