一种 p2p Botnet 在线检测方法研究

柴 胜[1],胡 亮[1],梁 波[2]
(1.吉林大学计算机科学与技术学院,吉林长春 130012; 2.国家质检总局信息中心,北京 100088)

摘 要: 文章详细分析了 p2p 僵尸网络的生命周期以及网络特征,利用改进的 SPRINT 决策树和相似度度量函数,提出了一种新的在线综合检测方法,并论述了虚拟机环境搭建、原型系统设计和实验结果分析.实验结果表明,检测方法是可行的,具有较高的效率和可靠性.

关键词: 僵尸网络; 对等网络; 检测
中图分类号: TP302.7 **文献标识码:** A **文章编号:** 0372-2112 (2011) 04-0906-07

The p2p Botnet Online Detect Approach Research

CHAI Sheng[1], HU Liang[1], LIANG Bo[2]
(1. *College of Computer Science and Technology Institute, Jilin University, Changchun, Jilin* 130012, *China*;
2. *AQSIQ Information Center, Beijing* 100088, *China*)

Abstract: The life cycle and network characteristics of p2p botnet was analyzed, improved sprint method and similarity function was discussed and new online detect approach was proposed. Secondly, the virtual machine network environment, the design of prototype system and the analysis of the experiment result was discussed. The result shows that the detection approach is feasible, high efficient and reliable.

Key words: botnet; p2p network; detect

1 引言

僵尸网络已经成为互联网安全的最大威胁之一.僵尸网络是指采用各种传播手段,将大量主机感染僵尸程序,从而在控制者和被感染主机之间所形成的一个可一对多控制的网络.控制者可以向僵尸网络中被感染主机下达各种恶意指令,例如扫描感染其它主机、发送垃圾邮件、偷取主机信息、发动 DDOS 攻击等.按照命令与控制机制[1],僵尸网络可以分为基于 IRC 协议、HTTP 协议和 p2p 协议三类.基于 IRC 协议和 HTTP 协议的僵尸网络[2]具有集中控制服务器,并且数量庞大,易于检测和防御.目前已被广泛应用的反制方法是阻止被感染主机与控制服务器通信,例如通过 CERT 部门协调关闭僵尸网络集中控制服务器.然而,基于 p2p 协议的僵尸网络没有集中控制点,更具韧性和隐蔽性,难于检测和防御.文献[3]通过对 Peacomm 僵尸程序的分析指出了僵尸网络向 p2p 协议结构发展的趋势.

基于 p2p 协议的僵尸网络中,每个僵尸结点往往与多个对等结点连接,结点间从开始连接到正常通信存在大量网络通信数据.大多情况下,由于每个结点的僵尸程序是相似的、自动运行的和无人工干预的,所以可以通过它们之间的网络数据来发现某些特征模式,进一步通过这些特征研究相关检测方法.本文将 p2p 僵尸网络生命周期分为三个阶段,并将三个阶段的网络数据分为行为特征和通信特征两部分,然后利用一种新的在线综合检测方法对其检测.

2 相关工作

僵尸网络的研究主要集中在四个方面,包括分析、跟踪、检测和反制.目前,僵尸网络的反制研究面临严峻的挑战.本文主要研究一种基于 p2p 协议的僵尸网络检测方法,研究者给出了许多方法来进行僵尸网络检测,主要分为两类,包括基于主机的检测和基于网络的检测.防病毒软件主要采用基于主机的检测方法,它们将各种僵尸网络及其变种当作木马或者病毒处理,主要采用特征码匹配技术或者恶意行为分析来检测各种僵尸程序.优点是对于流行僵尸网络能够相对及时发现并处理,缺点是不能发现智能变种僵尸网络,不能对远程控制者进行报告和防御等,另外各种杀病毒软件能力不同、终端用户需要安装杀病毒软件、终端用户安全意识

收稿日期:2010-03-02;修回日期:2010-09-26
基金项目:国家重点基础研究发展计划(973 计划)(No.2009CB320706);吉林省科技发展计划资助项目(No.20090111)

淡薄都影响僵尸程序的检测.基于网络的检测方法主要针对网络流量数据进行分析,通过流量数据的异常特征来判断网络内主机是否感染僵尸网络.目前这些方法主要侧重基于 IRC 协议和 HTTP 协议的僵尸网络检测[4~7],然而现在的僵尸网络已经由原来的集中式结构逐步向分布式结构(即采用 p2p 协议)发展[8],针对 p2p 僵尸网络的分析与检测已经成为该领域的研究热点,下面的一些方法和系统提供对 p2p 僵尸网络的检测支持.

