通过流量和数据包综合估计内网感染蠕虫概率的研究 *

王勇超[a]，谢永凯[a]，朱之平[a]，董亚波[b]
（浙江大学 a. 网络信息中心；b. 人工智能所，杭州 310027）

摘 要：提出了一种分析内网感染蠕虫可能性大小的方法。对通过内网交换机上的数据包使用蠕虫行为进行分析，得到行为异常的数据包数量，然后使用 AR 模型分析异常数据包的数量得到异常数据包的增长率；对内网异常流量和异常数据包增长率加权，并对它们综合估计得到内网中感染蠕虫概率的大小。实验表明该方法有效可行。

关键词：异常数据包；AR 模型；异常流量；加权；概率
中图分类号：TP393 **文献标志码**：A **文章编号**：1001-3695(2010)01-0237-03
doi:10.3969/j.issn.1001-3695.2010.01.070

Research of Internet worm infection probability estimation based on comprehensive analysis of net-flow and packets

WANG Yong-chao[a], XIE Yong-kai[a], ZHU Zhi-ping[a], DONG Ya-bo[b]
(*a. Center of Network & Information, b. Institute of Artificial Intelligence, Zhejiang University, Hangzhou* 310027, *China*)

Abstract: This paper presented a method of analyzing the probability of Internet worm infection in internal network. Analyzed the packets through the switch in network by using the active of Internet worm, acquired the amount of abnormal packets, then analyzed the number of abnormal packets to get the abnormal packets rise rate by using AR model. Weighted the abnormal flow and abnormal packets rise rate in internal network, comprehensively estimated them and got the probability of Internet worm infection in internal network. The experiment results show that the method is effective and feasible.

Key words: abnormal packets; AR model; abnormal flow; weight; probability

0 引言

随着计算机技术的发展和互联网的扩大，给整个社会带来了日新月异的变化，促进了整个社会的进化。然而，随着网络的普及和规模的不断扩大，蠕虫作为当前恶意代码中广为传播的一类，构成了目前对网络的最大威胁。蠕虫的泛滥及其传播手段的改进，留给对蠕虫的预警时间越来越短，因此，构架快速的蠕虫检测平台已经是当务之急，这对于减少蠕虫对网络的破坏并将其最小化，有着很大的现实意义。

蠕虫的异常检测是对收集到的数据进行分析，在此基础上判断蠕虫的感染情况。它认为具有攻击的行为与正常行为不同，从而发现异常行为时怀疑存在攻击。针对未知的蠕虫，可以利用异常发现来发现蠕虫导致的网络异常。目前常用的异常检测方法包括统计、抽样、神经网络以及其他一些利用蠕虫特性的方法[1]。

通过对异常行为的数据包进行筛选，检测出异常数据包产生的个数，通过使用 AR 模型对其进行判断从而得到异常数据包的增长率。由于异常数据包的增长率判断基于异常行为的筛选，在筛选的同时进行异常流量的检测，通过对异常流量和异常数据包同时进行分析，估计判断当前网络内感染蠕虫的可能性。该方法是基于统计的分析方法，所以对历史数据的依赖性较大。

1 介绍

蠕虫可以传播自身的副本，并且能够在远端机器上执行代码，是一种智能化、自动化的攻击载体。它会扫描和探测网络上存在服务漏洞的节点主机，一旦渗透成功会自我复制许多副本，通过网络传播从一个节点到另外一个节点。通过对多种网络蠕虫的分析，可以确定网络蠕虫的基本工作机制分为五步，包括收集信息、探测目标主机、攻击目标系统、自我复制、后续处理[2]。而蠕虫的一次攻击过程可以划分为三个阶段，即获取目标、提升特权和感染阶段。为了有效地检测并控制蠕虫的大规模爆发，应在蠕虫攻击的第一阶段分析出感染状况并在网络内作出预警。

蠕虫爆发时一般会呈现如下规律：

a)宿主机出现大量的连接数。受感染主机会在一定的时间内试图连接大量的不同 IP，扫描它们是否具有漏洞。在相同时间段内，感染网络蠕虫的主机所产生的不同连接数目远大

收稿日期：2009-05-06；**修回日期**：2009-06-18 **基金项目**：国家自然科学基金资助项目(60503061)；国家“863”计划资助项目(2008AA01Z416)

作者简介：王勇超(1975-)，男，讲师，硕士，主要研究方向为网络安全；谢永凯(1984-)，硕士，主要研究方向为网络安全(tsewingkai@gmail.com)；朱之平，副研究员，主要研究方向为网络安全；董亚波，副教授，博士，主要研究方向为无线网络、网络安全等.

