基于多维信息熵值的 DDoS 攻击检测方法

赵小欢[1]，夏靖波[1]，郭威武[2]，杜华桦[3]
(1.空军工程大学信息与导航学院,陕西西安,710077;2.93010 部队,辽宁沈阳,110015;
3.空军通信网络技术管理中心,北京,100843)

摘要　针对互联网中日益严重的分布式拒绝服务攻击行为,提出了一种基于多维信息熵值的 DDoS 攻击检测方法。首先根据 DDoS 攻击的特点,采用条件熵及相异熵构建具有良好区分度的多维攻击检测向量,在此基础上采用滑动窗口的多维无参数 CUSUM 算法放大正常流量与攻击流量的差异来实现 DDoS 攻击的检测。通过实际网络攻击流量及合成攻击流量测试表明:文中提出的算法能够检测到 LLS-DDoS 数据集及合成数据集中的全部攻击,算法对于 DDoS 攻击的响应速度快,能够应用于高速骨干网络中。

关键词　分布式拒绝服务攻击;条件熵;相异熵;多维无参数 CUSUM 算法;滑动窗口
DOI　10.3969/j.issn.1009-3516.2013.03.014
中图分类号　TP393　　**文献标志码**　A　　**文章编号**　1009-3516(2013)03-0058-05

Detection DDoS Attack Based on Multi-Dimensional Entropy

ZHAO Xiao-huan[1], XIA Jing-bo[1], GUO Wei-wu[2], DU Hua-hua[3]
(1. Information and Navigation College, Air Force Engineering University, Xi'an 710077, China; 2. Unit 93010, Shenyang 110015, Liaoning, China; 3. Communication Network and Technology Management Center of Air Force, Beijing 100843, China)

Abstract: In order to detect the increasingly serious distributed denial of service (DDoS) attack on the internet, an algorithm for detecting DDoS attack based on multi-dimensional information entropy is proposed. First of all, according to the property of DDoS attack, the multi-dimensional detecting vector which is capable of distinguishing attack from normal traffic is constructed based on conditional entropy and discrepant entropy. Then the sliding multi-dimensional non-parameter CUSUM algorithm with the capability of amplifying the discrepancy between normal and abnormal network traffic is adopted to detect DDoS attack. The experiments over actual and composite network attack traffic show that the proposed algorithm can detect all the DDoS attacks in both traces. Meantime, the proposed algorithm is capable of detecting DDoS attack quickly and it can be applied in the high backbone network.

Key words: distributed denial of service attack; conditional entropy; discrepant entropy; multi-dimensional non-parameter CUSUM algorithm; sliding window

分布式拒绝服务(Distributed Denial of Service,DDoS)攻击目前已经成为互联网络中最大的安全威胁之一,其通过控制互联网中大量傀儡主机向被攻击目标发送请求报文,消耗被攻击目标的资源,

收稿日期:2013-01-09
基金项目:陕西省自然科学基金资助项目(2012JZ8005)
基金项目:赵小欢(1984—),男,湖北枣阳人,博士生,主要从事网络流量测量、流量异常检测研究.
E-mail:zxhzxh_2012@163.com

阻止被攻击目标为合法用户提供服务。DDoS 的危害范围目前已经扩大到国家的经济、政治、军事、文化等各个重要领域。在近期 2 次较大规模网络战争(2007 年爱沙尼亚网络攻击战和 2008 年俄格冲突网络战)中,进攻方均通过僵尸网络发动 DDoS 攻击达到瘫痪对方信息基础设施的目的,Arbor Networks 在 2010 年首次监测到 DDoS 攻击流量达到 100 Gbps[1],因此,针对 DDoS 攻击的检测与控制方法的研究具有重要意义。