Gu 等人研究实现了 BotMiner[9]原型系统,提出了一种两阶段聚类方法,能够同时检测 IRC/HTTP/p2p 三种类型僵尸网络.BotMiner 目前停留在原型阶段,准确率不高,不能在线进行,并且针对僵尸网络有某些固定的假设,无法保证方法针对 botnet 新变种的适应性.Sang[10]等人对 p2p 僵尸网络特征进行建模,然后利用 Markov 模型进行状态转换与匹配,并将该方法应用到三种典型 p2p botnet 进行验证.该方法的特征建模并没有考虑僵尸网络的主流特征,例如发垃圾邮件,下载二进制代码等,并且误报率高.文献[11]提出利用 CUSUM 算法对 p2p 僵尸网络进行检测,检测的特征主要包括 ICMP 错误率、SMTP 协议包增长率、UDP 包增长率.首先该方法的特征建模并不完善,没有考虑 p2p botnet 的连接相似性、包大小、恶意扫描信息等特征;其次该方法重点考虑了 p2p 僵尸网络初始阶段的特征,对其它阶段的特征没有考虑.Mohammad[12]等人提出一种综合分类数据挖掘方法,通过不断的选择最优的分类器处理流数据,可以用作 p2p 僵尸网络的检测.该方法只考虑了 p2p 僵尸网络初始化阶段的一些特性,对僵尸网络结点间的通信相似性没有考虑,容易导致误报率较高.

综上所述根据是否能够在线检测,基于网络的 p2p botnet 检测方法可分为两种:一种是离线检测,即预先收集一段时间内的网络流量数据,然后利用各种算法对离线数据进行分析与检测;另一种是在线检测,即在线监控待检网络,接收相关流量数据,对网络数据流分析并检测,然后报告疑似感染主机.本文提出的是一种在线进行 p2p botnet 检测方法,在线检测与非在线检测相比,具有高效、快速检测、快速发现等特点,适合应用在安全等级较高的网络.

3 p2p 僵尸网络生命周期与特征

Wang[13]等人认为 p2p 僵尸网络分为三个阶段,包括选择成员阶段、组成僵尸网络阶段和等待指令阶段.选择成员阶段即利用各种方式感染大量主机,组成僵尸网络阶段即被感染的主机之间利用 p2p 协议组成僵尸网络,等待指令阶段即被感染主机等待僵尸网络控制者指令.这样的划分是根据僵尸程序执行顺序来理解的,站在网络数据流检测的角度,我们也将 p2p 僵尸网络分为三个阶段,包括初始阶段、发呆阶段和攻击阶段.其中初始阶段是主机感染 p2p 僵尸网络程序,并且加入到该 p2p 僵尸网络中,中间经历注入、感染、加入僵尸网络、下载二进制僵尸升级程序等过程.p2p 僵尸网络的初始阶段产生大量特殊的网络数据,在该阶段进行检测相对容易.发呆阶段是网络内感染 p2p 僵尸程序的主机启动后只是连接到对等结点,与僵尸网络保持简单通信和连接,不做其它事情.发呆阶段是僵尸网络最常见的阶段,僵尸网络生命周期的大部分时间处于发呆阶段,该阶段不产生过多的容易采集的网络数据,所以在该阶段进行网络检测难度比较大而且容易产生误报.攻击阶段是受控制的主机接收 p2p 僵尸网络控制者的指令进行破坏活动,常见的活动包括发送垃圾邮件、进行 DDOS 攻击、偷取受控主机信息等,其中发送垃圾邮件和 DDOS 攻击由于产生大量特殊网络数据相对容易检测.

总结上述各个阶段,p2p 僵尸程序有以下特点:

(1)ICMP 异常报文.某些主机感染 p2p 僵尸程序,在初始化时,僵尸程序会随机的向其他节点发送连接请求,这样就导致了出现大量的“找不到目的地址”的异常的 ICMP 报文错误.(初始阶段)

(2)ARP 异常请求.某些 p2p 僵尸程序内部存在多个固定连接 IP,初始化时,僵尸程序会向这些 IP 发送 ARP 请求,导致 ARP 协议报文突然增长.(初始阶段)

(3)主机端口连接建立速率.大部分 p2p 僵尸程序会在初始化时短时间内连接很多对等结点,如果这些连接建立成功,可以在主机端口检测某段时间的主机间连接建立情况来发现僵尸程序的行为特征.(初始阶段)

(4)特定时间段内数据包相似.网络内感染同一 p2p 僵尸网络的若干主机表现的行为相似,因为这些僵尸程序都是自动运行,不受使用者控制.无论这些主机之间通信,还是它们与外网控制主机通信,可以假设这些主机产生了大量大小相似的通信包.(发呆阶段)

