第41卷 第4期 西 安 交 通 大 学 学 报 Vol.41 №4
2007年4月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Apr. 2007

网络流的层次化分析及其在蠕虫早期检测中的应用

蔡忠闽[1,2]，秦涛[1,2]
(1.西安交通大学智能网络与网络安全教育部重点实验室，710049，西安；
2.西安交通大学机械制造系统工程国家重点实验室，710049，西安)

摘要：在NetFlow的基础上定义了2种新的网络流：主机层网络流和区域层网络流，并提出了相应的流量特征刻画方法.该定义从主机层和网络层对网络流量进行抽象，实现了底层网络流信息的合并，在不损失关键流量特征的前提下，可大大缩减网络流记录的数量.采用指数加权移动平均控制图模型对所提取的网络流特征进行预测和估计，并结合网络蠕虫的传播特征给出了一种网络蠕虫的早期检测方法.实验结果表明，所提检测方法可以准确地检测到蠕虫在局域网内的早期传播，进而为控制蠕虫传播赢得了宝贵的时间.

关键词：网络安全；网络流；指数加权移动平均；网络蠕虫
中图分类号：TP393 **文献标识码**：A **文章编号**：0253-987X(2007)04-0393-05

Stratified Analysis of Network Flows and Its Application to Early Detection of Internet Worms

Cai Zhongmin[1,2], Qin Tao[1,2]
(1. MOE Key Laboratory of Intelligent Network and Network Security, Xi′an Jiaotong University, Xi′an 710049, China;
2. State Key Laboratory of Manufacturing Systems Engineering, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: Based on NetFlow, the definition of two net-flows, Host-Net-Flow and Region-Net-Flow, as well as a depictive method for corresponding network traffic characteristics are proposed. The netflows are abstracted from the host level and network level by the definition. The incorporation of netflow information in bottom level is implemented, and the number of netflow records can significantly be reduced without losing key network traffic characteristics. EWMA (exponential weighted moving average) control chart models are employed to forecast and estimate the distilled network traffic features, then an early net-worm propagation detection method is presented considering the features of network worm propagation. The experimental results show that the early worm propagation in local networks can be detected both quickly and accurately by the proposed method, thereby worm propagation can be controlled as soon as possible.

Keywords: network security; net flow; exponential weighted moving average; net-worm

在信息化大潮的推动下，计算机网络在人们的生活和工作中扮演着日益重要的角色.但是，Stuart Staniford等人研究表明，精心编写的蠕虫能够在几分钟内传遍因特网，致使网络带宽耗尽而瘫痪，严重影响人们的正常生活[1].对网络流量进行有效监控，及时发现并控制包括网络蠕虫在内的网络异常行为，成为保证计算机网络正常运行的关键所在.为了适应网络流量监控的需要，思科公司提出了NetFlow的概念[2]，它可以刻画基于连接的网络行为，也可以较好地检测网络的异常行为[3-4]，但是所刻画

收稿日期：2006-07-31. 作者简介：蔡忠闽(1976～)，男，博士，讲师. 基金项目：国家高技术研究发展计划资助项目(2001AA142060)；国家自然科学基金资助项目(60243001).

的网络流量粒度过于细，在很短的时间间隔内会出现数万条乃至数十万条的网络流记录，海量数据给网络流的特征提取造成了极大的困难.

本文在 NetFlow 的基础上定义了 2 种新的网络流：主机层网络流和区域层网络流，并提出了相应的流量特征刻画方法以及网络蠕虫的早期检测方法. 实验结果表明，所提检测方法可有效、快速地检测传播的蠕虫.

1 主机层网络流量的刻画

在分析网络行为时，通常以联网主机作为基本的考察对象，因此有必要在连接层 NetFlow 的基础上给出准确的主机层网络流的描述.

1.1 主机层网络流

定义 1 在一定的时间间隔内，所有具有相同的源地址、目的地址以及目的端口的数据包称为主机层网络流.

通信时，由于网络源端口是根据本机资源随机打开的，所以去除源端口这一要素后并不影响网络流的特征，却在一定的程度上减少了网络流的数目，这有助于刻画主机的网络通信行为.

1.2 主机层网络流特征提取

在主机层网络流定义的基础上，可用网络流强度和网络流发散度这 2 个特征量来刻画主机层网络流的行为特征.

