doi:10.3969/j.issn.1000-436x.2013.z2.011

基于活跃熵的网络异常流量检测方法

穆祥昆[1,2]，王劲松[1,2]，薛羽丰[1,2]，黄玮[1,2]
(1. 天津理工大学 智能计算及软件新技术天津市重点实验室，天津 300384;
2. 天津理工大学 计算机视觉与系统省部共建教育部重点实验室，天津 300384)

摘 要：提出了一种基于活跃熵的网络异常流量检测新方法，将受监控的目标网络视为一个整体系统，对进出系统的网络数据流所形成的 NetFlow 记录进行分析，分别统计二者的活跃度并计算它们的活跃熵。在进行活跃熵的计算时，根据流量大小选择不同的尺度来降低误报率，从而能更有效地检测网络流量中存在的异常。在实际网络环境下的模拟实验结果表明，与传统检测方案相比，基于活跃熵的网络异常流量检测方法能够更有效地检测出具有随机特征的网络异常流量。

关键词：活跃熵；网络流量；异常流量检测；NetFlow 分析
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2013)Z2-0051-07

Abnormal network traffic detection approach based on alive entropy

MU Xiang-kun[1,2], WANG Jin-song[1,2], XUE Yu-feng[1,2], HUANG Wei[1,2]
(1.Tianjin Key Lab of Intelligent Computing & Novel Software Technology, Tianjin University of Technology, Tianjin 300384, China;
2. Key Laboratory of Computer Vision and System, Ministry of Education, Tianjin University of Technology, Tianjin 300384, China)

Abstract: A novel alive entropy-based detection approach was proposed, which detects the abnormal network traffic based on the values of alive entropies. The alive entropies calculated based on the NetFlow data coming from the network traffic of input and output of a whole system, which is essentially a monitored network. In order to decrease false positive rate of abnormal network traffic, different scales are selected to compute the values of alive entropies in different sizes of network traffic. With the low false positive rate of abnormal network traffic, the abnormal network traffic can be effectively detected. Experiments carried out on a real campus network were used to evaluate the effectiveness of the proposed approach. A comparative study illustrates that the proposed approach may easily detect the abnormal network traffic with random characteristics in comparison with some “conventional” approaches reported in the literatures.
Key words: alive entropy; network traffic; abnormal traffic detection; NetFlow analysis

1 引言

互联网自诞生之日起经过了迅猛的发展，规模与日俱增，人们越来越多地使用网络进行办公、学习与交流。但是，在网络为人们提供更加方便生活的同时，也带来了更多的隐患与威胁。其中，蠕虫、扫描、DDoS 等是产生这些隐患与威胁的主要原因之一。它们所产生的异常流量混杂在正常的网络流量之中，令人难以分辨。因此，如何在纷繁复杂的网络流量中检测出异常的数据流量，已经成为亟待解决的问题。

目前，用熵理论来分析网络流量成为近期的一个研究热点。国内外的研究人士根据熵理论提出了多种异常流量检测方案。Fonseca 等[1]用信息熵的方法发现流量空间上的信息单元也存在长相关特性。Lee 等[2]建议用几种信息理论来分析流量，包括熵、条件熵、相对熵。Feinstein 等[3]用熵来识别 DDoS 攻击，Wagne 等[4]用熵来检测蠕虫攻击。刘衍珩等[5]提出基于活跃熵的 DoS 攻击检测模型。除了基于

收稿日期：2013-09-04
基金项目：国家自然科学基金资助项目(61272450)；滨海新区科技小巨人成长计划基金资助项目(2011-XJR12005)
Foundation Items: The National Natural Science Foundation of China (61272450); BinHai New District Little Giant of Science and Technology Project(2011-XJR12005)

熵的检测方法之外，基于数据流的检测也受到了广泛关注。田杨[6]提出基于先验触发的改进型 BP 神经网络算法。王志[7]提出基于 NetFlow 的流量统计分析。在上述研究中，虽然基于熵理论的研究较多，但是大多是对网络数据分组的检测，而并非流层面信息的检测，效果受到了很大限制。而相关基于流量特征的检测方案也是主要依靠统计、聚类、挖掘等方法展开，计算时间相对较长。