目前已有较多文献对 DDoS 攻击检测方法展开了研究,DDoS 攻击的检测原理大致上分为 2 步,首先从大量网络事件中提取出评价网络攻击的指标参数,然后依据提取出的评价指标建立方法模型进行攻击检测。文献[2]选取可描述网络流量自相似性的 Hurst 参数作为评价指标,通过计算参数的变化来检测 DDoS 攻击,文献[3]依据 TCP 3 次握手时 SYN 包及 FIN/RST 包出现频率的约束关系,通过统计 SYN 与 FIN/RST 频率差异来检测 SYN Flooding 攻击,该方法需要 TCP 双向报文均通过检测点,只能适用于边界路由器,文献[4]指出大型高速网络流量的 IP、端口对应的信息熵值在一定范围内较为稳定,因此选取网络流五元组的样本熵值及源地址数量构成流量特征矩阵,通过支持向量机(SVM)算法实现正常流量和异常流量的分类,文献[5]选取网络流的流特征条件熵和流轮廓偏离度为指标,通过条件随机场(CRF)模型有效融合报文的上下文关系和多特征信息实现正常流量和 DDoS 攻击流量的区分,而文献[6]选择特定 UDP 端口应答包数、同一主机时间窗内连接数等 7 个特征评价指标,组合多个 RBP 神经网络分类器实现正常流量和 DDoS 攻击的分类,并依据尼曼-皮尔逊最小代价策略实现同一目标多个分类结果的有效融合。

可以看出,DDoS 攻击检测指标选取,由流量间的初级信息不断向初级指标间的关联融合发展,而在检测方法的选取上,一些新的机器学习算法也不断被引入 DDoS 攻击检测中。考虑到 DDoS 检测的实时性和准确性要求,本文构建了一组具有良好区分度的多维信息熵值作为评价指标,同时引入滑动窗口多维无参数 CUSUM 算法放大正常流量与 DDoS 攻击流量的差异,提高 DDoS 攻击检测精度。

1　检测向量的构建

在 DDoS 攻击检测时,检测指标构建的越全面,DDoS 攻击检测的精度越高,文献[7]从网络流中提取出 248 条流特征,这些特征能够全面描述网络流的状态,但文献[7]所需的时间和空间复杂度也是巨大的。考虑到 DDoS 攻击检测的时效性要求,本文选择 IP 包的源/目的 IP、源/目的端口 4 个基本特征进行变换,针对 DDoS 攻击的特点,构建具有良好区分度的多维熵值检测向量。

在 DDoS 攻击特征提取时,以 S 个连续的 IP 包作为单位流量,则对于一组数目为 S 的样本值,若 n_i 为样本 i 出现的次数,则有 $S=\sum_{i=1}^{N} n_i$,且该组样本的样本熵值为 $H(x)=-\sum_{i=1}^{N}\left(\frac{n_i}{S}\right)\log_2\left(\frac{n_i}{S}\right)$。

按照样本的源/目的 IP、源/目的端口等属性分别进行统计可以得到 IP、端口等不同的属性熵值,文献[4]指出这些样本熵值在一定的范围内较为稳定,因此采用样本熵值作为检测向量。而实际的 DDoS 攻击通常会同时对网络流量的多个属性造成影响,将网络流量的不同属性熵值分开来统计是不充分的,本文采用样本属性间的条件熵及相异熵来提高指标参数的区分度。

1.1　样本条件熵

对于变量 X 与 Y,变量 X 关于变量 Y 的条件熵 $H(X/Y)$定义为:

$$H(X/Y)=\sum_{y}p(y)H(X/Y=y)=-\sum_{y}p(y)\sum_{x}p(x/y)\log_2(P(x/y))$$

令 sip,dip,dport 分别表示 IP 包的源地址、目的地址、目的端口,根据 DDoS 攻击的特点,可采用以下 3 个样本条件熵检测 DDoS 攻击。

1) $H(\text{sip/dip})$:对于 DDoS 攻击而言,攻击流量中源地址相对于目的地址具有明显的多对一的映射关系,而正常流量间具有多对一、一对多与一对一 3 种映射关系,由 $H(\text{sip/dip})=\sum_{j}p(\text{dip}_j)H(\text{sip/dip}_j)$ 可知,当 sip 与 dip 存在多对一的映射关系时,$H(\text{sip/dip})$ 的值会显著增加。

