Journal of Computer Applications
计算机应用,2015,35(7):1892 - 1896
ISSN 1001-9081
CODEN JYIIDU
2015-07-10
http://www.joca.cn

文章编号:1001-9081(2015)07-1892-05
doi:10.11772/j.issn.1001-9081.2015.07.1892

基于统计特征的隐匿 P2P 主机实时检测系统

田朔玮*,杨岳湘,何　杰,王晓磊,江志雄
(国防科学技术大学 计算机学院,长沙 410073)
(*通信作者电子邮箱 tsw777@sohu.com)

摘　要:针对当前隐匿恶意程序多转为使用分布式架构来应对检测和反制的问题,为快速精确地检测出处于隐匿阶段的对等网络(P2P)僵尸主机,最大限度地降低其危害,提出了一种基于统计特征的隐匿 P2P 主机实时检测系统。首先,基于 3 个 P2P 主机统计特征采用机器学习方法检测出监控网络内的所有 P2P 主机;然后,再基于两个 P2P 僵尸主机统计特征,进一步检测出 P2P 僵尸主机。实验结果证明,所提系统能在 5 min 内检测出监控网内所有隐匿的 P2P 僵尸主机,准确率高达到 99.7%,而误报率仅为 0.3%。相比现有检测方法,所提系统检测所需统计特征少,且时间窗口较小,具备实时检测的能力。

关键词:对等网络;僵尸网络;统计特征;机器学习;检测系统
中图分类号:TP393.06　　**文献标志码**:A

Real-time detection system for stealthy P2P hosts based on statistical features

TIAN Shuowei*, YANG Yuexiang, HE Jie, WANG Xiaolei, JIANG Zhixiong
(*College of Computer, National University of Defense Technology, Changsha Hunan* 410073, *China*)

Abstract: Since most malwares are designed using decentralized architecture to resist detection and countering, in order to fast and accurately detect Peer-to-Peer (P2P) bots at the stealthy stage and minimize their destructiveness, a real-time detection system for stealthy P2P bots based on statistical features was proposed. Firstly, all the P2P hosts inside a monitored network were detected using means of machine learning algorithm based on three P2P statistical features. Secondly, P2P bots were discriminated based on two P2P bots statistical features. The experimental results show that the proposed system is able to detect stealthy P2P bots with an accuracy of 99.7% and a false alarm rate below 0.3% within 5 minutes. Compared to the existing detection methods, this system requires less statistical characteristics and smaller time window, and has the ability of real-time detection.

Key words: Peer-to-Peer (P2P); botnet; statistical feature; machine learning; detection system

0　引言

随着各类数字终端、服务器资源、网络带宽等资源的持续增长,对等网络(Peer-to-Peer, P2P)应用突破原有技术瓶颈有了长足的发展。由于 P2P 架构所具有的高效传输协议和良好鲁棒性,其技术已经被广泛地运用到文件共享、流媒体传播、视音频通话等众多领域。与此同时,一些受限于传统集中式架构中单点失效等问题的隐匿恶意程序,也转为采用 P2P 架构,这使得它们能够更好地应对当前的检测及反制措施。P2P 型僵尸网络是目前 P2P 隐匿恶意程序的典型代表。P2P 型僵尸网络主要由僵尸主机(Bot)组成,攻击者(Botmaster)通过特有的 P2P 协议发布命令与控制信息,远程控制所属僵尸主机协同发起诸如分布式拒绝服务攻击(Distributed Denial of Service, DDoS)、制造垃圾邮件、窃取敏感信息等多种网络攻击。

本文提出了一个基于流量统计特征来检测监控网络内隐匿 P2P 僵尸主机的实时检测系统。该系统通过在网络出口处捕获到的网络流量记录,首先通过 3 个流量统计特征并借助机器学习方法检测出所有进行 P2P 通信的主机;然后根据 P2P 僵尸主机特有的网络行为模式,提出两个流量统计特征,进一步检测出被 P2P 僵尸程序感染的 P2P 僵尸主机。与现有检测系统相比,本文系统仅关注网络流量的统计特性,无需检查报文载荷内容,因此不受报文负载加密及混淆的影响;其次,本系统能在 P2P 僵尸主机处于隐匿阶段(发起攻击前)时就将其精确检出;第三,本系统检测仅需 5 个统计特征,计算开销较小,具有较强的实时检测能力。