(5)内外网相同连接.经过观测,网络内正常主机与感染 p2p 僵尸网络的主机之间的通信连接呈现聚集特征.即假设网络内部分主机感染同一 p2p 僵尸网络,外网对等控制结点为 A、B、C,内网感染主机与 ABC 的连接呈现聚集特征.对于网络内主机感染不同的 p2p 僵尸网络,内网和外网都有对等控制结点的情况与上述假设类似.(发呆阶段或者攻击阶段)

(6)SMTP 报文.某些僵尸程序会建立大量 SMTP 连接并发送大量 SMTP 报文,根据这一特点可以检测相应 p2p 僵尸程序.(攻击阶段)

上述特征没有涵盖 DDOS 攻击特征,本文没有将

p2p 僵尸网络产生的 DDOS 攻击检测作为研究内容,因为:(1)僵尸网络发动 DDOS 攻击行为在整个生命周期中只占有很小的时间段;(2)DDOS 攻击检测是一个充分研究并且正在发展的课题,拟以后结合成熟成果研究该特征.

4　p2p botnet 在线检测方法

目前,无论僵尸网络采用自定义协议还是已有的应用层协议来进行通信,都离不开传输层 TCP 和 UDP 协议,因此,可以利用协议分析技术,只采集 TCP 或者 UDP 相关数据流来进行分析.从检测角度,将上面三个阶段的特征分为两类,一类是 p2p 僵尸网络结点间连接通信特征,包括特定时间段内数据包相似、内外网相同连接;另一类是僵尸结点行为特征,包括 SMTP 异常报文、ICMP 异常报文、ARP 异常请求、主机端口连接建立速率.基于 p2p botnet 三个阶段的通信特征和行为特征,本文提出一种新的基于网络特征的在线检测方法,该方法结合改进的 SPRINT 决策树分类算法和基于相似度度量函数的算法进行检测,然后对两方面进行综合分析进而决定相关主机是否是 p2p 僵尸主机.

4.1　改进的 SPRINT 决策树分类方法

p2p botnet 行为特征可以在虚拟机网络环境中采集和提取(参见第 5 节),然后通过分类方法进行训练产生决策树,最后在待检网络中利用相应决策树模型进行检测.数据分类方法有很多,其中决策树分类具有速度快和精度高特点,适合在线检测.针对待检网络收集到的 p2p botnet 行为特征是一种网络海量数据,传统分类方法在处理海量数据时会出现性能或精度降低的问题,IBM 研究人员提出了决策树分类算法 SPRINT[14],它是一种快速的、可伸缩的、适合处理较大规模的决策树分类算法,而且消除了所有的内存限制、性能较好.针对 p2p botnet 行为特征,本文采用改进的 SPRINT 决策树分类方法,在虚拟环境中部署典型 p2p bot,采集相关数据训练决策树,然后在待检网络中进行检测.

利用数据采集程序收集决策树训练的数据集,采用表 1 数据表结构,其中 TableHost 存储主机及其开放的端口,TablePort 存储在某一端口发送和接收的数据流数据.经过观测,p2p 僵尸程序感染后必然会监听本地主机的某一端口,所以只需要考虑网络内主机开放的端口即可,这样可以大大节约检测开销.

表 1　数据表结构

TableHost(ID,HOST,PORT)
TablePort(ID,PORT,DATA)

表 1 中 DATA 不是原始的通信数据,而是经过处理的特征向量数据 $\boldsymbol{D}$,$\boldsymbol{D}=(F_1,F_2,F_3,F_4)$,$F_i$ 即为上述 p2p botnet 行为特征,这些特征利用数据归约技术,使用直方图分箱来近似数据分布.举例来说,僵尸程序发送垃圾邮件表现为发送大量 SMTP 报文,可以采用 $\boldsymbol{F}_1=\{[0,2^5],\cdots,[2^f,2^{f+1}],\cdots,[2^n,\infty]\}$ 来表示,直方图中每个分箱代表发送 SMTP 垃圾报文的数量.ICMP 异常报文可以表示为 $\boldsymbol{F}_2=\{[0,1],\cdots,[N_1,\infty]\}$,$N_1$ 为直方图箱子数量;ARP 异常请求可以表示为 $\boldsymbol{F}_3=\{[0,1],\cdots,[N_2,\infty]\}$,$N_2$ 为直方图箱子数量;主机端口连接建立速率可以表示为 $\boldsymbol{F}_4=\{[0,1],\cdots,[N_3,\infty]\}$,$N_3$ 为直方图箱子数量.其中更新时间和直方图箱子数量根据训练和检测程序执行情况进行优化.

上述数据集作为输入,然后采用改进的 SPRINT 算法[15]生成决策树作为输出.SPRINT 算法目前存在的主要问题是最佳分割点计算量大,改进后的方法能够大大减少候选分割点的数目,从而有效地减少了计算量和计算时间.

输入:在虚拟机组成的网络中部署各种 p2p,采集行为特征数据存储为 TableHost 和 TablePort.