本文基于活跃熵[5]提出了一种网络异常流量检测方案。分析的对象不再是网络数据分组，而是网络 NetFlow 数据流。由于算法建立在分析高速统计性数据源之上，因此具有高效性和时效性的特点，提升了检测效率，很好地满足了网络数据检测的要求。该方法不仅在实验网络环境下取得很好的检测效果，而且在实际网络的测试中也有不俗的表现。

2 基于活跃熵的网络数据流检测算法

2.1 相关背景

网络流量就是网络上传输的数据量，其记录方式多种多样。本文采用 NetFlow 数据对网络流量进行记录。NetFlow 是由思科提出的一种数据交换方式。它的工作原理是利用标准的交换模式处理数据流的第一个 IP 分组数据，生成 NetFlow 缓存，随后同样的数据基于缓存的信息在同一个数据流中进行传输，不再匹配相关的访问控制等策略。可以借助多种方法对 NetFlow 记录进行研究与分析，挖掘记录背后隐藏的重要信息。

对于流量特征的相关测量[8~11]显示，一个普遍存在的规律是 IP 地址或流遵循齐夫定律（Zipf-type law），也就是一小部分流占总流量的绝大部分。相关文献[12~14]的研究表明， IP 及流长的总体分布呈现明显的重尾分布特征，而恶意网络行为或配置错误所导致的异常流量会对流长实际分布产生影响。

了解上述网络流量的分布规律有助于进行更进一步的分析。目前，分析网络流量的方法很多，本文基于活跃熵的方法对网络流量进行研究。而熵（entropy）这个概念来自于统计热力学，示一个系统内部的混乱程度。系统越是混乱，其熵值就越高；反之，若系统越是有序，那么所对应的熵值也就越低[15]。香农公式引入信息熵的概念以数字来量化样本的离散程度，当样本的值分布最集中时，熵值为 0，这时所有的样本值相同，当样本的分布最分散时，熵值最大，为 log$S$，这时所有样本值都不同。因此，在信息理论中经常用信息熵或 Shannon 熵来称呼熵。

2.2 算法描述

本文所设计的方法需要根据网络流量的变化来调整检测窗口尺度的大小。不妨对选取检测窗口大小的依据做出如下假设。

设长度为 T 的集合中存在 n 个不重复元素 $\{a_1,a_2,\cdots,a_n\}$，各值对应的出现次数为集合 $\{d_1,d_2,\cdots,d_n\}$。求熵公式为

$$H(x)=-\sum_{i=1}^{n}\left(\frac{d_i}{T}\right)\log\left(\frac{d_i}{T}\right) \tag{1}$$

则选取小尺度计算熵值达到极限的次数多于大尺度计算时达到极限的次数。

推导过程如下：

以 T 为单位长度时，可得熵值为

$$H_1(x)=-\sum_{i=1}^{n}\left(\frac{d_i}{T}\right)\log\left(\frac{d_i}{T}\right) \tag{2}$$

以 $T/2$ 为单位长度时，即将原来的集合分为两部分，此时计算可得这两部分熵值分别为

$$H_2(x)=-\sum_{i=1}^{n/2}\left(\frac{d_i}{T/2}\right)\log\left(\frac{d_i}{T/2}\right) \tag{3}$$

$$H_3(x)=-\sum_{i=n/2}^{n}\left(\frac{d_i}{T/2}\right)\log\left(\frac{d_i}{T/2}\right) \tag{4}$$

将 $H_1(x)$变形后可得

$$\begin{aligned}H_1(x)&=-\sum_{i=1}^{n}\left(\frac{d_i}{T}\right)\log\left(\frac{d_i}{T}\right)\\&=-\sum_{i=1}^{n}\frac{1}{2}\left(\frac{d_i}{T/2}\right)\left(\log d_i-\log\frac{T}{2}+1\right)\\&=\frac{1}{2}[H_2(x)+H_3(x)]-1\end{aligned} \tag{5}$$

若以 $T/2$ 为单位长度进行计算时，熵值达到极限的次数为 $2y$，则根据式(5)可得、以 T 为单位长度计算时，熵值达到极限的次数必小于 y。

从以上分析可知，当选取小区间尺度进行预警时，大区间尺度就不一定预警。故针对不同时段网络特点（流量分布特点）应选取不同的流窗口区间。

在上述证明的基础上，本文提出了基于活跃熵的异常流量检测方法。若将网络整体视为一个系统，其中包含 n 个主机个体，即 $S=\{s_1,s_2,\cdots,s_n\}$，则系统中每台主机的一次动作可记为一个动作单元。当系统内的主机与系统之外主机存在交互时，动作单元