1　相关工作

目前,针对僵尸网络的检测方法主要分为基于主机和基于网络两类。基于主机的检测方法分析每台受控主机的主机行为,记录所有出现的可疑行为,比如可疑系统 API(Application Programming Interface)的自动调用、非法访问系统文件等[1-4]。这种方法存在一个不可避免的缺陷,即缺乏可扩展性,因为它必须在每台受监测主机上都安装一个监控工具。基于网络的检测方法主要致力于挖掘网络流量的异常特征来判断主机是否感染僵尸程序。BotHunter[5] 和

收稿日期:2015-02-04;**修回日期**:2015-03-27。　**基金项目**:国家自然科学基金资助项目(61170286)。
作者简介:田朔玮(1984 -),男,内蒙古呼和浩特人,硕士研究生,主要研究方向:计算机网络与安全;　杨岳湘(1965 -),男,湖南岳阳人,研究员,博士,主要研究方向:计算机网络与安全、数据挖掘、信息检索;　何杰(1984 -),男,四川甘洛人,博士研究生,主要研究方向:计算机网络与安全;　王晓磊(1990 -),男,河南许昌人,硕士研究生,主要研究方向:计算机网络与安全;　江志雄(1985 -),男,云南昆明人,硕士研究生,主要研究方向:计算机网络安全。

BotMiner[6]通过检测扫描、发送垃圾邮件、传播恶意代码、DDoS 攻击等恶意活动来检测僵尸主机,但当 P2P 僵尸主机处于隐匿阶段未发起恶意攻击,或 P2P 僵尸网络的恶意攻击变得更加隐匿而难以检测时,此类方法的精度将大大降低。BotGrep[7]和 BotTrack[8]通过分析从多个大型网络(如互联网服务提供商(Internet Service Provider, ISP)网络)中提取出的网络流散布图(Traffic Dispersion Graph)来检测 P2P 僵尸网络。尽管这些方法不再依赖于僵尸网络的恶意活动,但其需要从诸如蜜罐等的第三方系统中收集到的被感染主机信息来引导检测过程,此外还需要大型网络的流量采集权限,而这通常难以获取。另一些研究[9-11]认为同属于一个 P2P 僵尸网络的僵尸主机在网络行为特性上会表现出一定的相似性。此类方法通过聚类表现出相似网络行为的主机来检测 P2P 僵尸主机。但是,此类方法仅当监控网络内存在多个同属于一个僵尸网络的僵尸主机时才能实施检测,当目标网络内仅存在一个同类僵尸主机时,此类方法即失效。此外,Rahbarinia 等[12]和 Zhao 等[13]等同样基于流量统计特征,并借助机器学习算法来检测 P2P 僵尸主机。但是,他们所提出的统计特征数量较多,特征向量维度较大,因此检测系统所需计算和存储开销较大。

本文首先在分析 P2P 网络工作原理的基础上,提出了检测 P2P 主机的 3 个流量统计特征,即:节点链接不稳定性、节点发现模式和远端 IP 分布广度,然后根据 P2P 僵尸主机特有的网络行为模式,提出了区分 P2P 僵尸主机和正常 P2P 主机的两个统计特征,即:通信报文占比和远端 IP 地址回访数。最后借助机器学习方法实现了对 P2P 僵尸主机的实时检测。实验结果证明,本系统能在 5min 内检测出监控网内所有隐匿的 P2P 僵尸主机,准确率高达到 99.7%,而误报率仅为 0.3%。