2) $H(\text{dport/dip})$:对于 DDoS 攻击而言,攻击者通常会向目标主机请求尽可能多的服务,目的端口和目的地址间具有多对一映射关系,而合法用户通常在一段时间内请求的服务较为单一,$H(\text{dport/dip})$ 可用来描述目的端口和目的地址间的多对一映射关系。

3) $H(\text{sip/dport})$:对于针对某一特定服务的 DDoS 攻击而言,大量主机会向某固定端口请求服务,源地址和目的端口间存在多对一的映射关系,于是采用 $H(\text{sip/dport})$ 条件熵描述源地址和目的端口间的多对一关系。

1.2　相异样本熵

由于 Flash Crowd 等高频率访问发生时,同样

会出现多个 sip 对应 1 个 dip、多个 sip 对应 1 个 dport 的情况,仅采用条件熵无法有效区分 Flash Crowd 与 DDoS 攻击。与高频访问发生时会出现大量重复地址不同,大规模 DDoS 攻击爆发期间,网络中短期内可能会出现大量相异数据项,因此可以通过计算单位流量中的相异样本熵来检测 DDoS 攻击。

令 $X=(x_1,x_2,\cdots,x_p)$ 及 $Y=(y_1,y_2,\cdots,y_q)$ 分别表示相邻的 2 个单位流量中出现的源地址,令 $f(\text{sip})$ 表示单位流量中源地址 sip 出现的频率,且有 $\sum_{i=1}^{p} f(x_i)=\sum_{j=1}^{q} f(y_j)=S$,对于 Y 执行以下删除操作:对于 $\forall\, y_j\in Y$,$j=1,2,\cdots,q$,若 X 中存在 $x_i=y_j$,$i=1,2,\cdots,p$,则 $Y=Y-y_j$。

令 $\text{sip}'=(\text{sip}_m,\text{sip}_{m+1},\cdots,\text{sip}_n)$,$n\geqslant m$ 表示 1 组 sip 执行删除操作后的结果,则源地址相异样本熵定义为 $H(\text{sip}')=-\sum_{i=m}^{n} f(\text{sip}_i)/S\log_2\left(f(\text{sip}_i)/S\right)$。

综上,DDoS 攻击四维检测向量 $\boldsymbol{H}$ 由 $H(\text{sip}/\text{dip})$,$H(\text{dport}/\text{dip})$,$H(\text{sip}/\text{dport})$ 及 $H(\text{sip}')$4 个特征组成,见式(1):

$$\boldsymbol{H}=\{H(\text{sip}/\text{dip}),H(\text{dport}/\text{dip}),H(\text{sip}/\text{dport}),H(\text{sip}')\} \tag{1}$$

2 多维无参数 CUSUM 算法

2.1 一维无参数 CUSUM 算法

记随机序列$\{X_n\}$表示第 n 组单位流量中四维检测向量的某一特征,在正常情况下,X_n 值较小且比较固定,令 $E(X_n)=\alpha$,当 DDoS 攻击发生时,X_n 值迅速增加。令 $Z_n=X_n-\alpha-\beta$,其中 β是大于 0 的常数,它用来保证 Z_n 在正常情况下为负,而在变化点发生时,由于 $X_n\gg\alpha$,Z_n 会突然变大且为正,无参数 CUSUM 算法通过不断累积 Z_n 为正的值来判断攻击是否发生。