定义 2 网络流中数据包个数与数据包总大小的比值称为网络流强度 X_i. 该强度刻画了一个主机层网络流中数据包的平均大小.

定义 3 一定的时间间隔内某个网络地址所拥有的最大的流数称为网络流发散度. 设某时间窗口内统计到的活动主机所拥有的网络流个数为 $F(i,j,k)$，其中 i 为时间窗口标志，j 为主机地址标志，k 为主机在此时间窗口内所拥有的网络流数目，则该时间窗口内的网络流发散度

$$Z_i = \max_{j,k}(F(i,j,k)) \tag{1}$$

它适用于描述主机通信活动的集中程度.

2 区域层网络流量的刻画

在刻画大规模的网络流量特性时，通常更关注于区域之间的数据交互行为，因为它能够直接反映出区域之间的通信关系.

2.1 区域层网络流

称具有不同子网号的网络为区域，称统计时间间隔内产生通信关系的区域数目为网络的连接度，其可以刻画一个网络与外部不同网络区域之间的连接情况.

定义 4 对一定时间间隔内的所有网络数据包，采用子网掩码对其源地址和目的地址进行处理，处理后所有具有相同源地址网络号和目的地址网络号的数据包属于一个区域层网络流.

两条网络流经处理后属于同一条区域流，其参数如表 1 所示.

2.2 区域层网络流特征提取方法

表 1 属于同一条区域流的两条网络流的参数

	地址	目的地址	协议	源端口号	目的端口号	数据包数目	字节数
主机源	219.229.118.111	202.117.15.55	TCP	10045	1998	17	1 138
	219.144.27.146	202.117.15.125	TCP	1821	17824	3	144
网络源	219.0.0.0	202.117.15.0	TCP		20	1 282	

设某一时间间隔的区域流为 $f_{\text{sector}}(s,b,p)$，其中 s 为原网络号(区域)标识，b 为流字节数，p 为流的数据包数目，则有如下定义.

定义 5 网络的区域连接度

$$d = \sum_{s_i} \frac{f_{\text{sector}}(s_i,b,p)}{s_i} \tag{2}$$

定义 6 网络的区域数据占有率

$$P(i) = \frac{f_{\text{sector}}(s,b,i)}{\sum_{p} f_{\text{sector}}(s,b,p)} \tag{3}$$

在正常的网络行为中，一个网络区域的连接度并不高，而且绝大部分的网络总流量却被少数的外联区域所占有，它们主导着网络流量的特性，称这部分网络区域为重区域，余下的网络区域称为轻区域. 采用这种网络流的划分方法不但能有效缩减巨大的网络流数量，而且使网络流量能够按照逻辑结构来划分、监控.

3 基于指数加权移动平均控制图模型的蠕虫早期检测算法

根据网络蠕虫的扫描传播规则[5]知，当蠕虫传

播时，网络中的流强度统计结果将明显增大，网络流的发散程度也将急剧增加. G. Box 等人的研究表明，指数加权移动平均(EWMA)控制图在各种条件下都能表现出优秀的预测性能，可以准确地预测和估计出随机过程中的统计特性的变化[6]，而且在其预测置信区间可有效刻画出行为的变化范围. 网络异常行为的出现意味着网络统计特性突变，观察到的行为将会超出 EWMA 的预测置信区间，据此设计合适的 EWMA 控制图模型，是实现网络蠕虫早期检测的关键.

3.1　EWMA 控制图算法

根据文献[6]，序列 $X(n)$的 EWMA 控制图表达式为

$$E_i = (1-\lambda)X_{i-1} + \lambda E_{i-1} \tag{4}$$

式中：λ 是遗忘因子；E_i是预测值. λ 的大小决定着预测模型对特征变化的响应速度. 设预测误差为$e_i = X_i - E_{i-1}$，根据该误差可以得到 EWMA 控制图的上下控制界

$$\left.\begin{aligned} C_x^{\mathrm{L}}(i) &= E_{i-1} - L\sigma_e(i-1) \\ C_x^{\mathrm{U}}(i) &= E_{i-1} + L\sigma_e(i-1) \end{aligned}\right\} \tag{5}$$

而

$$\sigma_e^2(i) = \lambda_1 e(i)^2 + (1-\lambda_1)\sigma_e^2(i-1) \tag{6}$$

其中 $0<\lambda_1<1$. 可采用 EWMA 控制图来预测估计方差的大小，这种方法能够平滑数据流的方差，通常是将式(5)中的 $L=1.96$ 保持 3%的误差.