2　系统设计

本文系统主要分为 2 个层次架构:训练阶段和检测阶段。其中包含 3 个模块:统计特征提取模块、机器学习训练模块和实时检测模块。系统结构如图 1 所示。

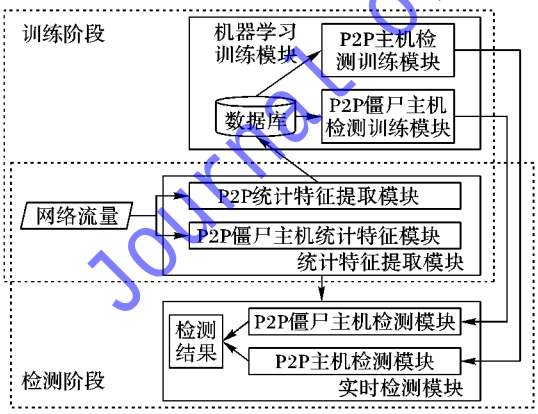

图 1　系统结构

训练阶段是检测阶段的基础,为检测阶段提供决策函数。检测阶段使用训练阶段提供的决策函数完成实时检测功能。

统计特征提取模块以 IP 地址标识的被监测节点为单位,对网络流量进行解析和计算,在每个时间窗口提取出全部统计特征,分别组成特征向量。

机器学习训练模块根据训练特征向量集训练机器学习算法,生成决策函数。

实时检测模块从特征提取模块获得实时特征向量进行检测并最终生成和输出检测结果。

2.1　统计特征提取模块

本模块负责从网络流量中提取出所有统计特征,包括用于检测 P2P 主机的统计特征和用于检测 P2P 僵尸主机的统计特征。

2.1.1　P2P 统计特征提取模块

通过分析 P2P 网络的固有特性,本文提出如下 3 个统计特征来区分 P2P 主机和非 P2P 主机。

1)节点链接不稳定性。P2P 节点为了获取所需要的资源必须主动向其他节点发起链接请求,而相当一部分的对等节点会因为线路故障、节点离开、协议版本不统一等原因而无法建立正常链接,因此,P2P 主机所产生的失败的网络链接数往往远高于非 P2P 主机,本文将该特性定义为“链接不稳定性”,并用链接失败率 *FailRate* 来表示。

特征定义 1　$FailRate = Flow_{Failed}/Flow_{All}$。其中:$Flow_{All}$ 表示主机 H 在时间窗口内产生的所有流的数量,$Flow_{Failed}$ 表示其中失败的流的数量。*FailRate* 的值越大,意味着节点 H 的链接不稳定性越高,则该主机越可能是 P2P 主机。

根据 NetFlow[14]的定义,本文通过五元组〈Proto, IP_{src}, $Port_{src}$, IP_{dst}, $Port_{dst}$〉确定流,即传输层协议、源 IP 地址及端口、目的 IP 地址及端口。本文主要关注传输控制协议(Transmission Control Protocol, TCP)流和用户数据报协议(User Datagram Protocol, UDP)流,并认为一条成功的 TCP 流必须包含完整的三次握手过程,而一条成功的 UDP 流至少要有一个请求报文和随之而来的响应报文;反之则记为失败的流。

当在满足下列条件之一时,本文认为一条流结束:

a)该流已有 10 min 未收到新的报文;

b)当流的活动时间超过 30 min 时,本文将其强行掐断;

c)检测到表示 TCP 流终止的标志位。

2)节点发现模式。P2P 节点在启动时大多直接根据内置 IP 列表建立链接,并通过相互学习路由表方式来扩大链接范围,而不需要经过 DNS 查询。但大多数的非 P2P 应用在发起链接之前,往往需要经过 DNS 域名解析以获取远端 IP 地址。本文用 $IP_{DNS}Rate$ 来表示 P2P 主机的此特性,并定义如下。

特征定义 2　$IP_{DNS}Rate = IP_{DNS}/IP_{ALL}$。其中:$IP_{ALL}$ 表示主机 H 在时间窗口内链接的所有远端 IP 地址的数量,IP_{DNS} 表示其中经过 DNS 域名解析后获得的 IP 地址数量。若 $IP_{DNS}Rate$ 的值越小,其经过 DNS 域名解析获得的远端 IP 地址越少,则其 P2P 主机特性越明显。

3)远端 IP 分布广度。P2P 节点为了保持其连通性和最大限度地获取资源,往往会尽可能地同多个对等节点建立链接。这些对等节点在分布上往往会分散到许多不同的子网之中。P2P 节点的这一特性与非 P2P 节点链接少且集中与少数服务器通信有很大的不同。根据此特性,本文定义远端 IP 分布广度 *Diversity*,如下所示。

