分散式阻断攻击的侦测

张维维
(湛江师范学院教育技术部　广东湛江 524048)

摘　要　本文提出一种有效侦测与过滤阻断服务攻击的方法 STTOPS(Sequential Testing with Tabulated Online Packet Statistics for DDos Detection),能够在一个紧密、固定尺寸的架构里使用有效的探索法监控许多网络地址。经实验证明,与 TOPS 相比较,STTOPS 在一个标准基准数据集方面有更高的平均准确率和更低的平均误报率,且使用较少计算资源,在攻击期间不会减慢下来。

关键词　分散式阻断攻击,TOPS,连续假设实验,STTOPS

The Detection of DDos Attack

ZHANG Wei-Wei
(EducationTechnology Department, Zhanjiang Normal College, Guangdong Zhanjiang 524048)

Abstract　We present an efficient method for detecting and filtering denial-of-service bandwidth attack. Our system called STTOPS (Sequential Testing with Tabulated Online Packet Statistics for DDos Detection) can monitor a large number of network addresses in a compact, fixed-size structure using several effective heuristics. Proved by experiment, STTOPS can detect bandwidth attack in a standard benchmark dataset with a higher average accuracy and a lower average inaccuracy than TOPS, and it uses less computational resources than TOPS and does not slow down during an attack.

Keywords　DDos, Tops, Sequential hypothesis testing, STTOPS

1　引言

随着科技的演变,电脑和网络已逐渐成为人们日常生活不可缺少的工具,然而伴随而来的是黑客入侵、蠕虫感染、分散式阻断服务攻击等等各式各样网络安全事件,其中的分散式阻断服务攻击正是网络安全界亟待解决的重要课题。

2　分散式阻断服务攻击

分散式阻断服务攻击通常简称为 DDos,即"Distributed Denial of Service"的缩写。顾名思义,它利用网络上已被攻陷的电脑作为傀儡机,向某一特定的目标电脑发动密集式的"拒绝服务"要求,借以把目标电脑的网络资源及系统资源耗尽。一旦目标电脑负荷过重而倒下,攻击者即可透过系统的漏洞而入侵目标电脑。

分散式阻断服务攻击依照攻击原理分为两类:频宽攻击和系统资源攻击。就频宽攻击而言,在攻击者端由于攻击的流量不大,不易侦测到流量的异常,但是经由网络慢慢汇集到受害者端时,攻击流量就相当惊人,导致受害的网络连线无法处理那么多封包,产生进出边缘路由器网络流量不对等的现象发生。就系统资源攻击而言,攻击者端会与受害者端建立许多半开的连线,受害者端可以缩短等待连线建立的时间、限制建立半开连线数目等方式来回应,有趣的是攻击者端送出许多建立连线的 SYN 封包,但是攻击封包上填的是假的源地址,导致 ACK 封包无法送回到攻击者端,这也存在进出边缘路由器封包不对等的现象,也可利用上述的侦测方式去侦测。

因此,本文将着眼于进出网络流量不对等这项重大特征,结合 TOPS 与(连续假设实验 Sequential hypothesis testing)两种算法的优点,简化分散式阻断服务攻击的侦测方式。

3　在线表列式封仓统计与连续假设实验

在线表列式封包统计(Tabulated Online Packet Statistics,TOPS)是放置于边缘路由器上(leaf router),采用固定大小的资料结构来监控网络流量的异常。它依据 IP 地址的特性(ex. 140. 113. 13. 112)建立四个大小含有 256 个格子(entry)的表格,每个格子里都有两个参数分别记载进来(Pin)与出去(Pout)子网络的封包流量,再依照 Pin/Pout 的比值去判断是否遭受攻击。为了提高发出警报的准确性,加入了确定门槛(certainty threshold),这是个不易实现的方法,必须要根据进来或出去目标网络的封包速率产生一个累积机率分布,再根据累积分布判断是否为遭受攻击。要储存此项资讯实为不易,是实现 TOPS 最大的困难。

连续假设实验方法经常被用于蠕虫侦测方面。善意的电脑主机只会和远端系统,试着建立应该建立成功的连线,但是受到蠕虫感染的电脑主机,则是会尽一切所能感染其他的电脑主机,但是因为有些电脑它的某个连接没开,甚至是电脑主机没开机,导致在一定时间内很多的连线是失败的,这个现象跟 TCP SYN Flooding 很像,差别在于受到蠕虫感染的电脑主机是一台对多台电脑进行感染,而 TCP SYN Flooding 则是多台电脑主机攻击一台电脑主机,衍生出来的侦测位置则有差异。遭受蠕虫感染的电脑主机在发动攻击时,会在它所经过的边缘路由器就被侦测到,而发动 TCP SYN Flooding