3.2　网络蠕虫早期检测算法

设 X_i 代表网络流强度，Z_i 代表网络流发散度. 若对所观测到的 X_i 和 Z_i 应用 EWMA 控制图算法，且 X_i、Z_i 的值落在 EWMA 置信区间之外，可记作一次异常. 在检测窗口 T 内，K_1、K_2 分别代表检测到 X_i、Z_i 异常的次数，则生成的预警信号

$$A(T+i) = \begin{cases} A(T+i-1) + 2^{K_1}, & X_i > C_x^{\mathrm{U}} \text{ and } Z_i > C_z^{\mathrm{U}} \\ A(T+i-1), & X_i \in [C_x^{\mathrm{L}} C_x^{\mathrm{U}}] \quad \text{or } Z_i \in [C_x^{\mathrm{L}} C_x^{\mathrm{U}}] \\ A(T+i-1) - 2^{K_2}, & X_i < C_x^{\mathrm{L}} \text{ and } Z_i < C_z^{\mathrm{L}} \end{cases} \tag{7}$$

式中：$A(T)=0$；K_1、K_2 的初始值为 0.

即使网络正常运行，网络行为的随机性也会导致出现超出 EWMA 置信区间的行为，式(7)中的设定充分考虑到了这一点. 采用上述预警条件之后，本文方法可以在具有一定容忍性的同时，还具有准确的预警能力.

4　实验结果

4.1　网络流量采集

在中国教育科研网(CERNET)西北网络中心设置了一个网络流量监控平台，监控平台的拓扑结构如图 1 所示. 本文的数据采集节点位于 CERNET 西北网络主干网出口的路由器之处，所用数据是外网与校园网中的一个 C 网段的所有通信数据，并利用 CoralReef[7]软件对数据进行离线处理.

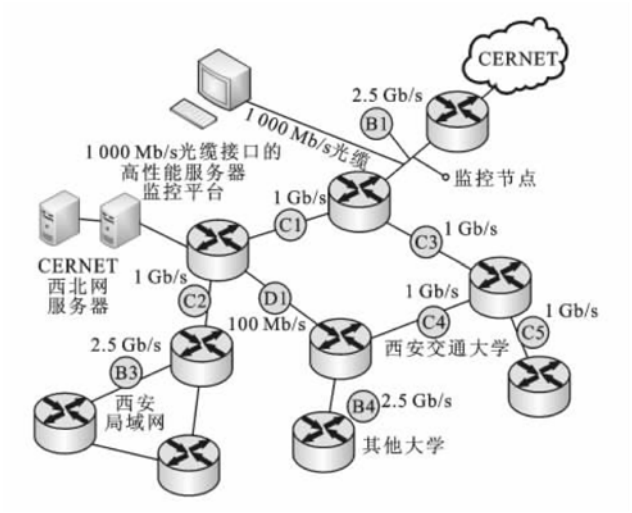

图 1　网络流量监控平台拓扑结构图

4.2　区域网络流分析

按照区域层网络流的定义，可对外网流入一个 C 网段的网络流量进行统计分析. 实验结果表明，网段的连接度主要集中在[40，60]区间，如图 2 所示，而后期的网络连接度出现浮动，这是因为0：00点以后用户对网络的使用有所减少，而且网络行为也不连贯. 每一个时间间隔内不同网络区域的网络数据包数之与此时间间隔内的总数据包数之比值如图 3 所示.

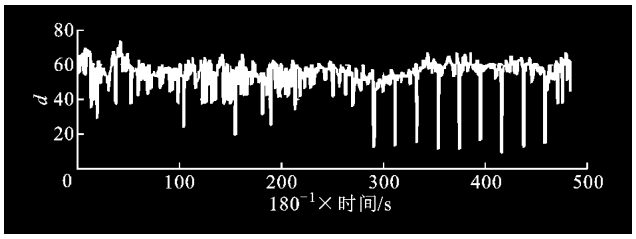

图 2　西安交通大学内部区域连接度

由图 2、图 3 结果可知，虽然网段的连接度在 60 左右，但其中只有 6 个网络区域占据了绝大部分的网络流量，这 6 个网络区域属于重区域，这部分区域的网络流量影响着整个网络流量的统计特性，即自

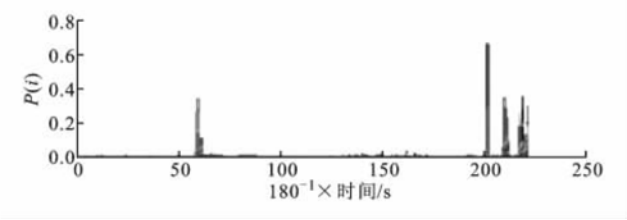

图3 西安交通大学内部区域的数据占有率

相似[8]以及重尾[9]等统计特性.

