文章编号:1001 - 9081(2010)12 - 3354 - 03

多阶段过滤的 P2P 僵尸网络检测方法

刘 丹,李毅超,胡 跃
(电子科技大学 电子科学技术研究院,成都 611731)
(liudan@uestc.edu.cn)

摘 要:提出基于流分析的 P2P 僵尸网络检测方法。首先基于节点连接分布性和突发性特征过滤掉非 P2P 节点,进而根据 P2P 节点对间连接度和流量的对称度,采用 K 均值聚类以发现各个 P2P 群,最后基于各 P2P 群内节点的流行为相似性检测是否为 P2P 僵尸网络。在局域网环境中的实验表明,该检测方法能够有效识别各种 P2P 僵尸网络,提高了检测效率和精度。

关键词:P2P 网络;僵尸网络;聚类;数据流;恶意行为;检测模型
中图分类号: TP393.02 **文献标志码**:A

P2P-Botnet detection based on multi-stage filtration

LIU Dan, LI Yi-chao, HU Yue
(*Electronic Science and Technology Institute, University of Electronic Science and Technology of China, Chengdu Sichuan* 611731, *China*)

Abstract: A new method for detecting P2P-Botnet, which was based on the analysis of network streams, was presented. Firstly, by using outburst and distributed characteristics of the P2P streams, the P2P nodes could be picked up from the common nodes. Then, based on the communication symmetry and cohesion characteristics of the pairs of nodes in a P2P network, the set of peers in one P2P network could be taken out by using the K-average cluster method. Finally, by contrasting with the common actions of the peers in every P2P network, a P2P-Botnet could be distinguished from the P2P networks. Plenty of experiments have been done in LAN environment and the results verified the efficiency and precision of the proposed method.

Key words: P2P network; Botnet; clustering; data stream; malicious behavior; detection model

0 引言

近年来对 P2P 僵尸网络的研究大多处于分析阶段[1-3],对其检测的研究刚刚起步[4],可分为 P2P 协议识别及僵尸行为检测两个方面。目前提出了很多基于网络流特征的 P2P 协议识别算法。文献[5-6]中提出根据数据块大小识别 P2P 流量的思想及根据 P2P 应用流协议特征的组合识别方案。文献[7]中提出基于连接成功率的识别方法;文献[8]对 P2P 节点的行为在社会层面、功能层面、网络层面进行了细致分析,通过 P2P 节点行为在这三个层面上的特点进行识别。

在僵尸网络检测方面,基本以流协议分析和终端行为监测为主,文献[1]中对 storm 进行了深入分析;文献[9]提出采用挂钩 SSDT 表等方法在主机终端上控制僵尸的思想;文献[10]中通过对网络流按协议分类统计识别 storm;文献[11]提出了通过加入 Overnet 网络并发布大量的键值来延迟僵尸体间通信的思想;文献[12]提出了一种基于逻辑分析基础的 storm 检测方法。

为提高对 P2P 僵尸网络的检测效率,本文提出一种基于网络数据流统计特征的 P2P 僵尸网络检测方法,提出并实现了 P2P 节点识别、P2P 群聚类和 P2P 僵尸网络检测三个核心算法。通过实验验证了该方法的良好特征。

1 基于网络数据流的 P2P 僵尸网络检测模型

1.1 P2P 僵尸网络检测模型

本检测模型如图 1,分为 P2P 特征识别、P2P 节点聚类和僵尸行为识别三个顺序执行的检测环节。P2P 特征识别以单节点网络流为检测源,通过统计节点网络流的分布性和突发性特征,提取出具有 P2P 应用的节点集。节点聚类以 P2P 节点集的网络流特征为检测源,统计 P2P 节点对的通信对称度、通信量和通信频度特征,基于这些特征,采用 K 均值聚类算法,聚类出各个 P2P 群。僵尸行为识别通过对同一 P2P 群中各节点的流行为相似度的统计,分离出普通 P2P 群和 P2P 僵尸网络。