输出:决策树分类器 T

采用改进的 SPRINT 算法,主要步骤如下:

(1)如果结点 S(S 代表主机,即该主机上通信端口上所有特征数据的集合)满足停止条件,则返回;

(2)对于各个属性 F_i,找到一个值或者值集,产生最佳分裂;

(3)比较各个属性的最佳分裂,选择最佳的将 S 分为 S_1 和 S_2;

(4)递归对 S_1 和 S_2 产生决策树 T;

在待检网络中采用上述训练过的决策树分类器进行检测,方法如下:

输入:待检网络中,采集某时间段内行为特征数据存储为 TableHost 和 TablePort

输出:疑似 p2p botnet 主机集合 K

(1)遍历 TableHost 和 TablePort,得到每个主机上所有开放端口的特征向量数据 $\boldsymbol{D}$;

(2)将 $\boldsymbol{D}$ 与决策树 T 中分支进行匹配,如果满足异常条件,则将该主机加入集合 K,否则继续遍历;

(3)得到疑似 p2p botnet 主机集合 K,$K=\{k_1\cdots k_i\cdots k_n\}$.

SPRINT 算法产生的二叉树检索时间复杂度为 $O(\log(n))$,由上述步骤不难看出检测方法的时间复杂度为 $O(h*p*\log(n))$,其中 h 为主机数量,p 为在主机上平均开放的端口数量.

4.2　基于相似度度量函数的检测方法

p2p 僵尸网络结点间连接通信特征,包括若干时间段内数据包相似、内外网相同连接.这些特征本质上可以归为一类,即 p2p 僵尸网络结点间相似性,发现这种

相似性是检测的本质.p2p 僵尸网络结点间相似性的发现可以通过计算某段时间结点发送或者接收数据包数量、字节大小来实现,如果在很多定义时间内,两个结点发送或者接收的数据包数量相似、每个包字节大小相似,则认为结点间具有某种相似性.目前,这种结点间相似性可以通过聚类分析来实现,例如文献[14]利用一种采集程序先收集网络数据,存储后利用聚类方法离线进行聚类分析来计算相似性,但是这种方法不具备在线分析和检测的能力.

利用数据采集程序收集到网络数据流具有大量、快速和连续到达等特点,要想在线处理和分析数据流,聚类分析方法必须一遍扫描并且使用有限的空间和内存,即采用专门针对数据流的聚类分析方法.目前,数据流在线聚类分析方法的研究处于初始阶段,相关现有算法存在效率和质量的矛盾,本文提出一种新的基于相似度度量的在线方法,检测 p2p botnet 结点通信相似性.

定义 1 设集合 $R=\{r_1,r_2,\cdots,r_i,\cdots,r_n\}$ 为待检子网某时间段内数据流集合,其中 n 为主机个数;r_i 为某个主机的相应时间段内数据流集合,包括发生时间,持续时间,源 IP 地址,源端口,目的 IP 地址,目的端口,包总数,字节总数等属性.

定义 2 根据定义 1 中集合 R,计算每个主机的通信特征,包括时间段内数据流记录数 a、每个数据流包含数据包数量 b、每个数据包字节大小 c、窗口时间内每秒传输字节大小 d;设集合 $H=\{h_1\cdots h_i\cdots h_n\}$,其中 n 为主机个数,H 为待检网络内主机集合,$\boldsymbol{h}_i$ 代表某个主机的通信特征向量,$\boldsymbol{h}_i=(a,b,c,d)$.

a,b,c,d 四个通信特征的数据表示并不一致,可以采用直方图将它们映射到同一个空间表示,设每个特征的直方图箱子数量为 10,这样 $\boldsymbol{h}_i=(a_1\cdots a_{10},b_1\cdots.b_{10},c_1\cdots c_{10},d_1\cdots.d_{10})$.

$\boldsymbol{h}_i$ 为高维数据,针对高维数据的聚类分析算法时间复杂度偏高,不适合应用在在线检测的环境中,本文提出一种新的基于相似度度量的方法来检测主机结点的通信相似性问题,其中采用的相似度度量函数 Hsim 来自文献[16],它是一种专门针对高维数据的相似度度量函数,能够克服其它距离函数在高维空间中的缺点.

基于相似度度量的方法分两阶段,包括训练阶段和检测阶段,其中训练阶段的网络环境为虚拟机环境,在该环境中部署若干 p2p botnet 及变种.

