第20卷第4期
2017年4月

软件工程　SOFTWARE ENGINEERING

Vol.20 No.4
Apr. 2017

文章编号：2096-1472(2017)-04-04-04

基于统计特征的SYN Flood检测方法

刘　云，何　翼
（贵阳学院数学与信息科学学院，贵州 贵阳 550005）

摘　要：SYN Flood是当前最流行的拒绝服务(DoS)与分布式拒绝服务(DDoS)攻击方式。从构造一个SYN攻击报文的角度分析，SYN Flood攻击会引起网络中基于IP地址、标志位、端口号、序列号的统计特征异常，因此提出一种基于统计特征的SYN Flood攻击检测的方法。该方法首先从半连接队列中获取半连接信息，从全连接队列中获取IP地址存入BloomFilter中，再分别提取其统计特征，最后使用LMBP神经网络得到检测结果。实验结果表明该算法与其他算法相比具有更好的检测效果。

关键词：SYNFlood检测；统计特征；BloomFilter；LMBP神经网络
中图分类号：TP393　　　**文献标识码**：A

The SYN Flood Detection Method Based on Statistical Properties

LIU Yun,HE Yi
(*College of Mathematics and Information Science*,*Guiyang University*,*Guiyang* 550005,*China*)

Abstract: As the most popular attack method of denial of service (DoS) and distributed denial of service (DDoS),SYN Flood attack will cause some statistical properties abnormalities in the IP address,the TCP flag,the port number and the serial number from the perspective of constructing a SYN packet.Therefore,a SYN Flood attack detection method is proposed in this paper.Firstly,the method acquires half connection information from the half-connection queue,obtains IP addresses from the whole connection queue and puts them in Bloom Filter,then individually extracts statistical properties based on IP addresses,and finally determines whether SYN Flood attack happens by using LMBP neural networks.Experimental results show that the method could improve the effectiveness of detection.

Keywords:SYN flood detection;statistical properties;bloom filter;LMBP neural network

1　引言(Introduction)

SYN Flood是当前最流行的拒绝服务(DoS)与分布式拒绝服务(DDoS)攻击方式之一，也是一种非常流行的利用协议漏洞的资源匮乏型攻击，它对网络破坏力惊人[1]。因SYN Flood攻击技术操作方便、容易达到目的、更难于防范和追查，越来越成为黑客最常使用的网络攻击方式之一[2]。它主要利用TCP协议三次握手缺陷和IP欺骗技术，向目标服务器发送大量带有伪造IP地址的SYN报文，使目标服务器的TCP半连接队列被迅速占满而不能及时释放，造成无法建立正常的TCP连接，从而导致拒绝服务。

由于完整的TCP连接包括连接的建立和终止两个过程，在这些过程中SYN报文、ACK报文、SYN/ACK报文、FIN报文、RST报文的数量具有一定对称性。目前，针对SYN Flood攻击的检测技术主要根据这种对称性是否被破坏，以判断网络中是否存在SYN Flood攻击[3]。另外，提出一种基于重尾特性的SYN Flood检测方法，该方法将SYN报文的比重与流量的重尾特性相结合提高检测的正确率[4]。在攻击者发动SYNFlood攻击时，如果有意传送对等的SYN/ACK报文、ACK报文、RST报文、FIN报文、FIN/ACK报文，上述检测方法将失去作用。

2　SYNFlood攻击特征分析(SYN flood attack feature analysis)

本文从构造一个SYN攻击报文的角度分析，SYN Flood攻击会引起网络中基于IP地址、标志位、端口号、序列号的统计特征异常。据此，提出一种基于统计特征的SYN Flood攻击检测的方法。该方法首先从半连接队列中获取半连接信息，从全连接队列中获取IP地址存储到BloomFilter中，再分别提取基于IP地址、标志位、端口号、序列号的统计特征，

最后使用LMBP算法得到检测结果。

根据SYN Flood攻击原理得知，攻击端在发动攻击使需要大量构造SYN报文，而构造一个完整的SYN攻击报文不仅需要在网络层伪造IP地址，而且在传输层伪造源端口号、目的端口号、序列号[5]。为了快速检测出SYN Flood攻击和新型防检测的攻击形式，本文将从IP地址、TCP半连接队列、端口号、序列号、标志位五个角度进行特征分析。