一维无参数 CUSUM 算法的定义为:$y_i=S_i-\min_{1\leqslant k\leqslant i} S_k$,其中 $S_k=\sum_{j=1}^{k} Z_j$,当 y_i 大于阀值时,判定序列发生了异常。通常情况下,为了便于计算,使用较多的是非参数 CUSUM 算法的递归版本:$y_i=(y_{i-1}+Z_i)^+$,其中,$y_0=0$,$x^+=\max(x,0)$。

由于 CUSUM 算法中 y_i 在变化点出现期间会累积很高的值,当攻击结束后即变化点结束的下沿出现时,y_i 无法迅速下降到正常水平,为了确保攻击期间累积的 y_i 能够迅速回归到正常水平,本文采用滑动窗口无参数 CUSUM 算法[8],该算法每次只累积有限窗口 T 内的 Z_i 值,滑动窗口无参数 CUSUM 算法相应的递归公式定义如下:

$$\begin{cases} y_i=(y_{i-1}+Z_i)^+, & i\leqslant T \\ y_i=(y_{i-1}+Z_i-(Z_{i-T})^+)^+, & i>T \end{cases} \tag{2}$$

最后通过判决函数 $d(y_i)$判断是否发生了异常,$d(y_i)$定义为:

$$d(y_i)=\begin{cases} 0, & y_i\leqslant N \\ 1, & y_i>N \end{cases} \tag{3}$$

当统计量 y_i 大于阀值 N 时,判决函数 $d(y_i)$为 1,表示序列$\{X_n\}$的统计特性发生变化。

2.2 多维无参数 CUSUM 算法

在 DDoS 攻击发生时,网络流量的多个特征通常会同时发生变化,通过融合多维观测特征,能够有效放大攻击流量与正常流量的差异,提高检测精度。基于此,本文扩展一维滑动窗口无参数 CUSUM 算法,在文献[9]的基础上引入多维滑动窗口无参数 CUSUM 算法检测 DDoS 攻击。

记随机序列$\{X_{n,m}\}$表示第 n 组单位流量中四维检测向量的第 m 个特征,其中 $m=1,2,3,4$,则多维滑动窗口无参数 CUSUM 算法定义如下:

$$\begin{cases} Z_{i,m}=X_{i,m}-\alpha_{i,m}-\beta_n \\ y_{1,m}=(Z_{1,m})^+ \\ y_{i,m}=(y_{i-1,m}+Z_{i,m})^+, & i\leqslant T \\ y_{i,m}=(y_{i-1,m}+Z_{i,m}-(Z_{i-T,m})^+)^+, & i>T \end{cases} \tag{4}$$

式中 $\alpha_{i,m}=(X_{1,m}+X_{2,m}+\cdots+X_{i,m})/i$,$\beta_n$ 为使 $Z_{i,m}$ 在正常情况下略小于 0 的正数。判决函数 $d(y_{i,m})$定义为:

$$d(y_{i,m})=\begin{cases} 0, & y_{i,m}\leqslant N_m \\ 1, & y_{i,m}>N_m \end{cases} \tag{5}$$

于是对于每组单位流量都能得到一个四维的判决向量 $\boldsymbol{D}_i=\{d(y_{i,1}),d(y_{i,2}),d(y_{i,3}),d(y_{i,4})\}$。

由于 DDoS 攻击的多样性,多个判决函数 $d(y_{i,m})$可能同时检测到变化或者部分判决函数检测到变化,简单地将多个判决函数加权求和[9]并不能精确描述 DDoS 的攻击特性。一种简单且能精确描述 DDoS 攻击特性的方法是构建 DDoS 攻击判决向量表,通过检查向量 $\boldsymbol{D}_t$ 是否与判决向量表匹配从而确定是否存在异常。以多个主机采用真实的 IP 地址向某一目标主机多个端口发出攻击为例,此类攻击对应的匹配向量 $\boldsymbol{S}=\{1,1,x,x\}$,其中 x 表示不确定,若在某一时刻 t 判决向量 $\boldsymbol{D}_t$ 与 $\boldsymbol{S}$ 匹配,则可判定 t 时刻网络中存在此种类型的 DDoS 攻击。