开始计数，访问系统内主机时，v 值加 1，反之 v 值减 1。定义状态集合 $U=\{\mu_1, \mu_2,\cdots,\mu_n\}$为系统在经过 n 次动作后状态序列的集合。从状态集合中可以得到其活跃度集合 $A=\{a_1,a_2,\cdots,a_k\}$，其中，a_i 表示系统经过规定时间或数量等尺度后，状态 μ_i 出现的次数。通过计算状态 μ_i 出现的概率 p_i，可以得到相关各状态的活跃度概率集合 $P=\{p_1,p_2,\cdots,p_k\}$。对以上概率集合应用熵值理论则可得到活跃熵的定义为

$$H_a=-\sum_{i=1}^{k}(p_i)\log(p_i) \tag{6}$$

算法流程如图 1 所示。图中可以清楚地看到，网络数据经过逐步变化与检测判断后，最终反映网络的安全状况。

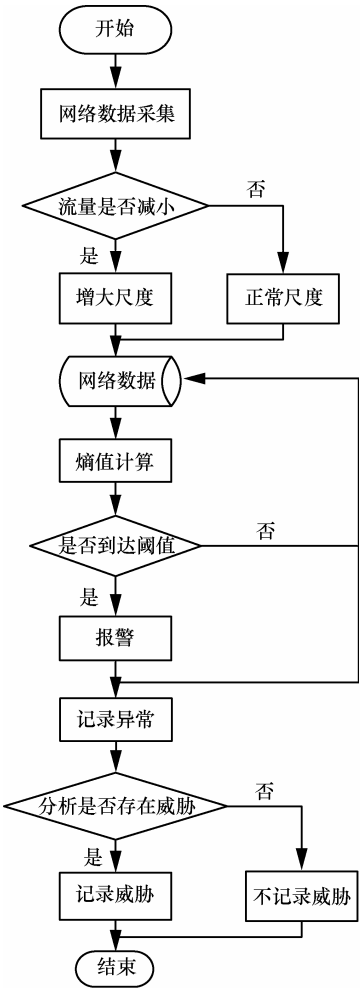


图 1　算法流程

流程描述如下。

算法 1　活跃熵网络异常流量检测

输入：网络数据

输出：报警或正常

Begin

1) 网络数据循环采集，收集网络流量的 NetFlow 信息。

2) 判断流量变化情况，若减小则增大窗口大小。

3) 将 NetFlow 数据构建成待检测网络数据。

4) 依据式(6)进行流量熵值计算。

5) 判断熵值是否达到阈值，若没有则跳转到步骤 3)。

6) 对威胁进行报警，跳转到步骤 3)。

7) 记录报警的异常。

8) 分析异常是否为威胁，若是威胁则记录该威胁。

9) End

3　实验及应用

3.1　算法验证

此部分实验主要介绍基于活跃熵的网络异常流量检测算法与其他检测方案的比较，比较对象为 Snort[16]。测试环境参照曾嘉在基于 NetFlow 的网络异常流量检测[17]一文中的设计，结构如图 2 所示。

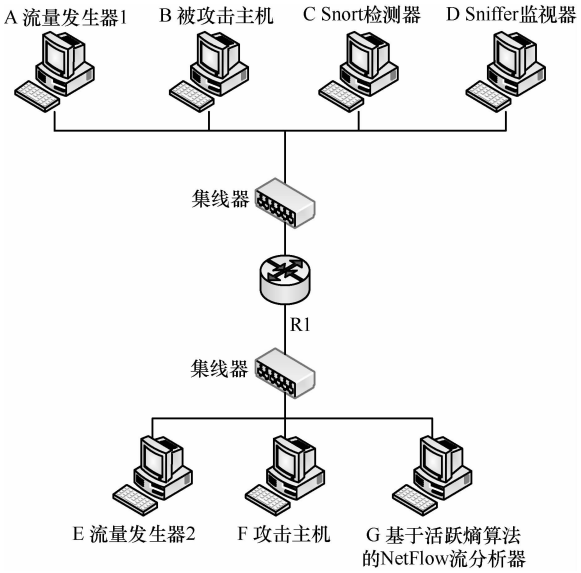


图 2　实验环境

主机 A 和主机 E 代表流量发生器，它们相互通信，负责产生正常网络流量。主机 F 代表攻击主机，装有攻击器，模拟相应的攻击。主机 B 为被攻击主机。主机 C 为装有 Snort 的检测主机，主机 G 为基于本文算法的流量分析器，对来自路由器 R1 的 NetFlow 流量进行分析。主机 D 为装有 Sniffer 的流量记录器，负责验证流量数据。在测试时，主机 F