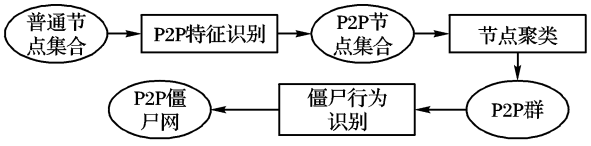

图 1 检测模型

1.2 P2P 节点特征识别

1.2.1 P2P 节点特征

P2P 节点通常会在短时间内发起大量连接,具有连接突发性特征。同时,这类节点产生的连接具有明显的分布性,其连接的目的子网数和普通节点比起来要高很多。本文用 S_a 描述节点连接突发度,S_d 描述节点连接分布度。

在一个 ΔT 的统计周期中包含 m 个采样点。对节点 S,在每个采样点上统计其网络连接数,形成连接数特征集 $\{N_{s1}, N_{s2}, \cdots, N_{sm}\}$,$N_{si}$ 表示第 i 个采样点上节点 S 的网络连接数。

在一个统计周期中,取平均连接数 $\overline{N_s} = \frac{1}{m}\sum_{k=1}^{m} N_{sk}$,第 i 个采样点的连接突发度定义为:$S_{ai} = N_{si} / \overline{N_s}$。

收稿日期:2010-05-15;**修回日期**:2010-07-25。 **基金项目**:中央高校基本科研业务费专项资金资助项目(ZYGX2009J090)。

作者简介:刘丹(1969-),男,四川成都人,副教授,博士,主要研究方向:网络安全、分布式计算; 李毅超(1969-),男,四川成都人,教授,主要研究方向:计算机网络、网络安全; 胡跃(1987-),男,四川成都人,硕士研究生,主要研究方向:网络安全。

取 $S_{a\max} = \max(S_{a1}, S_{a2}, \cdots, S_{am})$, $S^x_{a\max} = S_{a\max} \times (1 - x\%)$, $S^y_{a\min} = S_{a\max} \times y\%$;其中 x, y 为常量。

令 U_x 为所有符合 $\{S_{ai} \mid S_{ai} > S^x_{a\max}\}$ 的采样点集合域,N_{Ux} 为 U_x 包含的采样点数。L_y 为所有符合 $\{S_{ai} \mid S_{ai} < S^y_{a\min}\}$ 的采样点集合域,N_{Ly} 为 L_y 包含的采样点数。

定义 1　节点 S 的连接突发度 S_a。$S_a = N_{Ly}/N_{Ux}$, S_a 值越大,则突发度越高。

定义 2　节点 S 的连接分布度 S_d。$S_d = J_n/J_s$, J_s 为 S 在采样域 U_x 内连接的目的节点数,J_n 为 S 在采样域 U_x 内连接的目的子网数。S_d 越大,则节点连接的分布性越强。

1.2.2　P2P 节点识别算法

根据以上定义,P2P 节点识别算法如下:

1) 计算目标节点 S 的 P2P 特征 S_a 和 S_d;

2) 若 $S_a > C_{Sa}$ 且 $S_d > C_{Sd}$,则判定节点 S 为 P2P 节点,否则为非 P2P 节点。

其中 C_{Sa} 和 C_{Sd} 为 P2P 节点的突发性和分布性特征门限阈值参数,其取值特征来自于对已知 P2P 节点的突发性和分布性统计特征。

1.3　P2P 聚类

1.3.1　定义

定义 3　节点对连接度 Y_{ij}。$Y_{ij} = \lambda_1 \times (N_{ij} + N_{ji}) + \lambda_2 \times k_c$,其中:$N_{ij}$ 是节点 i 在统计周期内发送给节点 j 的数据包数量;K_c 表示节点 i 和节点 j 之间在统计周期内连接的次数;λ_1 和 λ_2 为常量。该连接度既反映了连接的频率,又反映了交换的数据量。其值越大,则表示节点间的连接度越高。

定义 4　节点对通信对称度 X_{ij}。$X_{ij} = \dfrac{N_{ij} + N_{ji}}{\mid N_{ij} - N_{ji} \mid}$,代表节点 i 和节点 j 之间单位时间交换信息的对等程度。其值越大,对称度越高。