(1)IP地址

数据显示，在正常网络中大约82.9%IP地址出现过；当网络中存在SYN Flood攻击时，只有0.6%—14%的IP地址是从前出现过。这是SYN Flood攻击带给IP地址的第一个统计特征。另外，当SYN攻击报文中IP地址为固定值时，IP地址的统计特征将呈现聚集趋势；当SYN攻击报文中IP地址为随机值时，这时IP地址的统计特征呈现发散趋势。这是SYN Flood攻击带给IP地址的第二个统计特征。

(2)端口号

当构造SYN攻击报文时，源端口号/目的端口号为固定值时，使得源端口号/目的端口号统计特征呈现聚集趋势；当构造SYN攻击报文时，源端口号/目的端口号为随机值时，使得源端口号/目的端口号统计特征呈现发散趋势。

(3)序列号

与端口号相似，当构造SYN攻击报文时，报文的序列号也存在固定值和随机值，也会导致序列号的统计特征也呈现聚集或发散趋势。

(4)标志位

根据TCP/IP协议，与TCP连接相关的标志位主要为ACK、RST、SYN、FIN。在TCP连接正常建立和终止过程中，网络中这些标志位的报文数量存在一定对称性。然而，当网络中存在SYNFlood攻击时，这种数量上的对称性将遭到破坏。

需要注意的是当恶意攻击者在大量发送SYN报文的同时，向目标服务器发送相应数量的SYN/ACK报文、ACK报文、RST报文、FIN报文、FIN/ACK报文时，会促使目前大多数SYN Flood检测技术失去作用。针对这种新型防检测的SYN Flood攻击，需要严格按照报文的源地址和目的地址进行报文统计分析。

(5)TCP半连接队列

由SYN Flood攻击原理可以看出，攻击的最终目的是耗尽TCP半连接队列，从而导致拒绝服务。因此，TCP半连接队列是SYN Flood攻击最根本目标，也能最直接、准确反应出SYN Flood攻击的存在。

3 基于统计特征的检测算法(Detection algorithm based on statistical features)

该算法主要分为获取TCP连接信息、提取统计特征、检测攻击三个部分。第一部分主要从半连接队列和全连接队列中获取所需的连接信息。第二部分根据第一部分的连接信息，提取半连接队列长度、IP地址、标志位、源端口号、目的端口号、序列号的统计特征。第三部分，根据第二部分的统计特征，通过BP神经网络判断网络中是否存在SYN Flood攻击。

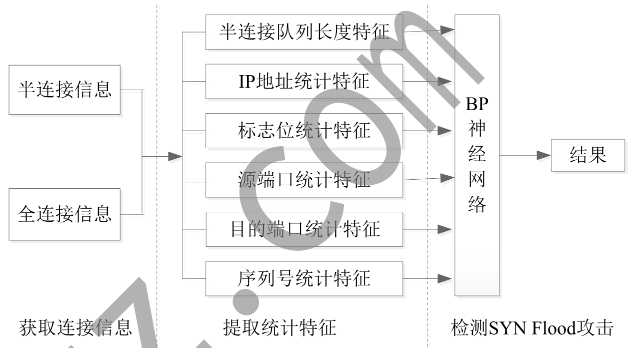

图1 基于统计特征的检测算法

Fig.1 Detection algorithm based on statistical features

3.1 获取TCP连接信息

为了根据上述统计特征快速检测到SYN Flood，需要从半连接队列提取半连接信息，从全连接队列中提取已经建立连接的IP地址库。本文分别采用以下两种策略提取连接信息。

(1)基于集合方式提取半连接信息

为了避免频繁访问半连接队列造成消耗过多的CPU和内存资源，本文采用周期性提取半连接队列中信息。同时，为了方便对半连接队列的快速处理，每间隔时间t将半连接队列中状态信息提取到状态集合Set_t中，其中集合定义如下：

$$Set_t = \{<ip_{t,1}, Links_{t,1}>, <ip_{t,2}, Links_{t,2}>, \ldots, <ip_{t,m}, Links_{t,m}>\}, m \ge 0 \quad (1)$$

其中，

$$Links_{t,i} = \{link_{t,i,1}, link_{t,i,2}, \ldots, link_{t,i,n}\}, 0 \le i \le m \quad (2)$$