向主机 B 发起攻击，攻击项及结果如表 1 所示。

表 1　　检测方案效果对比

检测方案名称	一般端口扫描	随机分布端口扫描	DoS	DDoS
Snort[16]	是	否	是	是
本文所提出的活跃熵算法	是	是	否	是

上述对比实验表明，在 DoS/DDoS 这项测试中，基于活跃熵的网络异常流量检测算法表现略低于 Snort，而在随机分布端口扫描这项上，本文算法优于 Snort。其中的原因主要是由于：在基于本文所设计算法的检测器中，主要针对网络流（NetFlow）整体态势进行分析，当出现 DoS 攻击时，实为发生在一个流内部的异常，不会反映到整体流记录层面；当出现 DDoS 攻击时，由于攻击规模骤增，会在流层面上有明显体现，此时 2 种检测方案均可检测到此威胁。而针对流记录层面的异常，如随机分布端口扫描这类异常就会轻松地被检测，此时 Snort 无法检测到这类威胁，本文所设计的算法在这方面效果明显优于 Snort 这类传统的 IDS(intrusion detection system，入侵检测系统)。

在与近期国内其他研究者所做的基于 NetFlow 流数据的异常检测方案[6,18]进行对比后发现：目前国内对于 NetFlow 流的研究一部分[6]主要是基于学习模型的算法，这类算法在检测之前要花费大量时间进行训练。另一部分[18]则是对流记录本身的数据进行分析。在与同类基于活跃熵检测的方法对比后发现，基于数据分组的活跃熵检测对 DoS 攻击具有较好的检测效果[5]，原因与之前分析 Snort 类似。而基于流的活跃熵检测则将检测放在流层面，对威胁更为严重的大规模 DDoS 攻击的检测能力较好。同时，本文算法无需花费大量时间进行特征训练，节约时间，提升了整体流量指标下的检测能力。

在上述实验和分析中可以看到，本文提出的基于活跃熵算法的网络异常流量检测算法更加适合于分析整体流量概念下的网络异常，而且对这类异常检测的效果比较理想。

3.2　一个校园网测试实例

为了验证本文算法的实用性，本文对校园网的不同时间段网络流量应用上述方法进行检测。网络详细参数\规模如下。

网络流量采集自学校教育网中的一台路由器。经过对路由器流量多天的收集后，计算得出以下 NetFlow 流数据流量平均信息：白天（10:00 左右）为 4.6×10^6 条/min，夜间（2:00 左右）为 1.9×10^6 条/min。