张维维　硕士,讲师,主要研究方向:计算机教育及应用。

攻击的意图，则是在远处受害者端的边缘路由器才被侦测到。但是两者的侦测原理都是一样的，在一定时间内若是建立连线失败的次数太多，我们就会怀疑是否发生攻击。

该方法设定 Y_i 是一个代表第 i 次连线结果的随机变量。如果连线成功，$Y_i=1$；如果连线失败，$Y_i=0$。又设定 H_1 是某台电脑主机在从事攻击的前提，H_0 是没有在从事攻击。假设以 H_j 为前提之下，随机变量 $Y_i|H_j$，$i=1,2,3,\cdots$，彼此是独立同分布。

$$\phi(Y_i)\equiv\frac{P_r[Y_i|H_1]}{P_r[Y_i|H_0]}=\begin{cases}\dfrac{\theta_1}{\theta_0} & Y_i=0\\ \dfrac{1-\theta_1}{1-\theta_0} & Y_i=1\end{cases}$$

其中 $P_r[Y_i=0|H_0]=\theta_0$　$P_r[Y_i=0|H_1]=\theta_1$，$\phi(Y_i)$ 为第 i 次的观察。

设定上限 η 和下限 η_0，$\phi(Y_i)$ 超过上限时遭到攻击，低于下限时没有遭到攻击。

该方法应用数理统计便于借由一段时间所发生的事件去判定实际的状况，弥补了 TOPS 的不足。

4　改进的推演算法

结合 TOPS 与连续假设实验两种方法的优点，我们提出 STTOPS(Sequential Test with Tabulated Online Packet Statistics)。首先，我们拟采用 TOPS 的储存架构去存取进出边缘路由器的封包数(Pin、Pout over some interval t)，假设采用受害者模式(victim mode)，TOPS 是利用 R=Pin/Pout 的比值来判断是否发生异常，若 R>Rmax，则发出警告(alarm)，再经过之前所储存的流量累积机率分布去判定是否发生攻击(attack)，困难点在于要储存累积机率分布不是一件容易的事，所以我们将处用连续假设实验的概念来简化所需的运算。一样针对每个 table entry 储存两个变量 Pin、Pout，再加上两个变量 A(A 初始值设为 0)与 Y(Y 初始值设为 0)。在之前 TOPS 是利用比值 R=Pin/Pout>Rmax，在此我们转换成 D = Pin-Rmax * Pout 简化计算之复杂度，若 D≥0，则发出警告(alarm)、把 Y 加一；若 D<0，则视为正常，把 Y 减一。每次更改 Y 值后，把 Y 与我们事先设定好的阈值(threshold)α 做比较。若 Y≥α，则判定为攻击状态(attack)，A 设为 1；Y 若大于 α，则 Y 设定为 α，是因为不想让 Y 无限制地往上升，导致前面的攻击侦测会影响到后面的攻击侦测。若 0≤Y<α，则 Y、A 值不变。在判定单一 entry 是否为攻击或正常状态之前，我们不对可疑封包采取任何的措施，一旦某个 IP 地址有四个的 table entries 判定为攻击状态(k=4)，则采取限流(throttle)等相关措施。

一开始先从某 tcpdump 档案里读出一个封包资料，利用 IP 标头里的目的地址资料来判断此封包是进入或出去子网络，再借由路由器上统计资料来判断此封包的目的地、地址是否遭受攻击，如果遭受攻击就发出警告，如果没有遭受攻击，就进行进出封包流量统计。直到读入的封包数达到 n 时，就计算全部表格的 Y 值且判断是否超过上限 α，并把进入封包数(Pin)与出去封包数(Pout)归零重新读数。反复经过数次统计，若 Y≥α，判定为遭受攻击并把 A 设为 1，如果 Y≥α，把 Y 设定为 α。若 0≤Y<α，则保留 Y、A 的值让之后的统计再去判断。

5　算法实现及结果

我们实验所采用的数据是 MIT Lincoln Lab 所制作的 DARPA off-line intrusion data set 入侵资料中，它主要包含了入侵资料，但也有若干阻断服务攻击的资料在里面。而这其中又包括了许多类型，我们的目标设定在侦测外部电脑主机对特定监视的网域所做的攻击与特定监视网域遭受攻击时进出不对等流量特性的正确性。从外部电脑主机发动的攻击分为两种类型：TCP SYN floods 和 ICMP floods。进一步监视 TCP SYN floods 发现大多数的 SYN 封包都会被受害者承认，因此这种类型的低强度 TCP SYN floods 将不会被侦测出来，因为进出流量不对等的比重太低，所以我们就专注属于 ICMP floods 类型名字叫做 pod 与 smurf。