特征定义 3　远端 IP 分布广度 *Diversity* 为主机 H 在单位时间窗口内所链接的远端 IP 地址分布的子网数量。

若 *Diversity* 的值越大,其链接的远端 IP 分布越广,则其 P2P 流量特征越明显。具体而言,本文取 IP 地址前 16 位前缀来近似表示其所属网络号[12],即若两个远端 IP 地址前 16 位不同,则本文认为它们隶属于不同的子网。

以上所列 3 个统计特征都由 P2P 网络的内在特性所决

定,不易被篡改,同时也不会受到加密流量的影响。

2.1.2 P2P 僵尸主机统计特征提取模块

虽然正常的 P2P 程序和隐匿的 P2P 僵尸程序都采用 P2P 模式架构,二者之间存在共同的网络行为模式,但由于使用协议、功能目的以及应用背景的不同,二者之间还存在着明显区别。通过对 P2P 僵尸主机特有的网络行为模式的分析,本文提出以下两个统计特征来区分 P2P 僵尸主机和正常 P2P 主机。

1)通信报文占比。本文注意到 P2P 僵尸主机为了保持自身的隐蔽性,大多数时间都处于潜伏阶段,在这一阶段,P2P 僵尸主机对外并不执行恶意行为,只是频繁与对等僵尸节点进行通信报文交互,如宣告自身存在、查询网络拓扑、寻求最新指令等,这些通信报文通常较小。而正常 P2P 应用除了这些通信报文,还存在大量用于数据传输的数据报文,为了达到较高的传输率,这些数据报文通常较大。

特征定义 4 根据实验观察,本文将报文大小小于 100 字节的报文看作通信报文[15],并统计单位时间窗口内通信报文占总报文数量的比例,称之为通信报文占比,用 $CPRate$ 来表示,即:$CPRate = Packet_{CP}/Packet_{All}$。其中:$Packet_{CP}$ 表示主机 H 在时间窗口内产生的通信报文的数量,$Packet_{All}$ 表示所有报文数量。若 $CPRate$ 的值越高,则其为僵尸主机的特性越明显。

2)远端 IP 回访数。P2P 网络中存在大量不同类型的网络活动,如节点发现、请求内容、资源通告等。同一类网络活动所产生的流往往具有相似的报文数量和大小[16]。此外,本文注意到 P2P 僵尸程序的同类网络活动所产生的流比正常的 P2P 程序更喜欢反复访问相同的远端 IP 主机。出现此现象的原因可能有以下两点:一是 P2P 僵尸网络的抖动性(节点加入或离开 P2P 网络的动态性)要比正常 P2P 网络要小得多[8]。P2P 僵尸主机为了从攻击者那里获取最新命令必须同其他对等主机保持稳定的连通性,而正常 P2P 主机为了获得稳定的下载速率,倾向于同更多新的对等主机建立链接,以获取资源。二是 P2P 僵尸主机为了规避检测,往往仅同有限的对等主机建立链接。

为利用此特性,本文将单位时间窗口内结束的所有流按流内报文平均大小进行分组。首先给定一系列连续区间 $[l_i, l_{i+1}]$,区间长度为 L,即 $l_{i+1} - l_i = L$,若流 f 的报文平均大小 b 满足 $0 \leqslant b - l_i < L$,则将该流 f 纳入此区间。最终每个区间内的流即为一组相似流,并认为它们由同一类 P2P 网络活动所产生。在此基础上,本文定义统计特征如下。

特征定义 5 主机 H 在单位时间窗口内结束的流记为集合 F,将 F 按报文平均大小分为一系列区间大小为 L 的相似流组 $G_1, G_2, \cdots, G_n$,将相似流组 G_i 中远端 IP 的集合记为 $\{IP_1, IP_2, \cdots, IP_q\}$,则该组相似流的远端 IP 重复访问次数为 $R_i = \sum_{x=1}^{q} (count(IP_x) - 1)$,定义远端 IP 最大回访数(Recall Max, RCM)为所有相似流组的 IP 重复访问次数 R_i 的最大值,即 $RCM = \max\{R_i\}_{1 \leqslant i \leqslant n}$。若 RCM 值越大,则其为僵尸主机的特性越明显。