训练阶段方法如下:

输入:虚拟机网络环境下 bot 主机数据流集合 R

输出:p2p bot 结点的通信特征集合 E

(1)遍历集合 R,计算 p2p bot 主机通信特征 $\boldsymbol{h}_i$,其中包括 $\boldsymbol{h}_i$ 的数据归约,得到 bot 主机的通信特征集合 $H=\{h_1\cdots h_i\cdots h_n\}$;

(2)利用两重循环遍历集合 H,利用相似度度量函数 Hsim 计算各个 bot 主机的通信特征,如果任意两个主机的通信特征相似度超过阀值,则去除 H 中重复的通信特征;

(3)得到新的通信特征集合 E,$E=\{e_1\cdots e_i\cdots e_j\cdots e_n\}$,其中 e_i 和 e_j 不相似.

检测阶段方法如下:

输入:p2p bot 结点的通信特征集合 E 和待检网络某时间段数据流集合 R

输出:疑似的 p2p bot 主机集合 S

(1)遍历集合 R,计算各个主机通信特征 $\boldsymbol{h}_i$,其中包括 $\boldsymbol{h}_i$ 的数据归约,得到所有主机的通信特征集合 $H=\{h_1\cdots h_i\cdots h_n\}$;

(2)利用两重循环遍历集合 E 和集合 H,利用相似度度量函数 Hsim 计算 e_i 和 $\boldsymbol{h}_i$ 的相似度,如果相似度超过阀值,则将该主机加入集合 S;

(3)得到疑似的 p2p bot 主机集合 S,$S=\{s_1\cdots s_i\cdots s_n\}$.

上述检测阶段算法的时间复杂性为 $O(n+a*b*h)$,其中 n 为计算主机通信特征时间,a 为集合 E 数量,b 为集合 H 数量,h 为函数 Hsim 计算时间,其中集合 E 元素数量通常很小.

利用上述两阶段方法发现的主机集合未必是 p2p 僵尸网络所控制的主机结点,必须将行为特征和通信特征的检测结果结合起来进行综合分析才能进一步提高判别的准确率.

4.3 综合分析

单独通过行为特征和通信特征检测都不足以准确地定位网络中的 p2p 僵尸主机,将二者结合起来进行综合分析能提高检测的准确率.

定义 3 设行为特征异常的主机集合 K,$K=\{k_1\cdots k_i\cdots k_n\}$,$n$ 为主机数量,权值为 a,$0<a\leqslant 1$;通信特征异常的主机集合 S,$S=\{s_1\cdots s_i\cdots s_m\}$,$m$ 为主机数量,权值为 b,$0<b\leqslant 1$;$g(h)=a*K(h)+b*S(h)$,其中 h 为待检网络中某主机,g 为该主机的疑似函数,$0\leqslant g\leqslant 2$;若 $h\in K$,则 $K(h)=1$,否则 $K(h)=0$;若 $h\in S$,则 $S(h)=1$,否则 $S(h)=0$.

p2p botnet 在线检测总体步骤如下:

(1)对某时间段内的网络数据流进行预处理,为行为特征检测和通信特征检测提供相应数据集;

(2)利用改进的决策树分类产生异常主机集合 K;

(3)利用基于相似度度量函数的检测方法产生异常主机集合 S;

(4)计算待检网络内主机的疑似函数 $g(h)$,如果超过预设阈值,则将该主机加入疑似主机集合 $\boldsymbol{G}$;到达指定检测时间,转到步骤 5,否则继续下一时间段的检测,返回步骤 1.

(5)算法结束时得到疑似主机集合 $\boldsymbol{G}$,$\boldsymbol{G}=\{g_1\cdots g_i\cdots g_n\}$,$g_i$ 是疑似函数超过预设阈值的疑似主机.

根据上述算法步骤不难看出,p2p botnet 在线检测的时间复杂度主要在于步骤(2)或者(3)(具体实现可以采用线程并行计算处理),而步骤(2)的时间复杂性为 $O(h*p*\log(n))$,步骤(3)的时间复杂度为 $O(n+a*b*h)$,可见总体算法的时间复杂度是满足在线检测需要的.

5　系统与实验

5.1　虚拟机环境搭建

很多研究者使用 ns2 建立仿真的入侵检测、网络蠕虫等测试环境,模拟大规模结点的网络环境,捕获相关数据进行研究.但是,上述仿真过程太过单一和理想化,p2p bot 从感染、加入 botnet、二次注入到发垃圾邮件、DDOS 攻击是一个复杂的过程,ns2 无法仿真上述真实情况,且对实验主机的性能要求比较高.因此,ns2 不适合应用于新型 p2p 僵尸网络的模拟和实验.最近几年,虚拟机技术不断发展并得到广泛应用,其中 VMware 是这方面的代表软件.许多研究人员,尤其是安全研究人员开始应用虚拟机技术开展相关研究,包括木马病毒等观测、入侵检测、僵尸网络研究等.首先,将虚拟机软件安装到物理主机上,然后在物理主机内可以模拟多个操作系统,并将这些虚拟机连接到网络上形成局域网.上述虚拟机实验环境能够完全模拟真实操作系统和主机,外部主机或者访问者基本不能察觉真实主机和虚拟主机的差异.目前,虚拟机最大的问题就是性能问题,VMware 具有复杂而庞大的功能,但是在性能上并不占有优势,本文采用 Sun 公司开源虚拟机工具 VirtualBox 来搭建虚拟机实验环境,VirtualBox 具备基本的虚拟机功能,最大的优势是节省物理主机的资源.