如图 3 所示。可以清楚地看到 NetFlow 流数据

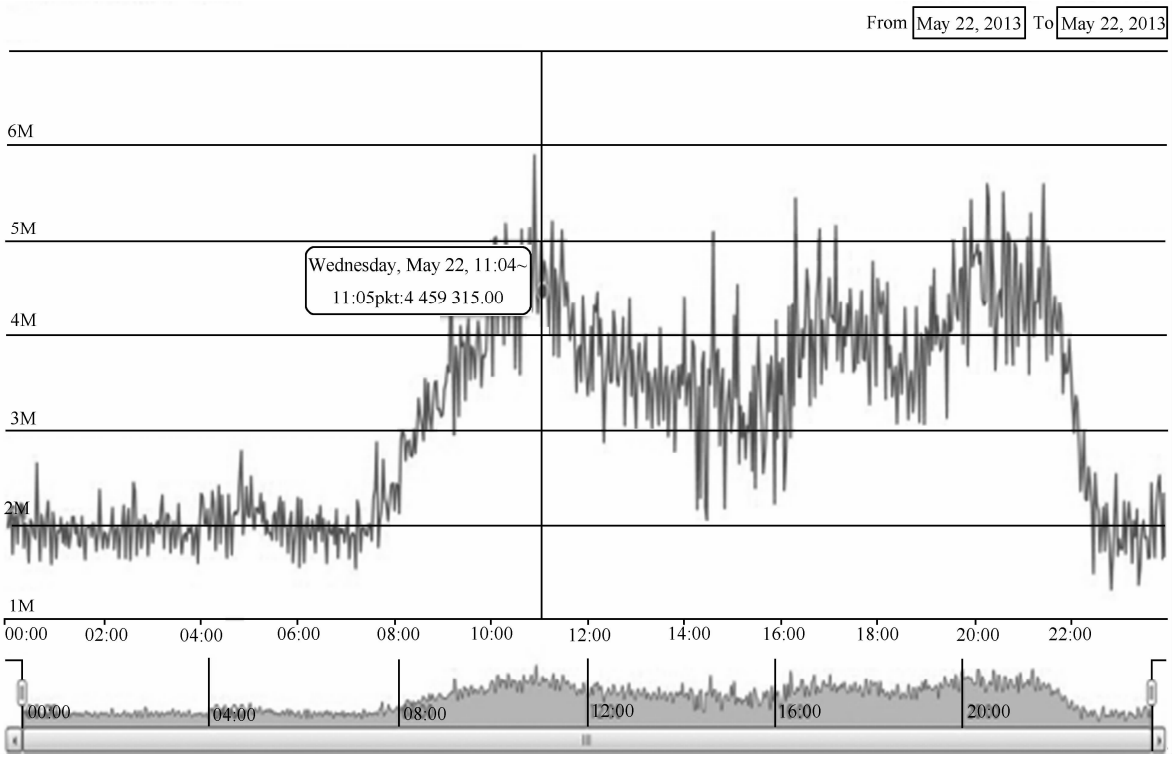


图 3　校园网日流量

流量分布存在白天多，夜间少的特点。

不同时间段流量存在明显的差异变化，流量随时间存在明显变化规律。白天流量密集夜间流量降低。在分析网络数据后发现，异常流量分布也存在一定的规律性。白天，异常流量的增长速度远不及正常流量的增长迅速，因此造成异常流量所占比例的急剧缩小。夜间，正常流量大幅减少，异常流量则有增长趋势，由此造成异常流量所占比例大幅升高。

可见，采取统一的流窗口区间进行熵值计算是存在问题的。过小的流窗口区间会造成报警率增加，误报率也随之增加，且当流窗口区间足够小时，误报率急剧上升。而过大的流窗口区间则会造成检测效率的大幅下降，这一点主要是由于过大的检测窗口包含更多的正常流量，因此造成异常流量被“淹没”，无法被检测。

