基于数值分析的异常扫描行为监测系统①

荆 涛[1,2], 李 俊[1]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学, 北京 100049)

摘 要: 提出了一种基于数值分析的异常扫描行为监测方法, 以 Netflow 网管数据为基础, 设计开发了监测系统, 实现了对网络中主流网络蠕虫病毒、IRC 僵尸木马的传播爆发以及黑客恶意扫描探测等异常行为的实时监测, 取得良好效果, 大幅提升了网络运营单位的网络安全支撑服务能力.

关键词: 异常扫描; 行为监测; 网络安全

Abnormal Scan Behavior Monitoring System Based on Numerical Analysis

JING Tao[1,2], LI Jun[1]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: This paper introduces an abnormal scan behavior monitoring methods based on numerical analysis. It designs and develops a monitoring system with Netflow network management data, to achieve real-time monitoring for abnormal behaviors such as network worms, IRC Trojan zombie outbreak and spreading of malicious scan by hackers, etc. The system shows significant implement results, and increase network security support service capabilities of ISP.

Key words: abnormal scan; behavior monitor; network security

1 引言

根据中国互联网络信息中心(CNNIC)发布的《第 31 次中国互联网络发展状况》[1], 截至 2012 年 12 月底, 中国网民规模达到 5.64 亿. 另外, 国家互联网应急中心(CNCERT/CC)发布的《2012 年我国互联网网络安全态势综述》[2]显示, 木马和僵尸网络依然对网络安全构成直接威胁. 2012 年, 国家互联网应急中心全年共发现 1419.7 万个境内主机 IP 地址感染了木马和僵尸程序; 分布式拒绝服务攻击依然危害网络安全, 2012 年, 分布式拒绝服务(DDoS)攻击依然是影响互联网运行安全最主要的威胁之一, 我国境内日均发生攻击流量超过 1G 的较大规模 DDOS 事件 1022 起, 约为 2011 年的近 3 倍, 呈现转嫁攻击和大流量攻击的特点.

伴随着网络攻击方式的日益复杂, 在高速网络环境下网络行为, 尤其是僵尸木马传播和 DDOS 攻击行为的快速诊断和甄别, 成为当前高速网络环境下重要的研究领域. 网络行为分析(Network Behavior Analysis-NBA)[3]往往与网络行为异常的探测相结合, 主要是利用对网络的被动观察和描述找到通讯和应用中违反安全策略的行为.

结合僵尸木马和网络蠕虫的传播特点和网络流量变化特征, 本文提出了一种基于数值分析的异常扫描行为监测实现方法, 能够对 Slammer、冲击波、震荡波、魔波、SSH 和 SQL Server 口令猜解扫描等异常探测行为进行实时监测, 取得很好的效果.

2 研究分析

为了能够对主流的网络蠕虫、病毒和扫描探测行为进行快速预警, 针对网络运营单位的实际情况, 可基于现有网管数据集合, 进行拟合分析和模型处理, 快速地发现网络中的流量异常, 将影响高速互联网络的各个因素进行量化描述, 从全局把握整体网络安全

① 基金项目:中国科学院“十二五”信息化专项——中国科学院网络安全保障与服务工程
收稿时间:2013-06-09;收到修改稿时间:2013-07-04

状况，及时发现比如 DDoS 攻击、僵尸网络、大规模蠕虫、木马及其他网络攻击行为.

目前对于网络异常行为的分析主要基于两种数据源: 协议嗅探捕获数据(利用嗅探技术捕获的原始协议报文)和网络管理数据(如: Cisco 的 Netflow)，利用协议嗅探进行捕获的数据对捕获的设备、实时分析以及处理分析能力方面，但在处理分析能力方面现阶段还无法跟上主流网络带宽的发展，存在瓶颈问题，在当前互联网高速发展的年代，嗅探捕获数据的处理能力的有限，无法对大带宽流量进行提取分析，从保存时间上来看，也存在存储空间的问题. 因此对高速网络的行为采用 Netflow 的网络管理数据作为研究分析对象是比较合适的. 对 Netflow 数据进行网络安全行为分析研究工作，我们可以按 Netflow 数据格式进行数据分析:

Netflow 数据格式包含了协议、端口、服务类型、链路流量、IP 地址等各类网络层和传输层协议信息(如图 1)，并由各种属性信息数据进行统计计算，得到流量、端口、IP 地址变化的分布规律[4]，并进行深入分析.