上述两个检测 P2P 僵尸主机的特征都是由 P2P 僵尸网络的内在本质特性所决定的,提取容易,计算开销小,且不受加密流量的干扰。

2.2 机器学习训练模块

机器学习训练模块依托现有软件环境 Weka(Waikato Environment for Knowledge Analysis)为平台[17],使用从统计特征提取模块获得的特征向量集训练相关的机器学习算法,为实时检测模块提供决策函数,该模块包含 3 个子模块:数据库模块、P2P 主机检测训练模块和 P2P 僵尸主机检测训练模块。

数据库模块是其他两个子模块的基础,用于存储另外两个子模块训练机器学习算法时所用的特征向量集。特征向量集由统计特征提取模块从事先采集好的训练数据集中提取所得。P2P 主机检测训练模块和 P2P 僵尸主机检测训练模块分别调用数据库中相应属性的数据,训练所选机器学习算法,生成 2 个决策函数,再将 2 个决策函数分别提交给 P2P 主机检测模块和 P2P 僵尸主机检测模块。

2.3 实时检测模块

实时检测模块是负责实现实时检测功能的关键模块。该模块包括 2 个子模块:P2P 主机检测模块和 P2P 僵尸主机检测模块。它们分别将统计特征提取模块从实时网络流量提取的特征向量输入机器学习训练模块得到的对应的决策函数,得到检测结果。检测过程分为 2 步。

1)使用 P2P 主机检测模块检测出监控网络内所有的 P2P 主机剔除非 P2P 主机。

2)使用 P2P 僵尸主机检测模块从已检测出的 P2P 主机中检测出 P2P 僵尸主机。

3 实验及结果分析

3.1 实验数据集采集

为测试本文所提出系统的检测性能,本文采集了 3 个实验数据集,即:包含各种非 P2P 应用流量的数据集、包含各种常见 P2P 应用流量的数据集和一个包含两个著名 P2P 僵尸程序流量的数据集。

非 P2P 流量数据集。为采集非 P2P 流量数据集,本文首先选取校园网环境下一个子网内部的 35 台独立主机分别随机地运行各类常见的非 P2P 应用,如:网页浏览、收发邮件、FTP 下载、网络游戏等,然后在该子网的网关路由器上捕获并存储该子网一天内产生的所有网络报文。整个过程由人工交互进行控制,以最大限度地保证实验数据集的代表性。

P2P 流量数据集。为增加实验所涉及 P2P 主机的多样性,本文选取了 6 个目前最流行的 P2P 程序,即:迅雷、eMule、Bitcomet、BitTorrent、PPTV 和 CBOX,其中不仅包括了 4 个最典型的 P2P 文件共享系统,还包括了两个应用广泛的 P2P-TV 平台。本文在校园网内部署了一个包含 12 台主机的实验子网。然后在每两台实验主机上同时分别运行一个上述的 P2P 程序,运行时间为 1 d。其产生的流量同样在子网网关路由器上捕获并存储。

P2P 僵尸网络流量数据集。本文从第三方获取到了 P2P 僵尸网络的流量数据集[10,12]。此数据集包含了 13 个 Storm 僵尸主机的所有流量和 3 个 Waledac 僵尸主机的所有流量。

3.2 P2P 主机检测测试

本节首先分别测试 3 个 P2P 统计特征的区分性。为此,本文从随机选取的 6 个 P2P 主机的流量(来自 P2P 流量数据集)和 2 个非 P2P 主机的流量(来自非 P2P 流量数据集)中提取出此 3 个 P2P 主机统计特征的特征集,时间窗口大小设置为 5 min。特征区分性实验结果如图 2 所示。

由图 2 可知,3 个 P2P 统计特征对 P2P 流量和非 P2P 流量基本可以实现线性可分,也就是说,这 3 个特征在识别 P2P 流量时都具有较强的区分性。

随后,本文将 3 个特征组成特征向量,并以 Weka 为实验

平台,选取不同的机器学习方法,采用 10 倍交叉验证方法[18],测试系统在不同时间窗口大小时对 P2P 主机的检测效果。本文选取支持向量机、贝叶斯网络、朴素贝叶斯、随机森林和 J48 决策树 5 种机器学习算法分别进行实验。实验结果如图 3 所示。