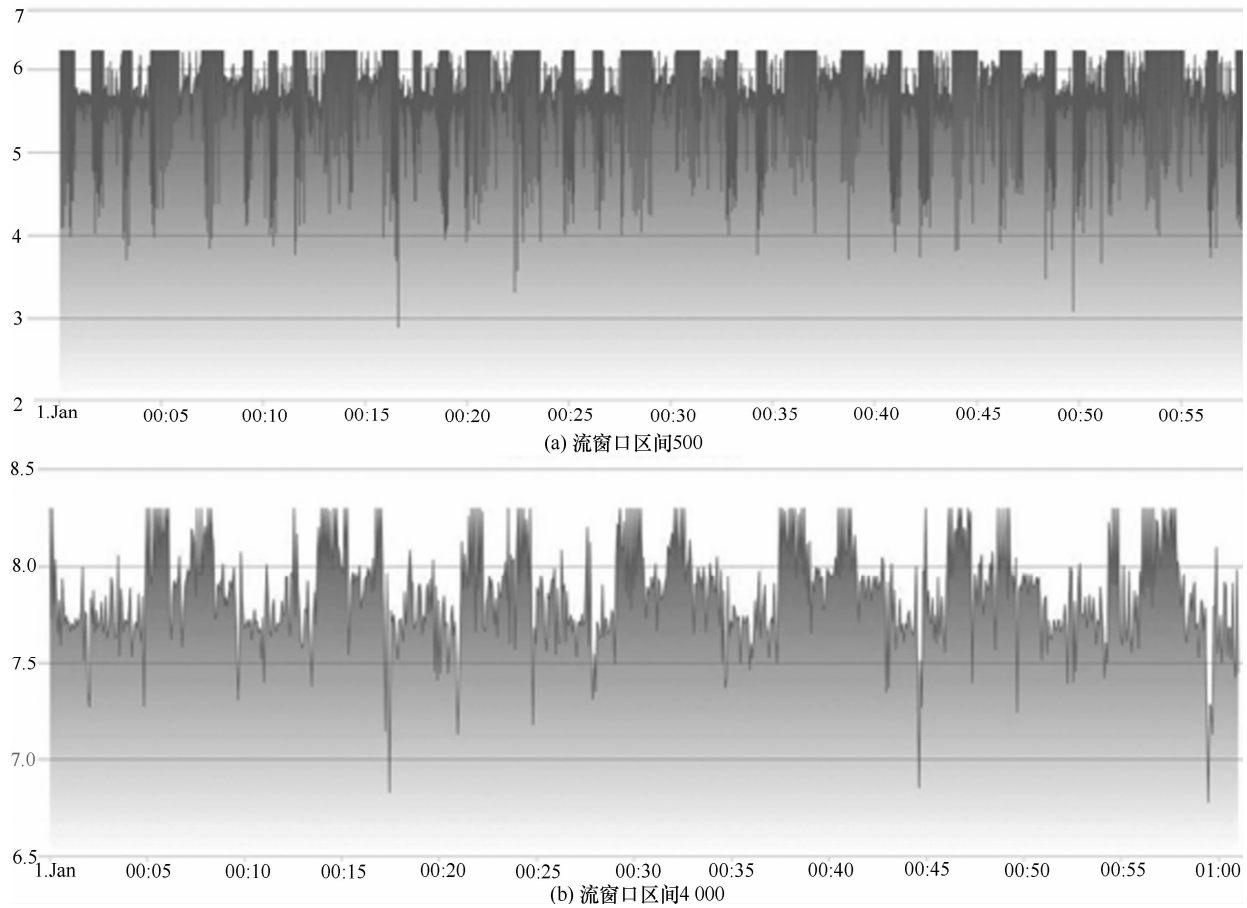
分别计算不同区间尺度下的活跃熵值，理想情况下各区间熵值情况如表 2 所示。

对上述网络流量不同时间区段分别采用不同流窗口区间进行实验测试，如图 4 所示，上方流窗口区间为 500 条，下方为 4 000 条。对照表 2 的各区间熵值极限值，从图 4 中可以明显看到，当选取不同流窗口区间时，达到顶点/低谷的熵值个数（异常点）所呈现的图形存在明显差异。印证了上文提到的流窗口区间选择对检测效果的影响。

表 2　　活跃熵熵值

NetFlow 流数	活跃熵熵值范围	理论正常值
500	0.115~6.215	5.120
1 000	0.180~6.908	5.812
2 000	0.260~7.601	6.503
4 000	0.318~8.294	7.196
6 000	0.301~8.700	7.601

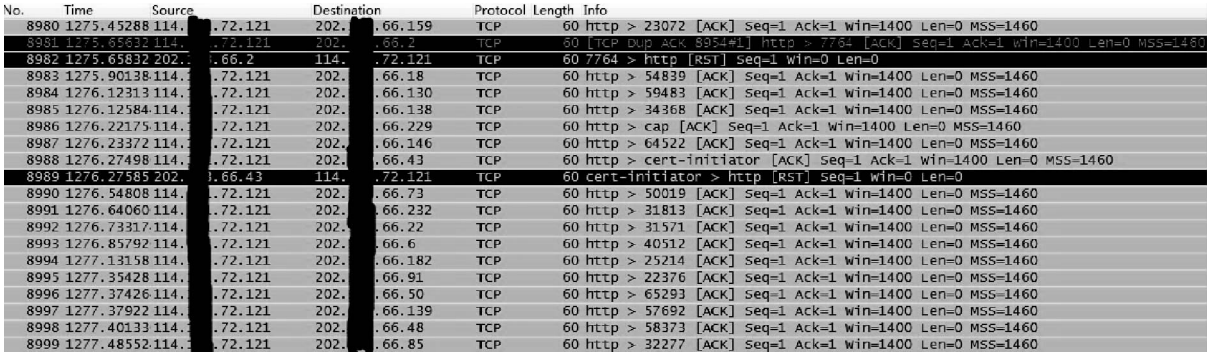
实验测试表明：正常流量 N（normal flow），异常流量 A（abnormal flow），流窗口区间 W（window）之间存在如下关系。