参数 k、n、Rmax、α 的确定如下：

不同的参数对实验结果产生不同的影响。因此，我们做了大量的实验以确定参数的取值，实验结果列表如下：

表 1　k 值对侦测率与错误警告的影响

Attack packet detection rate(%)		
	K=4	K=3
Rmax=2	60.63	60.63
Rmax=4	60.63	60.63
Rmax=6	60.63	60.63
Rmax=8	60.63	60.63
Rmax=10	60.63	60.63
Rmax=20	60.63	60.63
False alarms		
Rmax=2	610	614
Rmax=4	270	274
Rmax=6	0	1
Rmax=8	0	1
Rmax=10	0	1
Rmax=20	0	1

表 2　n 值对侦测率与错误警告的影响

Attack packet detection rate(%)			
	n=50	n=100	n=200
Rmax=2	90.85	80.85	60.63
Rmax=4	89.73	80.85	60.63
Rmax=6	89.73	80.85	60.63
Rmax=8	89.73	80.85	60.63
Rmax=10	89.73	80.85	60.63
Rmax=20	89.73	78.6	60.63
False alarms			
Rmax=2	1831	1326	610
Rmax=4	1371	989	270
Rmax=6	750	0	0
Rmax=8	0	0	0
Rmax=10	0	0	0
Rmax=20	0	0	0

表 3　α 值对侦测率与错误警告的影响

α	1	2	3	4	5
Attack packet detection rate(%)	X	97.6	96.47	95.35	94.22
False alarms	X	1879	1295	1106	1022
α	6	7	8	9	10
Attack packet detection rate(%)	93.1	91.98	90.85	89.73	88.6
False alarms	921	836	793	750	622
α	11	12	13	14	15
Attack packet detection rate(%)	87.48	86.36	85.24	84.11	82.99
False alarms	576	495	452	409	364

(下转第 123 页)

表 3 不同编码方案的编码效率比较

副本大小(MB)	m	n	Cauchy Reed-Solomon Code 编码时间(s)	Cauchy Reed-Solomon Code 编码时间(s)	Tornado Code 编码时间(s)	Tornado Code 编码时间(s)	Random Linear Code ($q=2^8$) 编码时间(s)	Random Linear Code ($q=2^8$) 编码时间(s)
10	4	8	0.485	0.233	0.213	0.105	0.901	0.242
20	4	8	1.197	0.552	0.530	0.252	2.401	0.561
30	4	8	1.582	0.776	0.734	0.346	3.242	0.798
40	4	8	2.254	1.012	0.989	0.498	4.095	1.021
10	8	16	0.645	0.312	0.245	0.120	1.253	0.320
20	8	16	1.466	0.701	0.572	0.257	2.877	0.782
30	8	16	2.008	1.022	0.806	0.363	4.102	1.084
40	8	16	2.639	1.714	1.051	0.478	5.242	1.832

实验采用的是 10-40MB 的副本数据，对于更大甚至上 G 级的副本而言，其编解码也是划分为 M 级的数据段进行的，因此大副本数据的编解码也符合相同的线性增长规律。

(2)副本数据的网络传输实验

测试平台为五台 P4 1.8G(512MB,Linux)台式机，GridFTP 版本为 2.1，网络环境为 100M 局域网。实验过程中 TCP Buffer 为 64kb，改变传输副本大小和传输模式进行实验，观察网格数据传输时间的变化。

表 4 不同传输模式的传输效率比较

副本大小(MB)	Parallelism=0 Striping=0 传输时间(s)	Parallelism=4 Striping=0 传输时间(s)	Parallelism=4 Striping=2 传输时间(s)	Parallelism=4 Striping=4 传输时间(s)
10	3.062	3.185	1.893	1.584
20	6.754	6.511	3.847	2.827
30	10.117	9.962	5.016	4.182
40	13.298	13.141	6.772	5.554

实验是在封闭网络条件下进行的，可以认为是同等条件的最大传输效率。对于第四种传输模式，则基本上达到了客户机的网络接受上限。

通过对比表 3 和表 4 的实验结果可以发现：即使是在最优的传输模式下，对相同大小的副本数据来说，编码时间仍接近于传输时间，解码时间则远低于传输时间。如此的编码开销与传输开销比值是可以接受的，因为可采用数据编码与数据传输同步进行的方式，使得编码开销只是消耗整个数据复制过程的计算资源，而几乎不消耗时间资源。那么即使是处理上 G 级的副本数据，相比于传统数据复制过程来说，编码数据复制将也不会有明显的操作时间增加。

总结 本文提出了基于线性分组编码机制的网格数据复制思想，通过对 Cauchy Reed-Solomon Code、Tornado Code 和 Random Linear Code 进行建模比较分析，证明基于编码机制的数据复制较传统数据复制有着传输性能和可靠性方面的优势，并通过实验证明副本编码开销与传输开销的比值在可接受范围内。本文研究表明，编码机制的数据复制具有很强的实际应用价值，是值得进一步深入研究的网格复制技术方案。