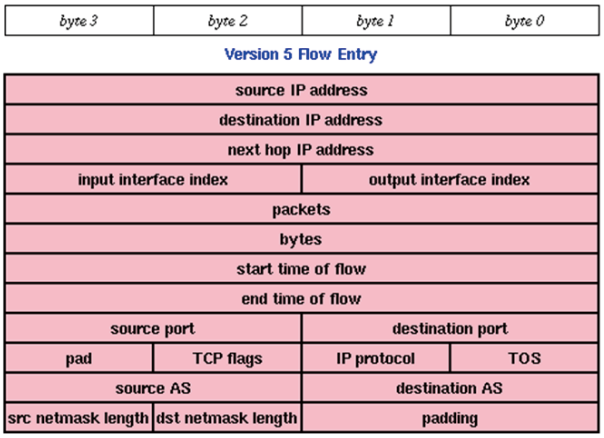

图 1　NetFlow V5 流记录格式

利用网络设备产生的 Netflow 数据分析蠕虫爆发、DDOS 攻击和僵尸木马扫描等行为，可将其数据包中包含的 IP 地址、端口、TCP 标志等网络层和传输层协议信息进行提取和整理，而网络行为的发起和结束的体现对象是 IP 地址，因此从 IP 地址作为研究对象出发，进行具体分析，是一个比较实际的研究思路，需要做的工作:

(1) 研究集中常见异常扫描行为的传输层特征，包括端口、报文大小等信息，并进行研究归类.

(2) 找出不同异常行为在源、目的 IP 以及端口信息的对应关系，并加以研究分析.

(3) 依照提取的数据集合和地址、端口的对应关系，设计算法模型，分析地址发生异常行为的判别概率.

(4) 设计系统功能模块，建立数据库进行系统实现.

在具体实现的时候，可以先将 Netflow 采集数据进行预处理，筛减无关数据类型，并同时整理计算特征属性值(特定短暂时间内，某一地址的目的地址个数、端口数、地址格式化)，预处理后进行数据的特征提取; 最后通过数据模型进行分析并得出结论.

对于现有的 Netflow 数据集所包含的信息而言，能够比较充分地涵盖当前网络节点及节点关系间的关键属性数据，通过构造的网络异常扫描行为分析模型，来发现新的网络行为和异常，也都是可行的.

3　系统实现

3.1　模块设计

我们通过以上分析，对系统进行模块设计，主要包括: 数据采集、数据筛选、数据存储、分析、展示 5 个功能模块(如图 2).

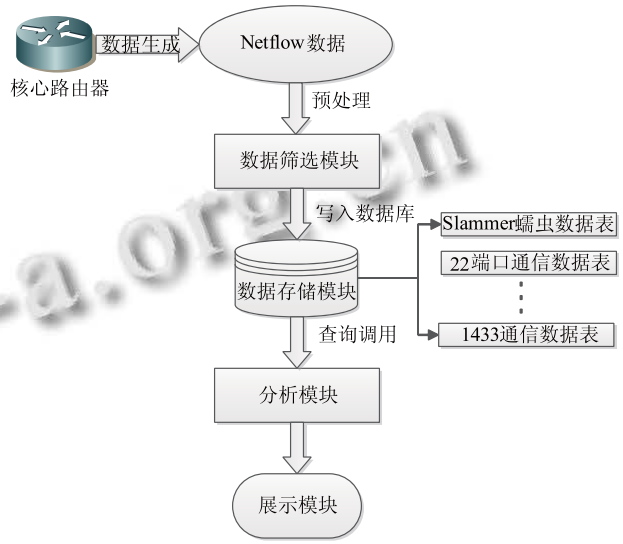

图 2　系统模块图

3.1.1　数据采集模块

数据采集模块是将核心路由器设置并生成的 Netflow 数据采集过来，由于 Netflow 数据通常是分时段生成，因此采集频率往往设置为 5 分钟一次. 在数据采集后，进行基本的预处理操作，将 flow 格式文件转换成文本文件，为之后的数据筛选做准备.

3.1.2 数据筛选模块

数据筛选模块在收到 Netflow 数据后, 按照蠕虫、僵尸木马扫描等行为特征, 并进行筛选. 具体的情况如下:

① Slammer 蠕虫: 利用的是微软 SQL Server 2000 的 SP3 以前版本的缓冲区溢出漏洞, 报文长度为 376 字节大小, 目的端口为 UDP1434;

② SSH 扫描: 黑客进行 Linux 的 SSH 应用口令猜解时, 往往先要探测 22 端口的开放情况, 再根据开放 IP 的列表进行口令列表的暴力破解行为, 因此我们可以针对其扫描探测的特点进行筛选. 具体为: TCP 协议, 目的端口 22, TCP 标志位为 SYN, 报文长度为含包头的最小长度(通常为 44、48、66 字节等);