在式(1)和式(2)中，$ip_{t,i}$表示第i个IP地址($0 \le i \le m$)；$Links_{t,i}$表示第i个IP地址的半连接信息集合($0 \le i \le m$)；$link_{t,i,j}$表示第i个IP地址第j个半连接信息($0 \le i \le m, 0 \le j \le n$)，其定义如下：

$$link_{t,i,j} = <sport_{t,i,j}, dport_{t,i,j}, seq_{t,i,j}, time_{t,i,j}, retry_{t,i,j}> \quad (3)$$

在式(3)中，$sport_{t,i,j}$、$dport_{t,i,j}$、$seq_{t,i,j}$、$time_{t,i,j}$、$retry_{t,i,j}$表示第i个IP地址对应的第j个半连接的源端口地址、目的端口地址、序列号、时间、重传次数（$0 \le i \le m$，$0 \le j \le n$）。

(2)基于BloomFilter提取全连接信息

BloomFilter是一种采用比特数组表示一个集合的数据结构，具有极高的时间和空间效率。为了减少存储空间的消耗和加快查找速度，本文将所有已建立连接的IP地址集合IP保存在BloomFilter中。

$$IP=\{ip_1,ip_2,\ldots,ip_w\},w\geq 0 \tag{4}$$

目前广泛使用的IPv4地址由4个8位比特组成，可表示为“$A_1.A_2.A_3.A_4$”形式。为了将IPv4地址保存在BloomFilter结构中，本文采用四个独立的哈希函数，四个256位比特数组分别对应于IP地址的四个字段，如图2所示。

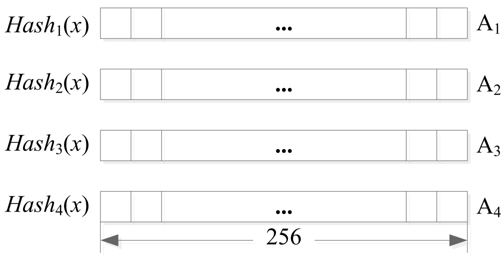

图2 BloomFilter结构

Fig.2 BloomFilter structure

3.2 提取统计特征

由于TCP半连接队列是SYN Flood最终目标，也最直接反映攻击的存在，本文采用比值方式反映半连队列的使用情况(式5)。

$$v_1=\frac{length_t}{Length} \tag{5}$$

其中，$length_t$表示半连接队列已使用长度值，$Length$表示半连接队列的最大长度值。

为了有效检测网络中SYN报文与SYN/ACK报文的对称性是否遭到破坏，采用式(6)反映对称的变化。

$$v_2=\frac{count_{t,syn}}{count_{t,syn/ack}+count_{t,syn}} \tag{6}$$

其中，

$$count_{t,syn}=length_t \tag{7}$$

$$count_{t,syn/ack}=\sum_{i=0}^{m}\sum_{j=0}^{n}retry_{t,i,j} \tag{8}$$

其中，$count_{syn}$表示收到来自客户端的SYN报文数量；$count_{syn/ack}$表示收到来自服务器的SYN/ACK报文数量。

从半连接状态集合Set_t中，统计每一个IP地址的半连接数量，得到集合$set_{t,ip}$，表示为如下形式。

$$set_{t,ip}=\{<ip_{t,1},count_{t,ip,1}>,<ip_{t,2},count_{t,ip,2}>,\ldots,<ip_{t,m},count_{t,ip,m}>\},m\geq 0 \tag{9}$$

其中，$count_{t,ip,i}$表示第i个IP地址的半连接数量。

根据已经建立过TCP连接的IP地址集合IP，计算半连接状态集合中新IP地址的比重，见式(8)。

$$v_3=\frac{\sum_{i=1}^{m}f_{t,i}}{Sum_{ip,t}} \tag{10}$$

其中，

$$f_{t,i}=\begin{cases}0 & (ip_{t,i}\in IP,0\leq i\leq m)\\ count_{t,ip,i} & (ip_{t,i}\notin IP,0\leq i\leq m)\end{cases} \tag{11}$$

$$Sum_{ip,t}=\sum_{k=1}^{m}count_{t,ip,k} \tag{12}$$

为了检测IP地址是否具有发散或聚集特征，本文采用熵的方式反映半连接IP地址的分布情况。

$$v_4=-\sum_{k=1}^{m}\left(\frac{count_{t,ip,k}}{Sum_{ip,t}}\right)\ln\left(\frac{count_{t,ip,k}}{Sum_{ip,t}}\right) \tag{13}$$