于正常主机的不同连接数目。

b)宿主机发送的数据包具有相似的行为。由于蠕虫发送数据包扫描端口的速度大大高于普通用户正常操作的发包速度,当网络中出现蠕虫的宿主机时,其发送数据包的目的端口相对其他主机具有高度的相似性和规律性。

2　流量与数据包的采集分析

本文通过统计分析的方法对经验流量建立动态临界线,以此分析出异常流量产生的情况。产生异常流量时通过使用 Netflow 对产生异常行为的节点发送的数据包进行统计,从而获得异常数据包的情况。

2.1　异常流量分析检测

在某一时刻 t,本文假设当前网络正常的最大流量为 $F_{\max}$,则正常情况下超过最大流量的概率分布为 $P(\xi_t \geqslant F_{\max})$,其中 ξ_t 为当前时刻 t 的流量。对于连续性的概率分布,一个变量如果受到大量微小独立的随机因素影响,那么这个变量一般满足正态分布 $N(\mu,\sigma^2)$。由于网络内存在大量独立用户,其行为相互独立且不同,本文也选择正态分布来描述单个时刻内网中发生流量的正常行为。

对于流量 ξ_t,过去 m 次正常流量的观测值为 $F_i(i=1,2,\cdots,m)$,取其平均值 $\overline{X}$ 作为子样观测的均值,有 $\overline{X}=\sum_{i=1}^{m}F_i^t/m$,并令其方差为子样观测方差,有 $D\xi_t=\sum_{i=1}^{m}(F_i^t-\overline{X})^2/N$,而对于符合正态分布的普通网络行为,$\mu_1=\overline{X}$,$\sigma_1=\sqrt{D\xi_t}=\sqrt{\sum_{i=1}^{m}(F_i^t-\overline{X})^2/N}$。

取动态基线为 μ_1^t,可信范围为 $3\sigma_1^t$。如果 $\mu_1^t-3\sigma_1^t>0$,则动态临界线的范围为 $[\mu_1^t-3\sigma_1^t,\mu_1^t+3\sigma_1^t]$,否则为 $[0,\mu_1^t+3\sigma_1^t]$。

假定当前 t 时刻的实际观测值为 F_D^t,对以往统计出的单个时刻主交换机上节点的连接数 $b_i^t(i=1,2,\cdots,m)$ 使用统计均值,得到 $\overline{b_t}=\sum_{i=1}^{m}b_i^t/m$,令 $\overline{l_t}=\mu_1^t/\overline{b_t}$ 为平均流量长度。由于网络中蠕虫传播的前期宿主机总是发送探测包用于寻找能够攻击的其他主机,此时网络内的流量会比正常流量时有所增加,即 $F_D^t>\overline{l_t}\times b_D^t$。其中 b_D^t 为当前网络中节点与交换机的连接数。

取 $\zeta=F_D^t-(\mu_1^t+3\sigma_1^t)$,若 $\zeta>0$,则可以判断当前网络内出现异常流量。当 $\zeta\leqslant 0$ 时,当前的连接数为 b_D^t,则一般可信的流量应为 $F_N^t=\overline{l_t}\times b_D^t$,令 $F_U^t=|F_D^t-F_N^t|$;同样,由于当前时刻网络内正常行为满足正态分布,在这种情况下置信范围仍然取 $3\sigma_1^t$,即当 $F_U^t>3\sigma_1^t$ 时,判断此时出现的流量为异常流量。

2.2　异常数据包筛选

蠕虫利用计算机操作系统存在的已知漏洞(对应着一个端口号)进行工作,并传染给存在该漏洞的计算机。尽管为保护自己不被发现,蠕虫往往会采用伪造自己源地址和源端口的策略,但却不能伪造目的端口号。在交换机上使用 Netflow 可以监控到数据交换的情况,分析数据包时一般应该注意如下的特点:

a)被感染的主机会持续发送大量的数据包;

b)同蠕虫的主机会发送相同大小的数据包;

c)蠕虫会发送数据包到相同的端口。

所以,当 Netflow 监测到网络内有主机长时间发送大量长度相同的数据包到相同的端口时,就可以怀疑为异常数据包并对其进行记录。其筛选方式如下:

首先在时间段 Δt 内检测并统计网络内发送的数据包长度,使用 c_{ij} 表示具有相同长度的数据包个数。其中:i 表示对应的网络内主机的编号;j 表示统计的时间段个数,j 对于每个节点都相同。对于同一个 i,如果存在 c_{ij} 在多个 j 内具有相同的大小,则统计其数据包所发送的端口。用 $p_{i\Delta}$ 表示节点 i 最频繁地针对某个目的端口发送数据包的个数(其中 Δ 表示目的端口),如果对于同一个 i 和 Δ 统计出的 $p_{i\Delta}$,则将此时的 $p_{i\Delta}$、i 和 Δ 写入数据列表,表示 i 对于端口 Δ 产生了 $p_{i\Delta}$ 个可疑数据包。最后,在时间段内对于多个 i 将其与 $p_{i\Delta}$ 作和,其 $k_t=\sum_t p_{i\Delta}$ 值即表示为在 t 时刻检测到 Δt 时间段内异常数据包的个数。