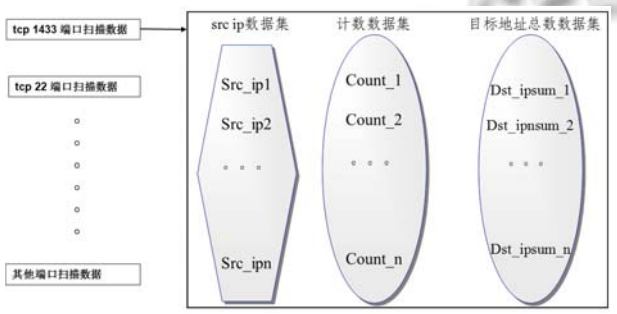

图 3 数据集结构图

③ SQL Server 端口扫描: 该扫描往往伴随着黑客对 SQL Server 的数据库服务端口的端口探测行为, 如该应用程序的守护程序监听端口开放, 则准备之后的暴力破解行为, 当然也有口令蠕虫的情况, 其具体特征为: TCP 协议、目的端口 1433, TCP 标志位为 SYN, 报文长度为含包头的最小长度;

④ 2967 端口扫描: 对 SQL Server 的扫描探测往往在扫描 1433 端口的同时也扫描 2967 端口, 当然, 也有像利用 Norton 杀毒软件溢出漏洞而进行传播的 witty 蠕虫, 也对 2967 端口有扫描探测行为. 其具体特征为: TCP 协议、目的端口 2967, TCP 标志位为 SYN, 报文长度为含包头的最小长度;

除了以上扫描情况外, 还有一些针对微软 RPC 溢出漏洞进行扫描探测的异常行为, 往往利用微软的 RPC 服务端口或者共享端口; 此外利用 IRC 协议远程控制的僵尸木马往往在进行扫描渗透的过程中, 也会利用 135、139、445 等服务端口[5], 当然也包括上述的 1433 和 2967 等服务端口. 因此, 类似的异常行为可按以下特征表述.

⑤ 139 端口扫描: TCP 协议、目的端口 139, TCP 标志位为 SYN, 报文长度为含包头的最小长度;

⑥ 135 端口扫描: TCP 协议、目的端口 135, TCP 标志位为 SYN, 报文长度为含包头的最小长度;

⑦ 137 端口扫描: TCP 协议、目的端口 137, TCP 标志位为 SYN, 报文长度为含包头的最小长度;

⑧ 138 端口扫描: TCP 协议、目的端口 138, TCP 标志位为 SYN, 报文长度为含包头的最小长度;

⑨ 445 端口扫描: TCP 协议、目的端口 445, TCP 标志位为 SYN, 报文长度为含包头的最小长度;

3.1.3 数据存储模块

通过数据筛选模块筛选后, 将有关 Netflow 数据写入数据库, 建立相关库表.

按照端口号或针对对象不同, 分别建立 Slammer 蠕虫数据表、22 端口扫描数据表、1433 端口扫描数据表等数据库表, 并进行数据库管理, 建立索引主键, 便于之后的分析查询.

3.1.4 分析模块

分析模块中主要将异常扫描数据进行数据归类, 并结合数学建模来得出相应的判别概率.

具体数学模型如下:

$$p=\frac{c_i}{c_{all}}w+\frac{n_i}{n_{all}}w$$

$$(c_{all}=c_1+c_2+...+c_n, n_{all}=n_1+n_2+...+n_n)$$

其中:

p: 为 src_ip_i 正在进行扫描的判别概率

c_i: 针对指定目的端口的 src_ip_i 的计数值

c_{all}: 针对指定目的端口的计数总和

n_i: 源 IP 地址均为 src_ip_i, 针对特定目的端口的目标 IP 总数

n_{all}: 全部针对特定目的端口的目标 IP 总数

w_i: 对扫描判别概率 p 的影响程度的权重

通过比对特定时间内扫描计数比率和目的地址分布情况进行数值计算, 根据实际情况赋予不同的权重, 得出相应的扫描判别概率.

4 实际应用

根据上述讨论研发网站信息分析系统, 并予以技术实现.

系统环境为: Intel(R) Xeon(R) CPU E5620 @

2.40GHz、48G 内存、3T SATA 硬盘、千兆以太网、操作系统采用 Linux 2.6.18-8.el5PAE 内核，脚本采用 Linux bash shell 编程语言，版本 3.1.17，数据库为 MYSQL (版本 5.0.22)，前台网页系统为 PHP (版本 5.1.6).

我们在骨干运营网络进行实际部署，从骨干路由器上生成 Netflow 数据，并发给系统服务器，系统服务器经过数据预处理、数据筛选、存储、分析等处理操作，进行前端展示.