为了检测源端口号、目的端口号、序列号的统计特征，从半连接状态集合set_t中分别提取源端口号集合$set_{t,sport}$、目的端口号集合$se_{tt,dport}$、序列号集合$set_{t,seq}$，并统计每个元素出现的次数，表示为如下：

$$\begin{aligned}set_{t,sport}=\{&<sport_{t,1},count_{t,sp,1}>,<sport_{t,2},count_{t,sp,2}>,\ldots\\&\ldots,<sport_{t,l},count_{t,sp,l}>\},l\geq 0\end{aligned} \tag{14}$$

$$\begin{aligned}set_{t,dport}=\{&<dport_{t,1},count_{t,dp,1}>,<dport_{t,2},count_{t,dp,2}>,\ldots\\&\ldots,<dport_{t,p},count_{t,dp,r}>\},r\geq 0\end{aligned} \tag{15}$$

$$\begin{aligned}set_{t,seq}=\{&<seq_{t,1},count_{t,seq,1}>,<seq_{t,2},count_{t,seq,2}>,\ldots\\&\ldots,<seq_{t,seq,z},count_{t,seq,z}>\},z\geq 0\end{aligned} \tag{16}$$

其中，$count_{t,sp,i}$表示第i个源端口号的出现次数($0\leq i\leq l$)；$count_{t,dp,i}$表示第i个目的端口号的出现次数($0\leq i\leq r$)；$count_{t,seq,i}$表示第i个序列号的出现次数($0\leq i\leq z$)。

为了检查源端口号、目的端口号、序列号是否具有聚集或发散趋势，同样使用熵反映其统计特征。

$$v_5=-\sum_{k=1}^{l}\left(\frac{count_{t,sp,k}}{Sum_{sport,t}}\right)\ln\left(\frac{count_{t,sp,k}}{Sum_{sport,t}}\right) \tag{17}$$

$$v_6=-\sum_{k=1}^{r}\left(\frac{count_{t,dp,k}}{Sum_{dport,t}}\right)\ln\left(\frac{count_{t,dp,k}}{Sum_{deport,t}}\right) \tag{18}$$

$$v_7=-\sum_{k=1}^{z}\left(\frac{count_{t,seq,k}}{Sum_{seq,t}}\right)\ln\left(\frac{count_{t,seq,k}}{Sum_{seq,t}}\right) \tag{19}$$

其中，

$$Sum_{sport,t}=\sum_{k=1}^{l}count_{t,sp,k} \tag{20}$$

$$Sum_{deport,t}=\sum_{k=1}^{r}count_{t,dp,k} \tag{21}$$

$$Sum_{seq,t}=\sum_{k=1}^{z}count_{t,seq,k} \tag{22}$$

3.3 检测SYNFlood攻击

BP神经网络常用于完成从多维到任意维的映射，而

LMBP既具有BP神经网络的自适应、自学习、自组织等优点，又能够避免BP神经网络收敛速度慢、容易陷入局部最小的缺点。本文采用三层LMBP神经网络将7个统计特征映射为检测结果(图3)。其中，输入层共7个节点，分别对应于7个统计特征；输出层共1个节点为[0,1]之间的实数$result_t$；隐藏层的初始化节点数根据式(23)计算：

$$node_{\text{hidden}} = \sqrt{node_{\text{input}} + node_{\text{output}}} + \delta, (0 \le \delta \le 10) \quad (23)$$

其中，$node_{\text{hidden}}$表示隐藏层节点个数；$node_{\text{input}}$表示输入层节点个数，在本文中为7；$node_{\text{output}}$表示输出层节点个数，在本文中为1；δ是常数。

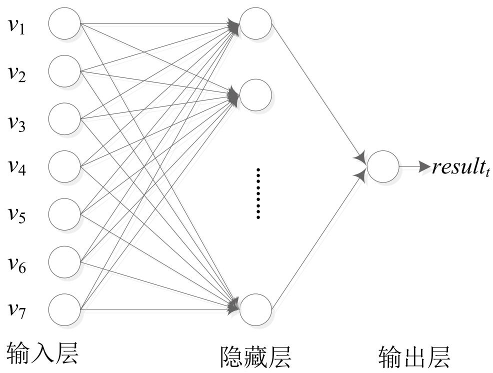

图3 三层BP神经网络

Fig.3 Three layer BP neural network

4 实验与结果(Experiment and results)