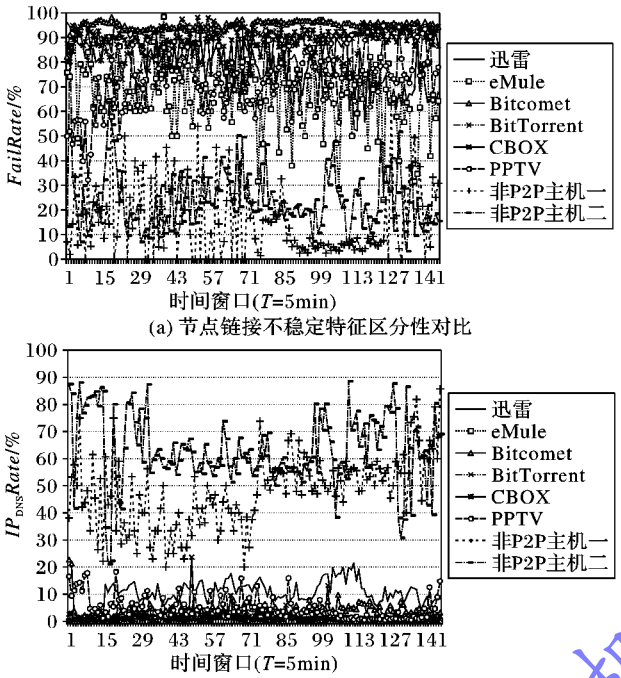

图 2　P2P 统计特征区分性实验结果

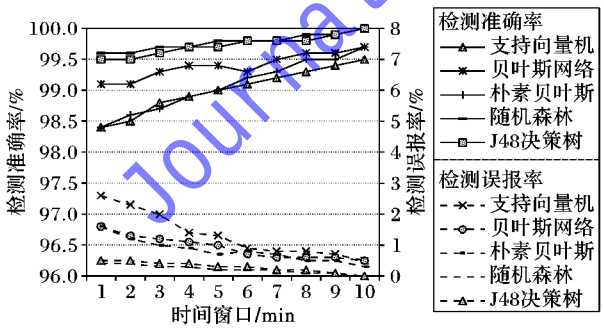

图 3　P2P 主机检测结果

从图 3 可以看出,时间窗口大小分别为 1 min 至 10 min 时,采用 5 种机器学习算法的检测准确率均能达到 98.4% 以上,误报率为 3% 以下。其中采用随机森林算法的检测准确率更是始终稳定保持在 99.6% 以上,误报率则始终低于 0.4%。

3.3　P2P 僵尸主机检测测试

本节同样先单独测试 2 个 P2P 僵尸主机统计特征的区分性。为此,本文从随机选取的 6 个 P2P 主机的流量(来自 P2P 流量数据集)和随机选取的 1 个 Storm 主机的流量及 1 个 Waledac 主机的流量(来自 P2P 僵尸网络流量数据集)中提取出此 2 个 P2P 僵尸主机统计特征的特征集,时间窗口大小设置为 5 min,相似区间长度 L 设置为 5。实验结果如图 4 所示。

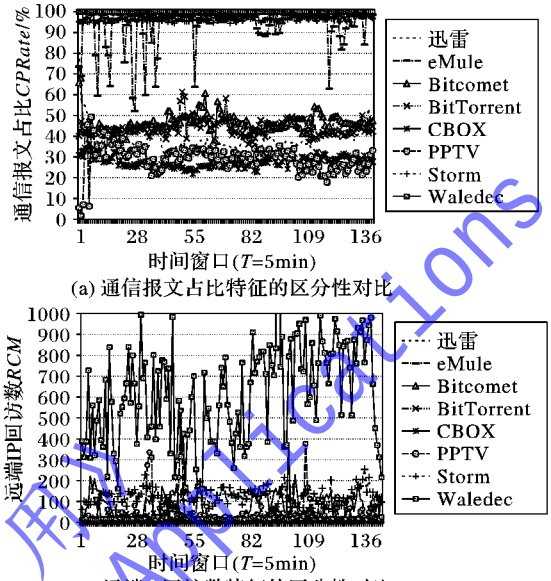

图 4　P2P 僵尸主机统计特征区分性实验结果

从图 4(a)可以看出,Storm 僵尸主机的通信报文占比一直稳定在 95% 以上,Waledac 僵尸主机的特征更是明显,一直处于 100%。而其他正常 P2P 主机所产生的通信报文占比基本处于 60% 以下,eMule 主机虽然通信报文占比也比较高,但其波动性很大,并不稳定。这可能是由于 eMule 主机下载的内容资源较少,eMule 程序长期处于通信报文交互阶段。