经过试验,一台 2G 内存的主流配置 PC 机可以安装 VirtualBox 模拟五个网络节点.目前,我们已经利用 10 台上述配置主机建立好相关虚拟环境,大概能模拟 100 ~ 120 个结点.该环境可以独立反复使用,能灵活地配置每个节点和网络拓扑拓扑,在感染 p2p 僵尸网络后可获取用于分析的网络数据和感染情况.

5.2　原型系统

如图 1 所示,原型系统包括训练子系统、检测子系统和中心控制子系统.

(1)图 1 中所述训练子系统包括网络接收模块、网络数据流预处理模块、分类训练模块和通信特征预处理模块,网络接收模块负责监控和接收虚拟机网络数据,数据流预处理模块对初始阶段 p2p 僵尸网络特征进行预处理,分类训练模块是根据 p2p 僵尸程序特征集合采用改进的 SPRINT 方法来训练分类决策树,通信特征预处理模块是根据已经部署的 p2p bot 及其变种计算相应结点的通信特征向量;

(2)图 1 中所述检测子系统包括初始化模块、网络接收模块、网络数据流预处理模块和综合检测模块,初始化模块用于初始化典型 p2p botnet 特征、训练好的分类决策树、p2p botnet 通信特征向量集合以及与中心控制模块保持数据同步;网络接收模块用于监控和接收待检网络通信数据并且负责按照窗口时间以及协议类型进行数据接收;网络数据流预处理模块用于对真实网络中接收的数据进行预处理;综合检测模块负责根据初始化的 p2p botnet 在线检测方法进行综合检测,检测的结果是所在网段内疑似感染 p2p botnet 的主机信息以及其连接的外网对等控制主机信息;

(3)图 1 中所述中心控制子系统包括在线更新模块统计报表模块,在线更新模块负责根据需要更新分类决策树和通信特征数据,通知检测子系统初始化模块,接收检测子系统的检测结果;报表统计模块给出疑似主机的报表和统计分析.

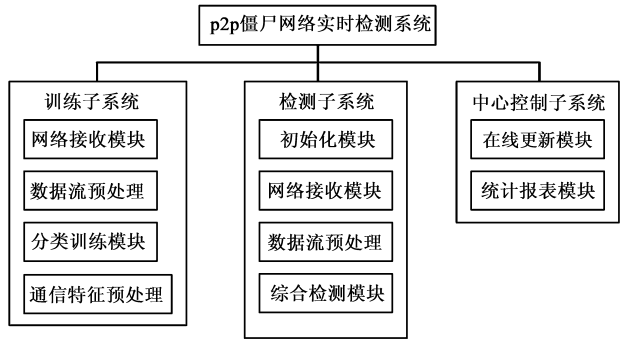

图1　原型系统组成

5.3　分类器训练和通信特征预处理

如图 2 所示,p2p botnet 训练子系统包括网络接收模块、网络数据流预处理模块、分类训练模块和通信特征预处理模块.

P2p 僵尸网络可能为表现出一些行为特征,例如 ICMP 和 ARP 异常报文、瞬时连接建立增加和 SMTP 垃圾报文等.针对上述特征,首先建立虚拟机网络环境,然后部署流行 p2p 僵尸程序,收集典型初始化数据流并进行数据预处理.针对这些预处理后的数据流,利用分类器训练方法展开训练,形成分类器决策树.

在虚拟机网络环境中,随机挑选若干虚拟结点部署 Peacomm、Nugache、Sinit 等僵尸程序或其变种(p2p 僵尸程序及变种可以下载获取),选择其余虚拟主机中任一台作为训练主机,安装部署 p2p botnet 训练子系统.其

中网络接收模块收集虚拟网络内通信数据流,然后网络数据流预处理模块进行数据预处理工作,即将数据流根据需要检测的特征进行数据概要和格式化.分类训练模块根据预处理好的数据流进行分类训练并产生较优的分类决策树,保存后的决策树可以导出到实际检测子系统中.

通信特征预处理模块是根据已经部署的 p2p bot 及其变种计算相应结点的通信特征向量,并且根据相似度度量函数计算这些特征向量间的相似度,去除相似的通信特征向量,保存后的 p2p bot 通信特征向量集合可以导到实际检测子系统中.