为了最大限度模仿互联网环境和检测本文方法的有效性，选择贵阳学院论坛作为实验对象，该论坛运行于Linux服务器上，面向全校师生开放，具有很高的访问量。

4.1 获取样本数据

由于BP神经网络在应用于检测之前必须学习，因此需要采集一定数量的半连接信息和全连接信息用于训练模型。

(1)获取半连接信息样本

为了采集无SYN Flood攻击的网络流量数据，收集了该论坛24小时的正常流量信息。在服务器上每间隔5分钟从半连接队列提取一次半连接信息存入MySQL数据库中。这样共采集到288条无SYN Flood攻击的样本数据。

为了获取带有SYN Flood攻击的网络流量数据，首先分别采用固定和随机的IP地址、端口、序列号组装SYN攻击报文，再采用不同的速率向服务器发送这些攻击报文，最后在论坛服务器采集半连接信息。与采集无SYN Flood攻击的流量一样，在服务器上每间隔5分钟从半连接队列中提取一次半连接信息存入MySQL数据库中，持续12小时，共得到144条含SYN Flood攻击的样本数据。

(2)获取全连接信息样本

Netfilter是从Linux2.4版本之后提供的一个通用防火墙框架，该框架在报文流经的几个关键点定义了大量HOOK，以方便用户注册钩子函数，实现对报文的自定义处理。NF_IP_LOCAL_IN是Netfilter在报文流入防火墙网络层处定义的HOOK。在NF_IP_LOCAL_IN处可以增加自定义处理，判断报文的类型，若是ACK报文则提取IP地址存入BloomFilter中。为了保证全连接信息与半连接信息同步匹配，在采集半连接信息的同时，将BloomFilter瞬时状态存入的MySQL数据库中。

4.2 实验结果与分析

本文选取2/3无SYN Flood攻击的样本和存在SYN Flood攻击的样本作为训练数据，剩余的1/3样本用于验证本文方法的有效性。训练中将LMBP神经网络的输出resultt分为两类：区间[0,0.5]为正常态，即不存在SYN Flood攻击；区间(0.5,1]为攻击态，即存在SYN Flood攻击。

为了验证本文方法的有效性，本文选择检测的误报率和漏报率作为评价标准，并与基于TCP缓存的检测方法[6]、基于状态防火墙的检测[7]、基于轻量级检测和混合连接策略的SYN Flood防御方法[8]，各种方法的检测结果如表1所示。

表1 SYN Flood攻击的检测效果

Tab.1 Detection effect of SYN Flood attack

SYN Flood检测方法	误报率(%)	漏报率(%)
基于状态防火墙的检测	11.80	14.58
基于TCP缓存的检测	9.28	10.41
基于重尾特性的检测	12.6	15.5
基于轻量级检测和混合连接策略	5.3	7.6
本文检测	4.86	6.25

通过表1可以看出，本文检测方法相对于其他两种方法无论是在误报率还是在漏报率方面均有不同程度的降低。通过以上实验结果可以看出，本文提出的检测方法相对于已有的SYN Flood攻击检测方法具有更好的检测效果。

5 结论(Conclusion)

针对分布式SYN Flood攻击会引起网络中基于IP地址、标志位、端口号、序列号的统计特征发生异常。本文提出一种基于统计特征的SYN Flood攻击检测的方法。该方法首先从半连接队列和全连接队列中获取连接信息，再分别提取基于IP地址、标志位、端口号、序列号的统计特征，最后使用LMBP算法得到检测结果。实验结果表明该算法具有较好的检测效果。下一步的工作是研究如何将此检测方法与防御策略结合适用，以提高防御效果。

参考文献(References)

[1] MING Yu.An Adaptive Method for Source-End Detection of Pulsing Dos Attacks[J].International Journal of Security and Its Applications,2013,7(5):279-288.

[2] MITKO B.Analysis of the SYN Flood Dos Attack[J].I.J.Computer Network and Information Security,2013,8(1):1-11.