图4显示了重区域和轻区域的网络流量情况,从中可以发现轻区域中的网络流量较重区域更为稳定,但当网络出现异常,轻区域中网络流量特性会产生显著的变化.因此,在进行网络行为监控时,如果将轻、重区域区分开来,着重关注轻区域的行为变化,则能有效地发现网络异常.

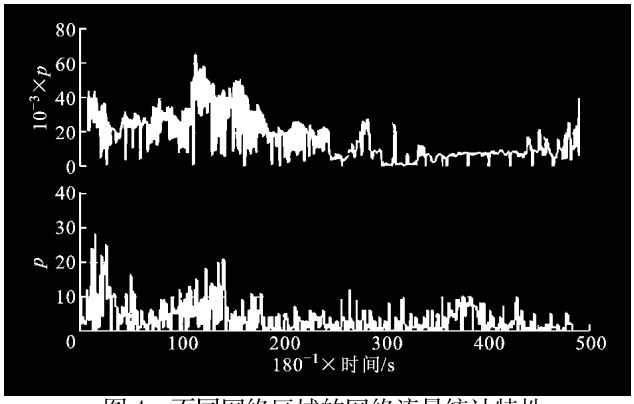

图4 不同网络区域的网络流量统计特性

4.3 主机层网络流量分析及蠕虫早期检测

按照主机层的网络流概念对网络流进行划分,对主机层的网络流强度和发散度进行统计分析,结果如图5所示.

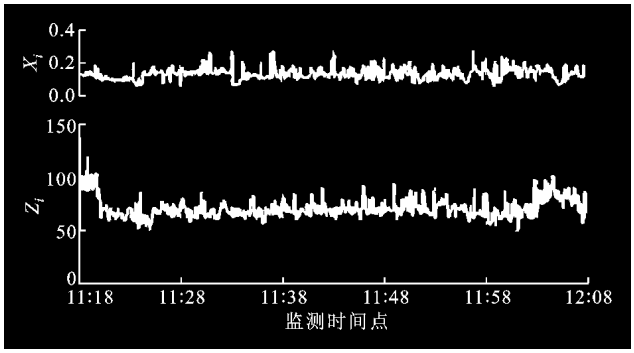

图5 某子网的网络流强度及网络流发散度

从图5可以看出,在网络正常运转的情况下,每个时间间隔内(5 s)活动的主机所拥有的最大流数目 $F_{i,j,k}\approx 65$,相应的 $X_i\approx 0.15$ 并具有一定的平稳性.如果网络中存在恶意行为,这种平稳性将被打破.图6给出了含有异常网络行为的统计结果,异常流量(CodeRed)是根据蠕虫行为特征用软件仿真产生的.通过扫描流量发现,CodeRed感染了C网段的一台主机并于11∶43开始向外传播,11∶48用户将主机关闭.

图6还显示了这台主机被CodeRed感染以后的 X_i 和 Z_i 的变化情况,从中可以看到,当蠕虫开始传播时,Z_i 有增加的趋势,这是因为主机被蠕虫感染以后,蠕虫扫描流量的目的地址便随机生成,而且这种扫描流量将导致大部分网络流只含有一个扫描数据包.图6所刻画的蠕虫行为利用EWMA进行了早期检测,检测结果如图7所示.