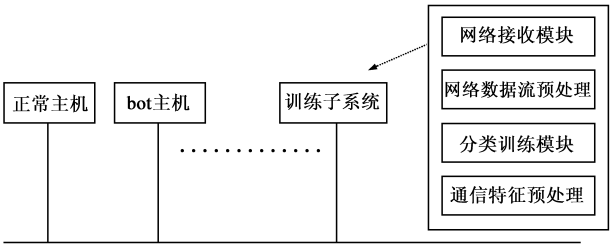

图2 虚拟机网络环境与p2p botnet训练子系统

5.4 检测系统在实际网络环境中的应用

如图 3 所示,系统包括检测子系统(初始化模块、网络接收模块、数据流预处理模块、综合检测模块)和中心控制子系统(在线更新模块和统计报表模块).实际应用时,管理人员在待检网络的出入口部署检测子系统和中心控制子系统.

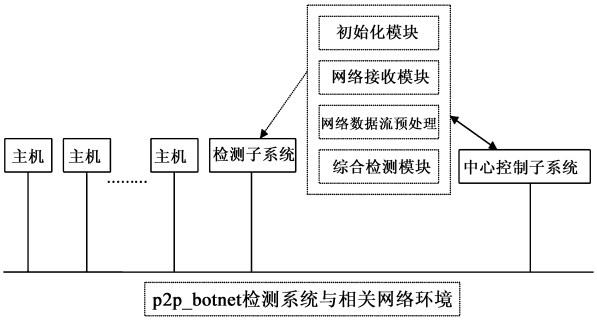

图3 p2p_botnet检测系统与相关网络环境

初始化模块负责初始化检测子系统需要的初始参数以及相关模型,包括滑动窗口时间、综合检测算法参数等,另外初始化模块与中心控制子系统建立通信连接,接收训练好的初始阶段分类决策树和通信特征向量集合等数据.

网络接收模块负责网络原始数据的采集,利用 libpcap 捕获函数库实现了网络接收模块.libpcap 是一个数据包捕获函数库,它能高效和快速的捕获和接收以太网网络内原始数据流.

数据流预处理模块在网络接收模块收集原始数据的同时,按照 p2p botnet 的特征对数据进行预处理工作.根据窗口时间设置不同,网络接收模块和数据流预处理模块处理完当前时间段数据,将预处理数据保存后,继续接收下一时间窗口数据.预处理后的格式化数据流分配给综合检测模块处理.

综合检测模块根据数据库中预处理后的格式化数据,采用第 4 节介绍的在线 p2p 僵尸网络检测方法进行疑似主机检测工作,疑似主机及相关信息上报给中心控制子系统进行处理.

5.5 实验分析

为了验证方法的有效性,我们选择一个具有 80 台主机的网络实验室,所有主机均具备连接互联网的能力.经过实验分析,将通信特征疑似函数阀值设为 0.8,通信特征和行为特征权值分别设为 0.9 和 0.7,其中收集网络数据的窗口时间段设为 30min.另外包括已经在虚拟环境训练好的决策树、p2p bot 通信特征数据集;然后利用一个定时自动化程序下载最新的 p2p 僵尸网络程序 Peacomm、Nugache、Sinit 变种执行,其中执行每种僵尸程序的主机数量为 5,共计 15 台主机,其它 65 台主机为正常主机.原型系统经过 60min 的检测,产生两个实验结果,经过人工分析,如表 2、表 3 所示.

表 2 前 30min 检测结果

类别	数量	检测疑似主机	检测率	误报率
Peacomm	5	5	100%	—
Nugache	5	5	100%	—
Sinit	5	5	100%	—
正常主机	65	2	—	3%

表 3 后 30min 检测结果

类别	数量	真实感染主机	检测疑似主机	检测率	误报率
Peacomm	5	4	4	100%	—
Nugache	5	2	2	100%	—
Sinit	5	3	3	100%	—
正常主机	71	0	2	—	2.8%

表 2 和表 3 结果的不同体现了 p2p botnet 的生命周期特性,一般主机在感染 p2p bot 程序后初始阶段会产生大量的行为特征数据,但是 p2p bot 连接对等结点有一定的错误率.表 2 的检测结果表明原型系统能检测出正在感染 p2p bot 的主机,表 3 的结果表明经过一段时间后原型系统检测的是最终感染的疑似主机,其中真实感染主机的数量是人工检查和分析的结果.

另外,继续增加检测的时间,经过原型系统 24 个小时的检测,又产生如下实验结果,统计每个时间窗口的实验结果表明原型系统的误报率在 3%左右.