[3] Behrouz A.Forouzan.TCP/IP Protocol Suite,Fourth Edition[M].Beijing:Tsinghua University Press,2014.

[4] 许晓东,杨海亮,朱士瑞.基于重尾特征的SYN洪流检测方法[J].计算机工程,2008,34(22):179-181.

[5] 陶建喜,等.非对称路由环境下SYN flood攻击防御方法[J].通信学报,2013,8(34):285-291.

[6] 胡鸿,袁津生,郭敏哲.基于TCP缓存的DDoS攻击检测算法[J].计算机工程,2009,35(16):112-114.

[7] 刘云.基于Linux平台的新的SYN Flood防御模型研究[J].计算机科学,2013,11A:210-213.

[8] 刘云.基于轻量级检测和混合连接策略的SYN Flood防御方法[J].计算机应用与软件,2016,11(33):310-313.

作者简介：

刘　云(1981-)，女，硕士，副教授.研究领域：网络安全.

何　翼(1981-)，女，硕士，副教授.研究领域：无线传感网络.

（上接第17页）

升了公交服务满意度及公交出行率，为智能公交系统的优化(线路站点规划、实时调度)提供了数据依据，对解决城市公共交通问题具有重要的现实意义。

参考文献(References)

[1] BinYu.WilliamH.K.Lama,MeiLamTam.Bus Arrival Time Prediction at Bus Stop with Multipleroutes[J].Transportation Research PartC,2011:1157-1170.

[2] ZHONG Huiling,ZHANG Meng,SHI Yongqiang.Efficient Path Planning Algorithm Based on the Strategy of Network Layer[J].Journal of Southwest University Jiao Tong,2011,46(4):645-650.

[3] ShafhiY,Khani A.A Practical Model for Transfer Optimization in a Transit Network:Model Formulations and Solutions[J].Transportation Research Part A:Policy and Practice,2010,44(6):377-389.

[4] Chang-Jiang Zheng.Improved Iterative Prediction for Multiple Stop Arrival Time Using a Support Vector Machine[J].TRANSPORT,2012,27(2):158-164.

[5] 王拙.基于有限状态自动机的公交车到站时间预测模型[D].哈尔滨:哈尔滨工业大学,2012.

[6]何巍楠.基于浮动车数据的城市常发性交通拥堵时空分布特征研究[D].北京交通大学,2012.

[7] 孙玉现,等.基于路况相似性的城市公交车到站时间预测机制[J].软件学报,2012,23(1):87-99.

[8] 王建.基于贝叶斯网络组合模型的公交驻站时间预测[J].城市交通,2012,10(5):78-83.

[9] 蒋新.基于多源数据的大城市常规公交运行评价研究[D].北京交通大学,2011.

[10] 李嘉.基于交通数据融合技术的行程时间预测模型[N].湖南大学学报,2014,41(1):33-38.

[11] 聂庆慧,等.基于多源ITS数据的行程时间预测体系框架及核心技术[N].东南大学学报,2011,41(1):199-204.

[12] 李川.基于复杂网络的城市公共交通网络连通性研究[D].成都:西南交通大学,2014.

作者简介：

李少伟(1981-)，男，硕士，讲师.研究领域：GIS/GPS在智能交通中的应用.

曹成涛(1981-)，男，博士，副教授.研究领域：智能交通技术.

杨　骥(1983-)，男，博士生，副高级工程师.研究领域：地理信息系统.

（上接第28页）

2015(06):9-12.

[5] 景再方,吕洁,孙向阳.营销专业大学生兼职行为及影响因素分析[J].高等农业教育,2010(05):73-78.

[6] 贺祖斌,杨红秀.大学生兼职功能的社会学分析[J].桂林师范高等专科学校学报,2008(01):87-90.

[7] Fazackerley Anna.Financial Fears Force Students to Juggle Work and Study[J].Times Higher Education Supplement,2006,10(3):11-12.

[8] 杨丰盛.Android应用开发揭秘[M].北京:机械工业出版社,2010.

作者简介：

张靖玉(1994-)，女，本科生.研究领域：电子商务.

张　琰(1982-)，女，硕士，副教授.研究领域：电子商务.

沈　洋(1995-)，女，本科生.研究领域：电子商务.

邢　燕(1996-)，女，本科生.研究领域：会计.