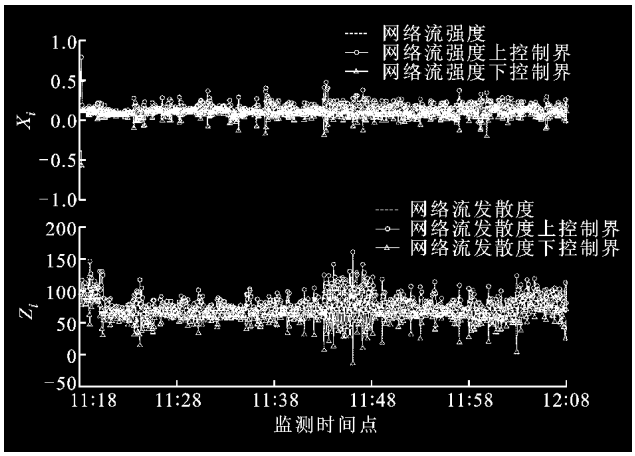

图6 网络流发散度及其强度的下控制界与上控制界

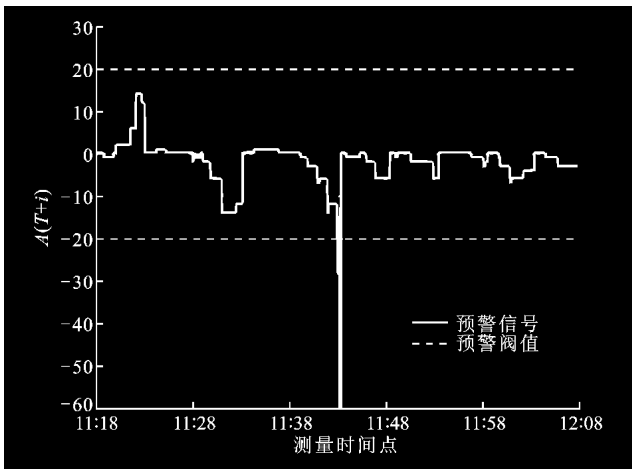

图7 CodeRed蠕虫检测结果

从图7可以看出,CodeRed是在11∶43感染了局域网内的一台主机并开始向外传播.本文算法在11∶44检测到了CodeRed(预警曲线出现了显著下降,超出了预警阈值的正常范围).这说明,在局域内部网络的蠕虫检测上本文算法具有准确的预警能力和快速的反应能力,同时由于它是基于网络流量特征的检测算法,所以不但可以检测到已知的蠕虫,而且对于未知蠕虫也有快速的响应能力.

5 结 论

通过分析网络流量特征，在 NetFlow 的基础上定义了 2 种网络流，并提出了相关网络流特征的刻画方法. 分析结果表明，区域层网络流在降低网络流数据维数的同时，也对网络流量的监控提供了新的视角. 通过主机层网络流的分析结果表明，主机层网络流强度和发散度具有较好的稳定性. 由此，本文进一步提出了基于 EWMA 控制图模型的网络蠕虫早期检测算法. 实验结果表明，该检测算法可以准确地检测到蠕虫在局域网内的早期传播，进而为控制蠕虫传播赢得了宝贵的时间.

参考文献：

[1] Staniford S, Paxson V, Weaver N. How to own the Internet in your spare time[C]// Proceedings of the 11th USENIX Security Symposium. Berkeley, USA: USENIX Assoc., 2002:149-167.

[2] Cisco systems Inc. CISCO NetFlow [EB/OL]. [2006-02-11]. http://www.cisco.com/en/US/products/ps6601/products_white_paper09186a00800a3db9.shtml.

[3] Laffy K C, Braun H W, Olyzos G C. A parameterizable methodology for Internet traffic flow profiling [J]. IEEE Journal of Selected Areas in Communications, 1995, 13(8):1481-1494.

[4] Lakhina A, Crovella M, Diot C. Characterization of network-wide anomalies in traffic flows, BUCS-2004-020 [R]. Boston: Boston University, 2004.

[5] Buchholz F, Daniels T E, Early J P, et al. Digging for worms, fishing for answers [C]// Proceedings of Computer Security Applications Conference. Los Alamitos, USA: IEEE Computer Society, 2002:219-226.

[6] Box G, Lceno A. Statistical control by monitoring and feedback adjustment [M]. New York: John Wiley & Sons, 1999:158-263.

[7] Moore D, Keys K, Koga R, et al. The CoralReef software suite as a tool for system and network administrators [EB/OL]. [2005-01-21]. http://www.caida.org/tools/measurement/coralreef.

[8] Leland W, Taqqu M, Willinger W, et al. On the self-similar nature of Ethernet traffic (extended version) [J]. IEEE/ACM Transactions on Networking, 1994, 2(1):1-15.

[9] Fiorini P M. On modeling concurrent heavy-tailed network traffic sources and its impact upon QoS [C]// 1999 IEEE International Conference on Communications. Piscataway, USA: IEEE, 1999:716-720.

（编辑　苗凌）