2.3　蠕虫检测的 AR 模型介绍

在蠕虫感染检测方面,文献[3]提出了基于网络流量趋势的 AR 模型估计,该网络蠕虫的传播过程在时间上描述如下:

$$z_t=z_{t-1}+\Delta t\alpha_t Z_{t-1}+e_t=(1+\Delta t\alpha_t)Z_{t-1}+e_t$$

其中:$t=1,2,\cdots,n$ 表示采样时刻;Z_t 表示扫描监测点在 t 时刻获取的网络扫描包数量;Δt 为时间间隔;α_t 为蠕虫感染率;e_t 为观测误差,此误差与网络正常访问量有关。将各时刻观测到的扫描包数量 Z_t 作为观测值输入量,就可以对感染率 α_t 进行估计。

利用 AR 模型进行蠕虫感染概率分析需要基于以下几个前提条件:

a)需要确定网络中数据包是否为扫描端口的扫描包。

b)根据 AR 模型所描述的情况,如果拟合的是蠕虫的感染概率,则当前网络中异常的扫描包主要是由于蠕虫发送大量探测包所引起的异常流量,而不是由于其他原因如普通用户自行发起连接所引起的异常,即所拟合的数据均需无异常误差干扰的。

c)网络中所有监测点的扫描数据应能客观反映空间分布方面的感染情况,而不能受某种异常信息污染。

d)数据样本应足够大。

2.4　异常数据包增长率分析

本文借鉴蠕虫检测的 AR 模型,分析当异常数据包产生时其增长率的大小。通过对异常数据包的分析可以获取当前异常数据包产生的概率。其步骤如下:

首先对异常的数据包套用 AR 模型,产生递归式:

$$k_t=k_{t-1}+\Delta t\beta_t k_{t-1}+e_t=(1+\Delta t\beta_t)k_{t-1}+e_t$$

其中:本文依旧使用 e_k 作为网络环境中带有噪声的观测误差,而 β_t 为此刻网络节点发送异常数据包的增长率。本文使用最小二乘估计法对采样的数据进行逼近,以求得 β_t。具体方法如下:

令 $s_t=\Delta t\times k_{t-1}$,$d_t=k_t-k_{t-1}$,则 d_t 的残差为

$$c_t=s_t\times\tilde{\beta}_t-d_t$$

将统计到的采样数据代入,并将其写成向量形式:

$$C_t=S_t\times\tilde{\beta}_t-D_t$$

其中:C_t、S_t、D_t 分别为由不同采样时间的 c_t、s_t、d_t 所组成的向量。根据最小二乘估计方法,令其约束条件 $\sum_t(c_t^2)$ 达到最小值,则会得到

$$\tilde{\beta}_t = (S'_t \times S_t)^{-1} S'_t D_t$$

根据数值分析的经验可以知道,当时间采样 t 足够大,即通过足够时间的观测估计后,$\tilde{\beta}_t$ 的值最终会收敛于 β_t,即为异常数据包的发生概率。

通过对浙江大学某宿舍楼服务器进行监控,时长为5 h,每5 min 生成一次报告。采取限制行为的措施,即采用了异常行为限制后异常数据包的增长率,如图 1 所示。

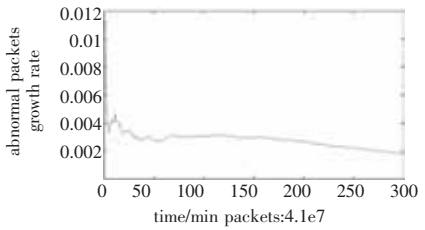


图 1　异常数据包增长率

如图 1 所示,异常数据包增长率是不断变化的。由于服务器中采用了异常主机限制策略,异常数据包增长率随时间增长而降低,而增长率可忍受阈值设置为 0.001 8,在采样后期异常包增长率接近于阈值,降低速度变缓。

3　内网感染蠕虫概率的判断

本文使用回归模型分析异常数据包,从而得出产生异常数据包的概率。对异常数据包产生概率以及异常流量的偏移值加权,并对它们的数据进行线性相关的分析,就可以进一步接近其真实的感染概率。

3.1　蠕虫感染可能性分析

网络内如果有蠕虫感染时会造成流量较普通时候偏大,但有较大流量的时候不能确定是否由蠕虫造成,甚至有时会出现正常流量时仍有蠕虫感染的情况。通过对异常流量的获取以及对可疑数据包产生概率的分析,可以减少对异常流量的误判。结合异常流量的大小、异常数据包出现的概率以及蠕虫的行为可知:当出现异常数据包的概率偏大的同时出现与其值相关的大量异常流量时,网络中存在蠕虫感染的行为概率较高。