从图 4 可知，某时刻 139 和 445 端口发生扫描异常的 IP 地址，该地址扫描计数和扫描判别概率值明显偏高，并在系统中高亮预警显示，通过用 Netflow 数据的具体数据整理验证，该地址在该时间段内确实有大量的目的地址段扫描情况发生(如图 5)，经过后续现场排查工作发现，该主机感染了僵尸木马程序，正在利用微软 ms06-040 高危漏洞进行扫描探测并渗透传播，该系统的运行效果良好.

图 4　扫描判别概率结果

5　结语

本文将数值分析的方法应用到异常扫描行为的监测中，大大提高了海量网管数据对异常网络行为的筛查和甄别效率，提升了互联网运营单位的服务质量和

ID	源IP	目的IP	协议	源端口	目的端口	字节	包	流	包大小
12991	159.[redacted].85	222.56.118.22	6	1581	445	4800	100	100	48
12993	159.[redacted].85	222.56.118.22	6	1992	445	8300	100	100	83
12996	159.[redacted].85	210.67.99.19	6	2293	445	4800	100	100	48
12999	159.[redacted].85	210.22.195.34	6	3633	445	8300	100	100	83
13000	159.[redacted].85	222.56.118.22	6	1749	445	38800	100	100	388
13001	159.[redacted].85	121.52.49.151	6	2037	445	4800	100	100	48
13003	159.[redacted].85	222.56.118.23	6	2128	139	4800	100	100	48
13004	159.[redacted].85	222.56.118.23	6	3816	445	4000	100	100	40
13005	159.[redacted].85	210.22.195.34	6	4661	445	21600	100	100	216
13006	159.[redacted].85	61.10.206.40	6	1937	445	4800	100	100	48
13007	159.[redacted].85	210.75.203.5	6	4014	445	4800	100	100	48
13008	159.[redacted].85	210.3.239.66	6	1232	445	8300	100	100	83
13009	159.[redacted].85	210.3.239.66	6	3551	139	4800	100	100	48
13010	159.[redacted].85	114.92.121.213	6	2536	445	4800	100	100	48
13012	159.[redacted].85	210.3.239.66	6	2759	139	4800	100	100	48
13013	159.[redacted].85	222.56.118.23	6	3306	445	4800	100	100	48
13014	159.[redacted].85	222.56.118.23	6	1167	139	4000	100	100	40
13015	159.[redacted].85	114.92.37.239	6	2453	445	4800	100	100	48
13016	159.[redacted].85	210.3.239.66	6	4077	445	8300	100	100	83
13019	159.[redacted].85	210.78.57.127	6	2795	445	4800	100	100	48
13020	159.[redacted].85	114.92.147.7	6	4599	445	4800	100	100	48
13021	159.[redacted].85	210.22.195.34	6	3759	445	4000	100	100	40
13024	159.[redacted].85	210.75.25.5	6	2603	445	4800	100	100	48
13026	159.[redacted].85	114.92.209.149	6	3371	445	4800	100	100	48
13027	159.[redacted].85	24.76.10.198	6	1801	445	2160	10	10	216
13028	159.[redacted].85	75.236.25.237	6	2011	445	480	10	10	48
13029	159.[redacted].85	20.33.98.245	6	1753	445	480	10	10	48
13030	159.[redacted].85	215.98.183.109	6	1769	445	480	10	10	48

图 5　疑似扫描地址通信地址对连接情况

支撑保障能力，但系统目前采用的数值分析函数仅为线性函数，且对权重的考量还需进行深入的研究，仍需在未来研究工作中进行完善和充实.

参考文献

1 CNNIC.第 31 次“中国互联网络发展状况统计报告”.http://www.cnnic.cn,2013,5.

2 CNCERT.“2012 年我国互联网网络安全态势综述”.http://www.cert.org.cn,2013:2–9.

3 李军,曹文君,李扬.FB-NBAS:一种基于流的网络行为分析模型.计算机工程,2008,34(3):165–167.

4 牛国林,管晓宏,龙毅,秦涛.多源流量特征分析方法及其在异常检测中的应用.解放军理工大学学报(自然科学版),2009,10(4):350–355.

5 杨奇.基于异常行为特征的僵尸网络检测方法研究[学位论文].西安:陕西师范大学,2010.

(上接第 18 页)

7 Cliff C. Correct Methods For Adding Delays To Verilog Behavioral Models. International HDL Conference 1999 Proceedings. 1999. 23–29.

8 Navabi Z. Verilog Digital System Design: Register Transfer Level Synthesis, Testbench, and Verification. 2nd ed., USA: McGRAW-Hill Companies, 2007: 83–85.

9 Samir P. Verilog HDL: A Guide To Digital Design and Synthesis. USA: PEARSON Education. 2003. 82–85.

10 刘波.精通 Verilog HDL 语言编程.北京:电子工业出版社,2007.247–250.