3 实验与评价

3.1 LLS-DDoS 数据集测试

本次实验数据集选用 MIT 林肯实验室提供的

分布式拒绝服务攻击数据集 LLS-DDoS-1.0 和 LLS-DDoS-2.0.2,2 个数据集均选用 inside 域中采集到的 dump 包,表 1 为采用 Plab 软件去除 IP 分片报文和非 IP 报文后 2 个数据集的基本信息,实验中以 2 000 个包为单位流量,图 1 和图 2 分别显示了LLS-DDoS-1.0和LLS-DDoS-2.0.2的四维检测向量$\boldsymbol{H}$中具有较高区分度的$H(\text{sip}/\text{sip})$、$H(\text{dport}/\text{dip})$、$H(\text{sip}')$3 个熵值的变化情况。

表 1　LLS-DDoS 数据集基本信息

Tab.1 The basic attribute of LLS-DDoS datasets

数据集	报文数	持续时间/s	攻击持续时间/s
LLS-DDoS-1.0	646 703	11 652	6
LLS-DDoS-2.0.2	346 672	6 166	8

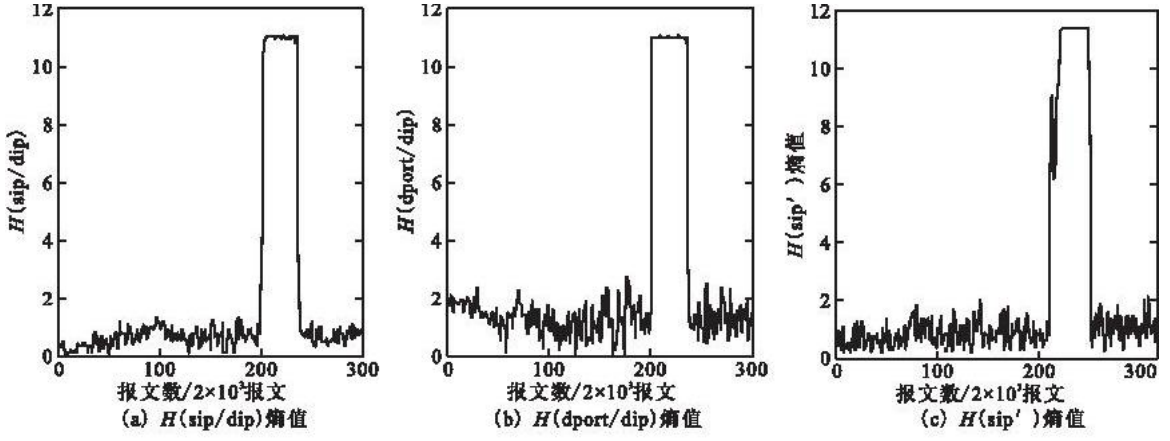

图 1　LLS-DDoS-1.0 熵值变化

Fig.1　The entropy sequence of LLS-DDoS-1.0

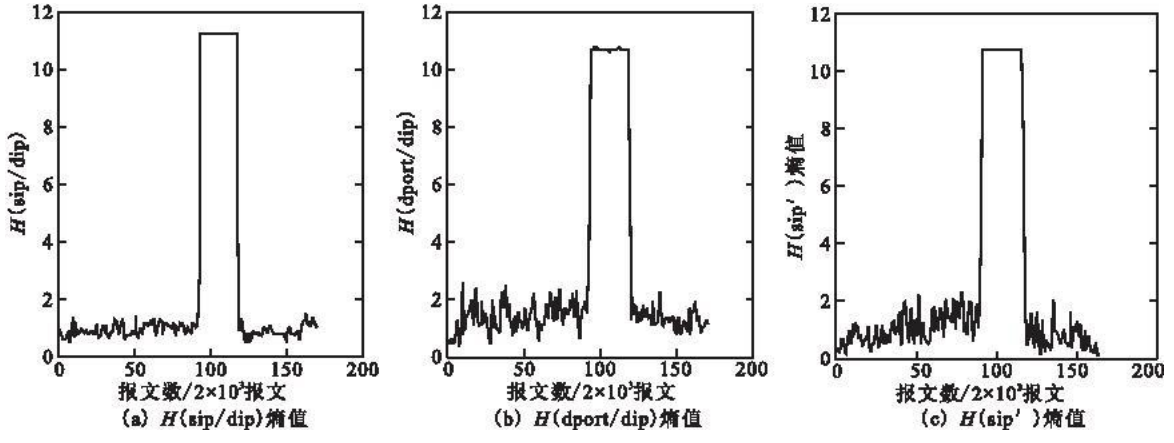

图 2　LLS-DDoS-2.02 熵值变化

Fig.2　The entropy sequence of LLS-DDoS-2.02

从图 1 和图 2 可以看出,$H(\text{sip}/\text{dip})$、$H(\text{dport}/\text{dip})$、$H(\text{sip}')$3 个特征对于 MIT 林肯实验室提供的 DDoS 攻击数据集具有良好的区分度。在滑动窗口多维无参数 CUSUM 算法中,对于 2 组 DDoS 数据集,参数$\{\beta,\beta,\beta,\beta\}$均设置为{1,1,0.5,1},滑动窗口 $T=5$,参数$\{N_1,N_2,N_3,N_4\}$设为{5,5,2,5}。