定义 5　节点对距离 D_{ij}。$D_{ij} = 1/Y_{ij}$,其值反映了节点间的连接距离大小,距离越小,连接度越高。

1.3.2　P2P 聚类算法

属同一个 P2P 网络的节点在较长时间内有稳定的信息交互,且交互具有高度对称性。基于对称度和连接距离的 P2P 聚类算法基本思想是,首先判断节点对的对称度是否符合 P2P 群节点对的基本要求,在满足对称性要求的条件下,采用 K 均值聚类算法进行 P2P 群聚类。算法如下。

1) 根据 P2P 节点识别算法,记录下所有在 U_x 采集域内连接的源目节点对,形成节点对集合:$\Phi_x = \{(i,j) \mid$ 连接 l_{ij} 在 U_x 采集域内存在$\}$。

2) 用 $N_{\Phi x}$ 表示 Φ_x 的节点对数量:

```
for(k = 0;k < N_Φx;k ++)
{
  对于每个连接 l_k(i,j) ∈ Φ_x,在全采集时间域内统计节点对 k(i,
  j) 的对称度 X_k(i,j) 和连接度 Y_k(i,j);
  if(X_k(i,j) > δ),以节点对距离 D_ij 为聚类属性,采用 K 均值聚类
  算法(见 1.3.3 节),对各节点进行聚类
}
```

δ 基于常用 P2P 群对称度的最低统计值选取。

1.3.3　K 均值聚类算法

类别数 c 的确定　把节点间距离 D_{ij} 看做分类数 c 函数,根据曲线 $\ln D_c$ 确定最佳分类数的间隙统计法[13],有:

$$Gap(c) = E(\ln D_c) - \ln D_c$$

其中:E 为某种分布下的期望。最佳聚类数 c 为满足不等式 $Gap(c) > Gap(c+1) - s(c+1)$ 的最小值。其中 s 为和分布有关的标准误差。具体算法如下。

1) 选取 c 个类的初始中心节点,$O^1_1, O^1_2, O^1_3, \cdots, O^1_c$,令 $k = 1$。

2) 在 k 次迭代中,将所有样本节点 x 依次归类到 c 个类别中的某一类,归类方法为:

若 $D_{xo^k_j} = \min\limits_{1<i<c} D_{xo^k_i}$,则 $x \in S^k_j$。其中 S^k_j 是以 O^k_j 为中心的节点聚类集,$j = 1, 2, \cdots, c$。

3) 用步骤 2) 得到的类 S^k_j,更新第 j 类的中心 O^{k+1}_j,使得 $\sum\limits_{j=1}^{c} \sum\limits_{x \in s^k_j} D^2_{xo^{k+1}_j}$ 达到最小。

4) 对于所有 $j = 1, 2, \cdots, c$,如果 $O^{k+1}_j = O^k_j$,则迭代结束;否则 $k = k + 1$,转到步骤 2) 继续执行。

1.4　P2P 僵尸网络检测

同一僵尸网络中僵尸体之间的行为具有较高相似性,本检测算法基于网络流提取僵尸体的相似性来检测 Botnet。僵尸体的典型恶意行为主要包括扫描行为、DDoS 攻击、发送垃圾邮件、二进制文件下载、exploit 等。对于扫描行为,其共性特征表现为对同一端口扫描;对于垃圾邮件行为,共性特征表现为出现有大量相同地址的 SMTP 报文行为;对于文件下载,共性特征表现为下载相同文件;对 exploit 行为,其共性特征表现为出现发送相同 exploit 的行为;对 DDoS 攻击,共性特征表现为有限时间内向相同地址发起大量连接。

设 P2P 群节点集合 $S_p = \{S_1, S_2, \cdots, S_i\}$。取其中一个子集 Ω_N 为分析样本。定义扫描行为 A_c、垃圾邮件行为 A_s、二进制文件下载 A_d、exploit 行为 A_e、DDoS 攻击行为 A_o。对于每种行为,比较 Ω_N 中各节点的行为相似度。