今后的工作重点将会是编码机制的数据复制管理研究，传统的复制管理对于编码数据复制而言有其局限性，改进复制管理策略将是编码数据复制实用化的关键一环。

参 考 文 献

1 Deb S, Choutte C, M'edard M, et al. Data harvesting: A random coding approach to rapid dissemination and efficient storage of data:[M. I. T. LIDS Technical Report]. 2004
2 Kubiatowicz J, et al. Oceanstore: An architecture for global-scale persistent storage. In: Proceedings of ASPLOS, 2000
3 Chang F, et al. Myriad: Cost-effective Disaster Tolerance. In: Proceedings of FAST, 2002
4 Frolund S, Merchant A, Saito Y, et al. A decentralized algorithm for erasure-coded virtual disks. In: Proceedings of DSN, 2004
5 Weatherspoon H, Kubiatowicz J D. Erasure Coding vs Replication: A Quantitative Comparison. In:Peer-to-Peer Systems: First International Workshop, IPTPS 2002, LNCS2429, 2002
6 Zhang Z, Lian Q. Reperasure: Replication Protocol Using Erasure-code in Peer-to-Peer Storage Network. In: Proc. of the 21st IEEE Symp. on Reliable Distributed Systems (SRDS 2002), 2002
7 Acedanski S, Deb S, Medard M, et al. How Good is Random Linear Coding Based Distributed Networked Storage. In:Proceedings of First Workshop on Network Coding, 2005
8 Luby M, Mitzenmacher M, Shokrollahi A, et al. Practical loss-resilient codes. In: 29th Annual ACM Symposium on Theory of Computing, ACM, 1997
9 http://www.icsi.berkeley.edu/~luby/erasure.html
10 Byers J W, Luby M, Mitzenmacher M. Accessing multiple mirror sites in parallel: Using tornado codes to speed up downloads. In: IEEE INFOCOM, New York, 1999
11 Byers J W, Luby M, Mitzenmacher M, et al. A Digital Fountain Approach to Reliable Distribution of Bulk Data. In: Proceedings of ACM Sigcomm '98, Canada, 1998
12 Li S Y R, Yeung R W, Cai N. Linear network coding. IEEE Transactions on Information Theory, Februray 2003
13 Gkantsidis C, Rodriguez P. Network coding for large scale content distribution: [Technical Report MSR-TR-2004-80]. Microsoft Research, 2004
14 Lee Byoung-Dai, Weissman J B. An Adaptive Service Grid Architecture Using Dynamic Replica Management. GRID, 2001. 63～74
15 Chen P, et al. RAID: High-Performance, Reliable Secondary Storage [A]. ACM Computing Surveys, 1994

(上接第 100 页)

由表 1 可知 k 值对实际攻击的侦测率不会造成影响，但是会影响错误警告，因为当 K=3 时，子网络内的某台主机遭受攻击会连带影响其它主机的封包也一并被丢弃，造成错误警告增加。由表 2 得知 n 值越小越能快速侦测到攻击发生，但相对的错误警告会增加。借由提升 Rmax，我们可以消除错误警告的负面效果。在 n=50、Rmax=6 条件下会产生 750 个错误警告，因为当时的流量分布是 8 个 http 封包才有一个 ack 封包，又加上 n=50 使得快速累积到攻击门槛。由表 4 得知 α 值对侦测率与错误警告的影响很小，随着 α 值变大，错误警告会降低但侦测率也相对下降。

由以上的各种实验，得出一个重要结论，想要快速侦测攻击发生则 n 值越小越好，但相对的误判机率就会升高，此时需将 Rmax 值调高来减少误判。我们建议各个参数值的设定为 $K=4, n=50, Rmax\geqslant 8, \alpha\leqslant 10$。

小结 我们在阻断服务攻击侦测方面，结合 TOPS 的储存架构与连续假设实验的概念，使得平均侦测效果比原本的 TOPS 还好，演算法的实现与运算上也更容易。未来希望可以持续改进演算法，将演算法实现在硬件上，借由实际的封包测试来看整体的表现。

参 考 文 献

1 陈晶，崔国华，洪亮，付才. 一种 Ad Hoc 网络中的安全匿名按需路由协议. 计算机科学，2007,34(1)
2 陈云开，孙小林，马君华. 外汇领域的洗钱侦测系统及关键算法研究. 计算机科学，2007,34(3)
3 Jung J, Paxson V, Berger A W, et al. Fast Portscan Detection Using Sequential Hypothesis Testing
4 [美]Cole E 著.《黑客——攻击透析与防范》. 电子工业出版社，2002.2