对于 LLS-DDoS-1.0 数据集,通过多维无参数 CUSUM 算法能够检测到数据集从报文数为 203 的时刻其对应的 $y_{i,m}$ 迅速增加,相应的判决向量变为{1,1,0,1};对于 LLS-DDoS-2.0.2 数据集,通过多维无参数 CUSUM 算法能够检测到数据集从报文数为 94 的时刻其对应的 $y_{i,m}$ 迅速增加,相应的判决向量变为{1,1,0,1}。由于向量{1,1,0,1}与 DDoS 攻击判决向量表匹配,因此基于信息熵的多维 CUSUM 算法不仅能够有效检测到 DDoS 攻击,同时还能够准确检测出 DDoS 攻击开始时刻,但检测到的攻击结束时刻有 3～4 个时延的滞后。

3.2　合成数据集测试

为了使实验数据集更具一般性,在攻击数据集 LLS-DDoS-1.0 的基础上叠加背景流量,背景流量取自 MIT 实验室 1999 年的正常数据集,从 LLS-DDoS-1.0 攻击流中随机提取 35 个攻击样本,从正常数据集中随机提取 1 000 个正常样本,多维无参数 CUSUM 算法参数设置与 3.1 中相同,按照文献[10]的要求提取出数据集中全部 1 035 个样本的源/目的 IP、源/目的端口、协议的信息熵,将本文提

出的算法与文献[10]提出的基于信息熵的方法进行对比,对比结果见表 2。

表 2 2 种算法检测结果对比

Tab.2 Performance comparison between the two algorithms

算法	实际攻击数量	算法返回攻击数量	正确检测	错误检测	检测率/%	误报率/%
多维 CUSUM 算法	35	38	35	3	100	7.9
信息熵算法	35	42	33	9	94.3	21.4

从表 2 可以看出,本文提出的滑动窗口多维无参数 CUSUM 算法与基于信息熵的方法相比对 DDoS 攻击具有更好的检测效果,该算法能够检测到 1 035 个样本中的全部攻击样本,但算法存在少量的误报情况,主要原因是 CUSUM 算法不断累加正值导致攻击结束后累加值无法马上回归到 0。而在实际的应用中我们更加关心的是能否检测到 DDoS 攻击以及攻击开始的时刻,而攻击结束时刻的较小时延滞后不会对系统性能造成影响。另一方面,通过降低滑动窗口大小 T 可减小时延滞后,从而减小误报率,但过小的 T 值可能使得 CUSUM 算法累加范围过窄而造成部分攻击样本的漏检,在实际应用时需要根据具体情况设置合适的窗口大小。