针对僵尸恶意行为五元组 $\{A_c, A_s, A_d, A_e, A_o\}$,定义相似度为 $\{R_c, R_s, R_d, R_e, R_o\}$。$N_{\Omega_N}$ 为分析样本数。其中,$R_c = C_{\Omega N}/N_{\Omega_N}$, $C_{\Omega N}$ 为具有相同扫描行为的节点数。$R_s = S_{\Omega N}/N_{\Omega_N}$, $S_{\Omega N}$ 为具有相同垃圾邮件行为的节点数。$R_d = D_{\Omega N}/N_{\Omega_N}$, $D_{\Omega N}$ 为具有相同下载行为的节点数。$R_e = E_{\Omega N}/N_{\Omega_N}$, $E_{\Omega N}$ 为具有相同 exploit 行为的节点数。$R_o = O_{\Omega N}/N_{\Omega_N}$, $O_{\Omega N}$ 为具有相同 DDoS 攻击行为的节点数。

定义 6　样本集 Ω_N 的流行为相似度 R, $R = R_c + R_s + R_d + R_e + R_o$。

基于以上定义,P2P 僵尸网络识别算法如下:

1) 计算 Ω_N 的相似度 R;

2) 若 $R > C_R$,则判定节点集 S_p 为 P2P 僵尸网络。

C_R 取值来自于对已知 P2P 僵尸网络节点的统计特征。

2　仿真实验

2.1　P2P 流节点识别

为验证该算法和原型系统的有效性,在局域网接入处采集网络流分析。在测试的节点集中选择性地部署了 Bittorrent/Emule/Kazaa 三种 P2P 协议的应用,并对该环境的网络流进行了统计分析。实验中,被测网络节点总数为 100,其中部署了 P2P 应用的节点数为 30。选取 $\Delta T = 60$ m, $x = 5$, $y = 40$,如图 2 所示节点以 S_a 和 S_d 为坐标的分布。

从图 2 中可以看出,基于节点的 S_a 和 S_d,能够有效分离出具有 P2P 应用的节点。从图 2 中可以得到,$C_{Sa} = 5$, $C_{Sd} = 0.6$

是本实验环境中恰当的突发性和分布性检测阈值。

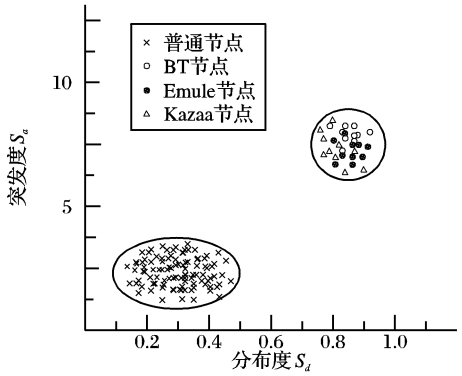

图2 节点分布示意图

2.2 P2P节点聚类

在测试节点集中选择性地部署了Bittorrent/Emule/Kazaa三种P2P协议的应用,采用P2P聚类算法进行聚类效果如表1。实验中,被测网络节点总数为100,其中部署了Bittorrent/Emule/Kazaa的P2P应用节点各30个,选取$\lambda_1 = 0.2$,$\lambda_2 = 0.8$。可见,基于节点间连接度的k聚类算法,在聚类时间达到一定程度(1 d)时,具有很高的聚类率,能有效地识别出不同的P2P网络。

表1 基于连接度的聚类实验结果

聚类时间	协议类型	节点数	聚类节点数	聚类率/%
1 h	Bittorrent	30	16	53
	Emule	30	18	60
	Kazaa	30	12	40
1 d	Bittorrent	30	24	80
	Emule	30	30	100
	Kazaa	30	26	87
10 d	Bittorrent	30	29	97
	Emule	30	30	100
	Kazaa	30	30	100
30 d	Bittorrent	30	30	100
	Emule	30	30	100
	Kazaa	30	30	100

2.3 P2P僵尸网络识别

僵尸网络恶意行为一天中通常会产生若干次。在测试的节点集中选择性地部署了Storm和Nugache、Slapper三类P2P僵尸网络,基于网络流共性的P2P僵尸网络检测算法检测结果如表2(统计时间为1 d的数据分析范围)。

表2 基于流行为相似度的P2P僵尸网络识别实验结果

样本类型	R_c	R_s	R_d	R_e	R_o	R
无bot	0	0	0.07	0	0	0.07
Storm	0.1	0.8	0.40	0	0.9	2.20
Nugache	0.9	0	0.60	0	0.3	1.80
Slapper	0.9	0	0.50	0	0.4	1.80