(a) 流窗口区间500

(b) 流窗口区间4 000

图 4　不同尺度检测对比



No.	Time	Source	Destination	Protocol	Length	Info
8980	1275.45288	114.[redacted].72.121	202.[redacted].66.159	TCP	60	http > 23072 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8981	1275.65632	114.[redacted].72.121	202.[redacted].66.2	TCP	60	[TCP Dup ACK 8954#1] http > 7764 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8982	1275.65832	202.[redacted].66.2	114.[redacted].72.121	TCP	60	7764 > http [RST] Seq=1 Win=0 Len=0
8983	1275.90138	114.[redacted].72.121	202.[redacted].66.18	TCP	60	http > 54839 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8984	1276.12313	114.[redacted].72.121	202.[redacted].66.130	TCP	60	http > 59483 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8985	1276.12584	114.[redacted].72.121	202.[redacted].66.138	TCP	60	http > 34368 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8986	1276.22175	114.[redacted].72.121	202.[redacted].66.229	TCP	60	http > cap [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8987	1276.23372	114.[redacted].72.121	202.[redacted].66.146	TCP	60	http > 64522 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8988	1276.27498	114.[redacted].72.121	202.[redacted].66.43	TCP	60	http > cert-initiator [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8989	1276.27585	202.[redacted].66.43	114.[redacted].72.121	TCP	60	cert-initiator > http [RST] Seq=1 Win=0 Len=0
8990	1276.54808	114.[redacted].72.121	202.[redacted].66.73	TCP	60	http > 50019 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8991	1276.64060	114.[redacted].72.121	202.[redacted].66.232	TCP	60	http > 31813 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8992	1276.73317	114.[redacted].72.121	202.[redacted].66.22	TCP	60	http > 31571 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8993	1276.85792	114.[redacted].72.121	202.[redacted].66.6	TCP	60	http > 40512 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8994	1277.13158	114.[redacted].72.121	202.[redacted].66.182	TCP	60	http > 25214 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8995	1277.35428	114.[redacted].72.121	202.[redacted].66.91	TCP	60	http > 22376 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8996	1277.37426	114.[redacted].72.121	202.[redacted].66.50	TCP	60	http > 65293 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8997	1277.37922	114.[redacted].72.121	202.[redacted].66.139	TCP	60	http > 57692 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8998	1277.40133	114.[redacted].72.121	202.[redacted].66.48	TCP	60	http > 58373 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460
8999	1277.48552	114.[redacted].72.121	202.[redacted].66.85	TCP	60	http > 32277 [ACK] Seq=1 Ack=1 Win=1400 Len=0 MSS=1460

图 5 攻击分组信息

$$W \sim \frac{A}{N} \tag{7}$$

流窗口区间大小与异常检测效率随异常流量所占该时段网络的比例有关。当网络中异常流量占总流量较小时，选取小流窗口区间；当网络中异常流量占总流量比例较大时，选取大的流窗口区间。这样可以有效地降低误报率，提升检出率。

图 5 给出了一组攻击分组的真实实例，内容是根据所检测到的异常信息得到的数据分组信息。此异常被本文算法所检测，Snort 未能检出。

该攻击伪装成 HTTP 响应分组，对本校 66 网段进行全网段的攻击。并且此攻击具有随机式分布的特征，经过一段时间后会覆盖整个 66 网段的所有主机及特定的部分端口。

4 结束语

基于活跃熵的网络异常流量检测方法适用于对大流量异常的检测。经实验验证，表明该方法能较好地满足网络数据检测的要求。未来进一步工作将就算法进行改进，对不同网络应用造成的网络流量变化特征进行提取和分析，区别处理。

参考文献：

[1] FONSECA N, CROVELLA M, SALAMATIAN K. Long range mutual information[J]. ACM SIGMETRICS Performance Evaluation Review, 2008, 36(2):32-37.

[2] LEE W, XIANG D. Information-theoretic measures for anomaly detection[A]. Proceedings of IEEE Symposium on Security and Privacy[C]. Oakland, CA, 2001.130-143.

[3] FEINSTEIN L, SCHNACKENBERG D, BALUPARI R, *et al*. Statistical approaches to DDoS attack detection and response[A]. Proceedings of DARPA Information Survivability Conference and Exposition (DISCEX)[C]. Washington DC, USA, 2003. 303-314.