从图 4(b)可以明显看出,两个 P2P 僵尸主机远端 IP 回访数 RCM 明显高于其他正常 P2P 主机,特别是 Waledac 的 RCM 要远远超出其他程序,Storm 的 RCM 则在 100 左右波动,相比其他正常 P2P 主机区分性也比较明显。

最后,本文将此两个统计特征组成特征向量,同样以 Weka 为平台采用 10 倍交叉验证方法来评估系统对 P2P 僵尸主机的检测性能。图 5 给出了分别采用上述 5 种机器学习算法在不同时间窗口大小时系统对 P2P 僵尸主机的检测准确率和误报率。

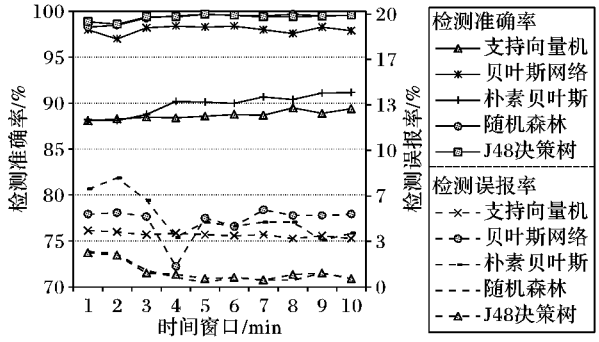

图 5　P2P 僵尸主机检测结果

从图 5 可以看出,采用随机森林算法时的检测效果最好,且在时间窗口大小为 5 min 时达到最佳。综上所述,本文选取随机森林算法作为本文系统的机器学习方法,系统的时间

窗口大小设置为5 min。系统检测准确率为99.7%,而误报率仅为0.3%。

4 结语

本文通过分析P2P网络和P2P僵尸网络的基本特性,提出了3个用于检测P2P主机的统计特征(节点链接不稳定性、节点发现模式、远端IP分布广度)和两个用于检测P2P僵尸主机的统计特征(通信报文占比和远端IP回访数)。在此基础上,本文设计并实现了一个P2P僵尸主机实时检测系统。相比已有的系统,本文系统具有以下特点:1)不需要检测报文负载,能检测加密流量;2)不依赖于僵尸主机的恶意活动,能检测处于隐匿阶段的P2P僵尸主机;3)所用特征少,时间窗口小,检测精度高,实时性强。实验结果证明,本文系统能在5 min内检测出监控网络内的所有P2P僵尸主机,检测准确率达到99.7%,误报率仅为0.3%。实验结果有效地证明了本文系统在准确率和实时性上的优越性。

参考文献:

[1] LIU L, CHEN S, YAN G, *et al*. Bot Tracer: execution-based bot-like malware detection [M]// ISC'08: Proceedings of the 11th International Conference on Information Security, LNCS 5222. Berlin: Springer, 2008: 97 – 113.

[2] SZYMCZYK M. Detecting botnets in computer networks using multi-Agent technology [C]// DepCos-RELCOMEX'09: Proceedings of the Fourth International Conference on Dependability of Computer Systems. Piscataway: IEEE, 2009: 192 – 201.

[3] STINSON E, MITCHELL J C. Characterizing bots' remote control behavior [C]// DIMVA'07: Proceedings of the 4th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Berlin: Springer, 2007: 89 – 108.

[4] XU K, YAO D, MA Q, *et al*. Detecting infection onset with behavior-based policies [C]// NSS 2011: Proceedings of the 5th International Conference on Network and System Security. Piscataway: IEEE, 2011: 57 – 64.

[5] GU G, PORRAS P, YEGNESWARAN V, *et al*. BotHunter: detecting malware infection through IDS-driven dialog correlation [C]// SS'07: Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium. Berkeley: Usenix Security. 2007: 1 – 16.