可见,Storm僵尸网络的主要特征是会发送大量的垃圾邮件,而Slapper/Nugache有大量的扫描行为。通过这种基于节点流行为相似度的关联分析,能够准确检测出一个P2P网络是否感染了僵尸体。

3 结语

本文提出了基于P2P节点识别算法、P2P群聚类算法和基于流行为相似性的僵尸体识别算法为核心的P2P僵尸网络络识别模型。模型以网络数据流为分析源,以节点发起网络流的突发性和分布性为检测特征值,从网络数据流中抽取出P2P流;以同一P2P群内节点的连接性特征为K均值聚类算法的聚类属性,从P2P流中聚类出各P2P群;针对同一P2P群内各节点,统计它们发出流的恶意行为相似性,作为P2P僵尸网络检测的特征值。不同于传统的检测算法和模型,本模型的各检测特征来源于对网络流的统计特征,能有效检测出未知协议和加密协议的P2P僵尸网络络。在局域网环境对本检测模型进行了验证实验,表明基于P2P流的分布性和突发性特征能有效地从网络流量中提取出P2P流;基于连接性特征的P2P聚类在较长的统计周期内能获得良好的聚类效果;通过节点群恶意行为相似度特征来检测P2P僵尸网络,能有效分辨出僵尸网络和非僵尸网络。

参考文献:

[1] GRIZZARD J B, SHARMA V, NUNNERY C, *et al*. Peer-to-peer botnets: Overview and case study [DB/OL]. [2007-04-01]. http://usenix.net/events/hotbots07/tech/full_papers/grizzard/grizzard.pdf.

[2] STEWART J. Sinit P2P trojan analysis [EB/OL]. [2003-12-08]. http://www.secureworks.com/research/threats/sinit/.

[3] LEMOS R. Bot software looks to improve peerage [EB/OL]. [2006-05-02]. http://www.securityfocus.com/news/11390.

[4] KANG JIAN, ZHANG JUN-YAO, LI QIANG, *et al*. Detecting new P2P botnet with multi-chart CUSUM [C]// International Conference on Networks Security, Wireless Communications and Trusted Computing. Wuhan, Hubei: [s.n.], 2009: 688-691.

[5] KARAGIANNIS T, BROIDO A, BROWNLEE N, *et al*. File-sharing in the Internet: A characterization of P2P traffic in the backbone [R]. Riverside, USA: University of California, 2003.

[6] KARAGIANNIS T, BROIDO A, FALOUTSOS M. Transport layer identification of P2P traffic [C]// Proceedings of International Measurement Conference. Sicily, Italy: [s.n.], 2004: 121-134.

[7] ZHOU L, LI Z. P2P traffic identification by TCP flow analysis [C]// Proceedings of International Workshop on Networking Architecture and Storages. Shenyang, China: [s.n.], 2006: 47-50.

[8] KARAGIANNIS T, PAPAGIANNAKI K. BLINC: Multilevel traffic classification in the dark [J]. ACM SIGCOMM Computer Communication Review, 2005, 35(4): 229-240.

[9] NUMMIPURO A. Detecting P2P-controlled bots on the host [C]// Seminar on Network Security. Espoo, Helsinki: [s.n.], 2007: 151-156.

[10] STEGGINK M, IDZIEJCZAK I. Detection of peer-to-peer botnets [M]. Amsterdam, Netherlands: [s.n.], 2008.

[11] HOLZ T, STEINER M. Measurements and mitigation of peer-to-peer-based botnets: A case study on storm worm [C]// Proceedings of the 1st USENIX Workshop on Large-Scale Exploits and Emergent Threats. San Francisco: [s.n.], 2008: 33-39.

[12] PORRAS P, SAIDI H, YEGNESWARAN V. A multi-perspective analysis of the storm (peacomm) worm [EB/OL]. [2007-11-12]. http://www.cyber-ta.org/pubs/StormWorm.

[13] TIBSHIRANI R, WALTHER G, HASTIE T. Estimating the number of cluster in a dataset via the gap statistic [J]. Journal of the Royal Statistics Society, Series B, 2001, 32(2): 411-423.