[4] WAGNER A, PLATTNER B. Entropy based worm and anomaly detection in fast ip networks[A]. Proceedings of the 14th IEEE International Workshops Enabling Technologies: Infrastructure Collaborative Enterprise[C]. Washington DC, USA, 2005.172-177.

[5] 刘衍珩,付枫,朱建启等. 基于活跃熵的 DoS 攻击检测模型[J].吉林大学学报(工学版), 2011, 41(4):1059-1064.
LIU Y H, FU F, ZHU J Q, *et al*. DoS detection model base on alive entropy[J]. Journal of Jilin University(Engineering and Technology Edition),2011,41(4):1059-1064.

[6] 田杨. 基于 NetFlow 的异常流量检测研究与实现[D].长沙：国防科学技术大学,2009.
TIAN Y. Anomaly Traffic Detection Research and Implementation Base on NetFlow[D]. Changsha: National University of Defense Technology, 2009.

[7] 王志. 基于 NetFlow 的流量统计分析系统设计与实现[D]. 北京：北京邮电大学，2007.
WANG Z. Design and Implementation of a Network Traffic Measurement and Analysis System Based on NetFlow[D]. Beijing: Beijing University of Posts and Telecommunications, 2009.

[8] FELDMANN A, GREENBERG A, LUND C, *et al*. Deriving traffic demands for operational IP networks: methodology and experience[J]. IEEE/ACM Transactions on Networking, 2001,9(3):265-279.

[9] FANG W, PETERSON L. Inter-AS traffic patterns and their implications[A]. Proceedings of the 4th Global Internet Symposium[C]. 1999. 1859-1868.

[10] BROWNLEE N, CLAFFY K C. Understanding Internet traffic streams: dragonflies and tortoises[J]. Communications Magazine, IEEE, 2002, 40(10): 110-117.

[11] KOHLER E, LI J, PAXSON V, *et al*. Observed structure of addresses in IP traffic[J]. IEEE/ACM Transactions on Networking, 2006,14(6): 1207-1218.

[12] DUFFIELD N, LUND C, THORUP M. Estimating flow distributions from sampled flow statistics[J]. IEEE/ACM Transactions on Networking, 2005,13(5):933-946.

[13] KUMAR A, SUNG M, XU J J, *et al*. Data streaming algorithms for efficient and accurate estimation of flow size distribution[A]. Proceedings of ACM SIGMETRICS[C]. 2004.177-188.

[14] YANG L, MICHAILIDIS G. Sampled based estimation of network traffic flow characteristics[A]. INFOCOM 2007 the 26th IEEE International Conference on Computer Communications[C]. 2007.

1775- 1783.

[15] 刘华文.基于信息熵的特征选择算法的研究[D].长春：吉林大学，2010.
LIU H W. A Study on Feature Selection Algorithms Using Information Entropy[D]. Changchun: Jilin University, 2010.

[16] Snort, the open sou rce network intru sion detection system[EB/OL]. http://www. snort.org.

[17] 曾嘉, 金跃辉, 叶小卫. .基于 NetFlow 的网络异常流量检测[J].微计算机应用, 2007, 28(7):709-713.
CENG J, JIN Y H, YE X W. NetFlow-based anomaly traffic analyzer[J]. Microcomputer Applications, 2007, 28(7):709-713.

[18] 张国祥, 基于 NetFlow 的校园网异常流量检测方法的实现与分析[D]. 呼和浩特: 内蒙古农业大学, 2011.
ZHANG G X. Implementation and Analysis of Campus Network Anomaly Traffic Detection Based on NetFlow[D]. Huhehot: Inner Mongolia Agricultural University,2011.

作者简介：

穆祥昆（1988-），男，回族，天津人，天津理工大学硕士生，主要研究方向为计算机网络、信息安全。

王劲松（1970-），男，天津人，博士，天津理工大学教授，主要研究方向为计算机网络、信息安全、互联网技术。

薛羽丰（1990-），男，广东深圳人，天津理工大学学生，主要研究方向为计算机网络、信息安全。

黄玮（1980-），男，江西抚州人，博士后，天津理工大学讲师，主要研究方向为计算机通信、智能计算、神经模糊系统。