表 4 一天后检测结果

类别	数量	真实感染主机	检测疑似主机	检测率	误报率
Peacomm	5	4	4	100%	—
Nugache	5	2	2	100%	—
Sinit	5	3	3	100%	—
正常主机	71	0	3	—	4.2%

6 结束语

传统僵尸网络正朝着以 p2p 为代表的分布式结构发展,本文详细分析了 p2p 僵尸网络的生命周期以及网络特征,提出了一种新的在线综合检测方法,在此基础上论述了虚拟机环境的搭建、原型系统的设计和实验分析.实验分析结果表明检测方法是可行的.目前上述实验仅仅是针对 p2p 僵尸网络来进行研究,下一步准备涵盖基于 IRC 和 HTTP 协议的僵尸网络,同时研究提高检测方法的效率和减少原型系统误报率,并进一步分析新的 p2p botnet 变种以及研究原型检测系统在真实网络环境中的应用.

参考文献

[1] 诸葛建伟.僵尸网络研究[J].软件学报,2008,19(3):702 – 715.
ZHUGE Jian-Wei. Research and development of botnets[J]. Journal of Software,2008,19(3):702 – 715.(in Chinese)

[2] M A Rajab,J Zarfoss,F Monrose,A Terzis. A multifaceted approach to understanding the botnet phenomenon[A]. Proc of the 6th ACM SIGCOMM on Internet Measurement Conference[C]. Rio de Janeriro: Association for Computing Machinery, 2006.41 – 52.

[3] J B Grizzard, V Sharma, C Nunnery, B B Kang, D Dagon. Peer-to-peer botnets: Overview and case study[A]. Proc USENIX HotBots'07[C]. Berkeley: USENIX Association, 2007.78 – 83.

[4] J R Binkley, S Singh. An algorithm for anomaly-based botnet detection[A]. Proceedings of USENIX SRUTI'06[C]. Berkeley: USENIX Association,2006.43 – 48.

[5] J Goebel, T Holz. Rishi: Identify bot contaminated hosts by irc nickname evaluation[A]. Proceedings of USENIX HotBots'07[C]. Berkeley: USENIX Association,2007.53 – 60.

[6] C Livadas, R Walsh, D Lapsley, W T Strayer. Using machine learning techniques to identify botnet traffic[A]. Proceedings of the 2nd IEEE LCN Workshop on Network Security[C]. Tampa: IEEE Computer Society,2006.967 – 974.

[7] W T Strayer, R Walsh, C Livadas, D Lapsley. Detecting botnets with tight command and control[A]. Proceedings of the 31st IEEE Conference on Local Computer Networks (LCN'06)[C]. Tampa: IEEE Computer Society,2006.195 – 202.

[8] R Lemos. Bot software looks to improve peerage[EB/OL]. Http://www.securityfocus.com/news/11390,2006.

[9] Guofei Gu, Roberto Perdisci, Junjie Zhang, Wenke Lee. BotMiner: Clustering analysis of network traffic for protocol- and structure-independent botnet detection[A]. Proceedings of the 17th USENIX Security Symposium (Security'08)[C]. Berkeley: USENIX Association,2008.139 – 154.

[10] Sang-Kyun Noh, Joo-Hyung Oh, Jae-Seo Lee, Bong-Nam Noh, Hyun-Cheol Jeong. Detecting p2p botnets using a multi-phased flow model[A].3rd International Conference on Digital Society[C]. Cancun: Computer Society,2009.247 – 253.

[11] Jian Kang, Jun-Yao Zhang, Qiang Li, Zhuo Li. Detecting new p2p botnet with multi-chart CUSUM[A]. International Conference on Networks Security, Wireless Communications and Trusted Computing[C]. Wuhan: Computer Society,2009.688 – 691.

[12] Mohammad M Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani Thuraisingham. A multi-partition multi-chunk ensemble technique to classify concept-drifting data streams[A]. Proceedings of The 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'09)[C]. Bangkok, Thailand: Springer Verlag,2009.363 – 375.

[13] Ping Wang, Lei Wu. A systematic study on Peer-to-Peer Botnets[A]. Computer Communications and Networks[C]. San Francisco: Institute of Electrical and Electronics Engineers Inc,2009.121 – 128.

[14] Shafer J, Agrawal R, Mehta M. SPRINT: A scalable parallel classifier for data mining[A]. Proceedings of the 1996 International Conference on Very Large Data Bases. Bombay[C]. San Francisco: Morgan Kaufmann Publishers Inc,1996.544 – 555.

[15] Liu, Youjun. Improvement of SPRINT algorithm[J]. Jisuanji Gongcheng/Computer Engineering,2006,32(16):55 – 57.

[16] 杨风召.高维数据挖掘技术研究[M].南京:东南大学出版社,2007.

作者简介

柴　胜　男,1976 出生,吉林长春人,讲师,主要研究领域为安全软件工程.　E-mail:chaisheng@jlu.edu.cn

胡　亮(通信作者)　男,1968 出生,吉林长春人,教授,主要研究领域为网络安全.　E-mail:hul@jlu.edu.cn