此时,设置异常流量与异常数据包概率的加权值分别为 p_1 和 p_2,则蠕虫感染概率可以近似表示为

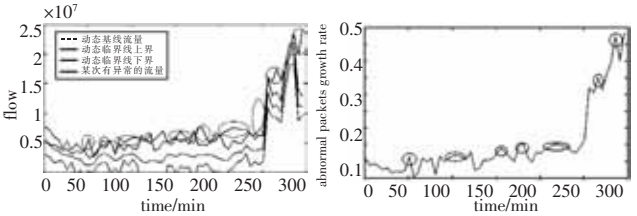
$$\alpha_t = p_1 (F^t_U / F^t_N) + p_2 \beta_t$$

其中:权值的选取要根据具体的网络环境,其限制条件为 $p_1 + p_2 = 1$。例如,在网络比较宽松不拥挤的情况下可以对 p_1 设置较小值;相反,当网络带宽资源比较紧张时可以设置其为较大值,要根据网络管理人员的经验进行设置。

最后,对感染蠕虫概率的阈值进行设置,如果超出阈值时发出危险报警。

通过对浙江大学宿舍楼服务器建立异常数据包和流量综合分析的方法,得到的结果如图 2 所示。交换机带宽为百兆,检测时长为 300 min,每 5 min 进行一次检测报告。感染率阈值设置为:当网络内感染的可能性大于等于 0.19 时产生警报。由于带宽比较大,将权值设置为 $p_1 = 0.25, p_2 = 0.75$,即对异常数据包的分析较流量分析更加重要。检测时间为某天13:00 ~ 18:00。

在图 2(b)中,在综合异常流量和异常数据包增长率分析后,圆圈标注出感染概率超过阈值发出报警的时段,在它们的时间段中,计算所得感染蠕虫的概率均超过阈值。



(a) 异常流量的检测　　(b) 产生的警报图

图 2　分析效果图

3.2　蠕虫检测系统结构

本文根据异常主机的行为描述了一个简单的蠕虫检测系统,该系统由数据处理器(DP)、数据包和流量感应器、网络事件处理中心三部分组成。首先由感应器感知网络内的整体流量和客户端流量、客户端和服务器的数据包数量,DP 将收集到的数据结合历史同期数据进行处理,并把结果发送至网络事件处理器,网络事件处理中心的宿主机列表接收并将数据推入历史统计表中更新,同时事件处理中心对检测到的蠕虫主机发出警告并对其进行隔离。检测系统结构框图如图 3 所示。

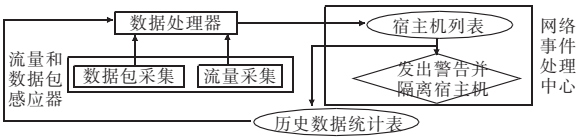


图 3　检测系统结构框图

4　结束语

根据近年来网络的发展,网络蠕虫的产生以及变种速度大大加快,使得网络的防御体系不能及时对网络进行保护。本文从异常流量和异常数据包判断这两者同时出发,通过对两者加权来判断当前内网中感染蠕虫的概率大小。其优点在于,判断方法简单易用,而且综合流量和异常数据包行为来进行判断可以减少误报。但是由于分析之前需要进行大量的数据采集,会涉及到大量的历史数据,这使得该方法依赖于长时间的检测,而检测历史越长,其检测的准确度也越大。

提高异常数据包的筛选算法可以大大提高本文所提出方法的速度,从而增加系统报警的实时性,实现有效的蠕虫感染早期预警。

参考文献:

[1] 郭晔. 面向 agent 的蠕虫防御系统研究[D]. 杭州:浙江大学,2007.

[2] 汪伟. 网络蠕虫检测技术研究与实现[D]. 杭州:浙江大学计算机科学与技术学院,2006.

[3] ZOU C C, GONG Wei-bo, TOWSLEY D, *et al*. The monitoring and early detection of Internet worms[J]. IEEE/ACM Trans on Networking, 2005, 13(5): 961-974.

[4] 魏宗舒. 概率论与数理统计[M]. 北京:高等教育出版社,1983: 394-399.

[5] 张　嵩. AR 模型在预测中的分析[J]. 襄樊学院学报,2008,29(5):13-14.

[6] 汪伟,鲁东明,董亚波,等. 面向内网的网络蠕虫检测系统设计与实现[J]. 计算机工程,2006,32(17):205-206.

[7] ZOU C C, GONG Wei-bo, TOWSLEY D. Code red worm propagation modeling and analysis[C]//Proc of ACM Conference on Computer and Communications Security. 2002: 138-147.