4 结语

分布式拒绝服务攻击已经成为互联网中最大的安全威胁之一,准确及时地检测出 DDoS 攻击对于网络管理和网络安全具有重要的意义。本文通过分析 DDoS 攻击的特点,构建了具有良好区分度的多维信息熵值作为 DDoS 攻击的检测向量,在此基础上,采用滑动窗口多维无参数 CUSUM 算法放大正常流量与 DDoS 攻击流量的差异来实现攻击检测。通过实际网络攻击流量及合成攻击流量实验表明,本文提出的算法对于 DDoS 攻击具有较好的检测效果,算法能够应用在大型骨干网络中。

参考文献(References):

[1] Darren Anstee. DDoS attack trends through 2010, infrastructure security report & ATLAS initiative [EB/OL]. [2013-01-09]. http://ripe62.ripe.net/presentations/88-Darren-Anstee-AA-RIPE-2011-DDoS-Trends.ppt.pdf, 2011.

[2] 郑康锋,王秀娟. 利用边际谱 Hurst 参数检测 DDoS 攻击[J]. 北京邮电大学学报, 2011, 34(5): 128-132.
ZHENG Kangfeng, WANG Xiujuan. Detection DDoS attack with Hurst parameter of marginal spectrum [J]. Journal of Beijing university of posts and telecommunications, 2011, 34(5): 128-132. (in Chinese)

[3] Wang H, Zhang D, Shin K G. Change-point monitoring for the detection of Dos attacks[J]. IEEE trans on dependable and secure computing, 2004, 1(4): 193-208.

[4] 朱应武,杨家海,张金祥. 基于流量信息结构的异常检测[J]. 软件学报, 2010, 21(10): 2573-2583.
ZHU Yingwu, YANG Jiahai, ZHANG Jinxiang. Anomaly detection based on traffic information structure[J]. Journal of software, 2010, 21(10): 2573-2583. (in Chinese)

[5] 刘运,蔡志平,钟平,等. 基于条件随机场的 DDoS 攻击检测方法[J]. 软件学报, 2011, 22(8): 1897-1910.
LIU Yun, CAI Zhiping, ZHONG Ping, et al. Detection approach of DDoS attacks based on conditional random fields[J]. Journal of software, 2011, 22(8): 1897-1910. (in Chinese)

[6] Arun Raj Kumar P, Selvakumar S. Distributed denial of service attack detection using an ensemble of neural classifier[J]. Computer communications, 2011, 34(11): 1328-1341.

[7] Auld Tom, Moore Andrew W, Gull Stephen F. Bayesian neural networks for internet traffic classification[J]. IEEE transactions on neural networks, 2007, 18(1): 223-239.

[8] 孙知信,李清东. 基于源目的 IP 地址对数据库的防范 DDoS 攻击策略[J]. 软件学报, 2007, 18(10): 2613-2623.
SUN Zhixin, LI Qingdong. Defending DDoS attacks based on the source and destination IP address database[J]. Journal of software, 2007, 18(10): 2613-2623. (in Chinese)

[9] 康健,宋元章. 利用多维观测序列的 KCFM 混合模型检测新型 P2P botnet[J]. 武汉大学学报:信息科学版, 2010, 35(5): 520-523.
KANG Jian, SONG Yuanzhang. Application KCFM to detect new P2P botnet based on multi-observed sequence[J]. Geomatics and information science of Wuhan university, 2010, 35(5): 520-523. (in Chinese)

[10] Lakhina A, Crovella M, Diot C. Mining anomalies using traffic feature distributions[J]. Computer communication review, 2005, 35(4): 217-228.

(编辑:徐楠楠)