[6] SINGH K, GUNTUKU S C, THAKUR A, *et al*. Big data analytics framework for peer-to-peer botnet detection using random forests [J]. Information Sciences, 2014, 278: 488 – 497.

[7] JIANG H, SHAO X. Detecting P2P botnets by discovering flow dependency in C&C traffic [J]. Peer-to-Peer Networking and Applications, 2014, 7(4): 320 – 331.

[8] SILVA S S C, SILVA R M P, PINTO R C G, *et al*. Botnets: a survey [J]. Computer Networks, 2013, 57(2): 378 – 403.

[9] YU X, DONG X, YU G, *et al*. Online botnet detection based on incremental discrete Fourier transform [J]. Journal of Networks, 2010, 5(5): 568 – 576.

[10] ZHANG J, PERDISCI R, LEE W, *et al*. Building a scalable system for stealthy P2P-botnet detection [J]. IEEE Transactions on Information Forensics and Security, 2014, 9(1): 27 – 38.

[11] EN T F, REITER M K. Are your hosts trading or plotting? Telling P2P file-sharing and bots apart [C]// ICDCS 2010: Proceedings of the 2010 IEEE 30th International Conference on Distributed Computing Systems. Piscataway: IEEE, 2010: 241 – 252.

[12] RAHBARINIA B, PERDISCI R, LANZI A, *et al*. PeerRush: mining for unwanted P2P traffic [J]. Journal of Information Security and Applications, 2014, 19(3): 194 – 208.

[13] ZHAO D, TRAORE I, GHORBANI A, *et al*. Peer to peer botnet detection based on flow intervals [C]// SEC 2012: Proceedings of the 27th IFIP TC 11 Information Security and Privacy Conference on Information Security and Privacy Research. Berlin: Springer, 2012: 87 – 102.

[14] FINSTERBUSCH M, RICHTER C, ROCHA E, *et al*. A survey of payload-based traffic classification approaches [J]. IEEE Communications Surveys & Tutorials, 2014, 16(2): 1135 – 1156.

[15] LIU C, YANG Y, TANG C. A classification method of unstructured P2P multicast video streaming based on SVM [C]// MINES'09: Proceedings of the 2009 International Conference on Multimedia Information Networking and Security. Piscataway: IEEE, 2009: 68 – 72.

[16] HE J, YANG Y, QIAO Y, *et al*. Accurate classification of P2P traffic by clustering flows [J]. China Communications, 2013, 10(11): 42 – 51.

[17] HALL M, FRANK E, HOLMES G, *et al*. The WEKA data mining software: an update [J]. ACM SIGKDD Explorations Newsletter, 2009, 11(1): 10 – 18.

[18] KOHAVI R. A study of cross-validation and bootstrap for accuracy estimation and model selection [C]// IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1995: 1137 – 1145.

(上接第1869页)

[14] ZHU X. Security and privacy preservation mechanisms in vehicular Ad Hoc network [D]. Hefei: Hefei University of Technology, 2013. (朱晓玲. VANET安全和隐私保护机制研究[D]. 合肥: 合肥工业大学, 2013.)

[15] SWEENEY L. *k*-anonymity: a model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557 – 570.

[16] CAMENISCH J, GROTH J. Group signatures: better efficiency and new theoretical aspects [C]// Proceedings of the 4th International Conference on Security in Communication Networks, LNCS 3352. Berlin: Springer, 2005: 120 – 133.

[17] FAN L, CAO P, ALMEIDA J, *et al*. Summary cache: a scalable wide-area Web cache sharing protocol [J]. IEEE/ACM Transactions on Networking, 2000, 8(3): 281 – 293.

[18] BLOOM B H. Space/time trade-offs in Hash coding with allowable errors [J]. Communications of the ACM, 1970, 13(7): 422 – 426.

[19] BONOMI F, MITZENMACHER M, PANIGRAHY R, *et al*. An improved construction for counting bloom filters [M] // ESA 2006: Proceedings of the 14th Annual European Symposium on Algorithms, LNCS 4168. Berlin: Springer, 2006: 684 – 695.

[20] GUO D, LIU Y, LI X, *et al*. False negative problem of counting bloom filter [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(5): 651 – 664.