基于 TCP 报文数量比值分布的网络异常快速预警

陈亮，龚俭，彭艳兵

(东南大学计算机科学与工程系，江苏 南京 210096)

lchen@njnet.edu.cn

摘　要：随着网络结构复杂程度越来越高，特别是近年来由于攻击频度的不断增长，攻击规模的不断扩大，对网络的监测和控制，以及对异常行为的监测也变得更为复杂与困难。因此，如何快速地发现网络上的异常行为显得更为重要。本文对 CERNET 江苏省省网主干的 TCP 标志报文的数量进行了统计，首先分析了一般情况下 TCP 各种标志报文数量的比值分布，在此基础上进行了对比，发现各种标志报文数量的统计比例在一般情况下（网络上无攻击等异常的发生）是相对稳定的，随后表明以这个相对稳定的比例作为一个指标，并快速衡量出以某种标志报文数量突变为现象的扫描攻击的可能性。最后以某一时间 CERNET 江苏省网主干 TCP 标志报文的实例对比了正常与异常的情况，并分析了异常情况的具体原因。

关键词：TCP、标志报文、比值分布、扫描攻击

中图分类号：TP393.2　　**文献标识码**：A

The quick way for the network abnormal behaviors upon the ratio distribution of the number of TCP packets

CHEN Liang，GONG Jian，PENG Yan-bing

（Computer Science Dept., Southeast University，Nanjing, Jiangsu Province, 210096）

Abstract：Along with the sharp increase in the bandwidth and application of Internet, the complexity of the net framework has become much higher. Especially in recent years, with the continuous growth of the frequency and the continuous spread of the scale that network attack took place, it has been more and more complicated to monitor and control the network and abnormal behaviors. Hence, it becomes more important to discover the abnormal behavior on the internet as soon as possible. This paper focuses on the statistics of the number of TCP flag-packets on the Jiangsu's trunk of CERNET. At first, it analyzes the ratio distribution of the numbers of flag-packets in the normal condition and makes a comparison upon that, which then educes that the ratio distribution is comparatively steady in the normal condition (no abnormal behaviors on the net). If using such comparatively steady ratio as a kind of index, it becomes possible for us to quickly detect the scan attack with the phenomenon of the number of a certain flag-packet has a great change. At last, this paper compares the normal condition with the abnormal condition upon an example of the number of TCP flag-packets on the Jiangsu's trunk of CERNET during a certain time, and then analyzes the detailed cause of such abnormal phenomenon.

Keywords：TCP, flag-packet, distribution of ratio, scan attack

[1]

收稿日期：2004-09-29

基金项目：973 大规模网络监测

1、引言

在长期与信息安全专家的较量中，黑客的攻击手法不断升级翻新，对开发隐蔽的计算机网络攻击技术更加得心应手，这不仅使攻击更难于察觉，也更难于防范。同时，随着入侵技术的发展，入侵或攻击的规模将不断扩大，技术也更加分布化[1]，对网络异常的监测就显得更为困难。如去年出现的“脉冲蛇神”以及“反射式 DDoS”等攻击形式[2]，以其攻击的隐蔽性和破坏性使得众多安全措施变得无效。于是，如何能更好的监测网络行为，及时的发现异常情况，是目前网络行为学重要的研究方向之一。

对于上述的大规模及分布式的网络攻击，现有的研究文献主要集中在滥用检测上，如 Kumar 在其博士毕业论文中所陈述的四类攻击特征[3]，东南大学的陆晟在其博士论文中描述的基于规则的入侵检测[4]，20CN 网络安全小组所提及的规则匹配[5]等。而这些方法或工具对于主干网络，由于性能或者价格的限制，没有太大价值。

Ohio 大学的 Marina Bykova 等人曾提出一种基于 TCP 报文统计的异常检测机制[6]。其统计 TCP 各种报文的数量以及 TTL、端口号等多种属性的分布情况及其在整体中所占的比重，并基于这样的结果分析网络的异常行为。这种方法可以在事后详细分析的过程中发现多种类型的攻击，但由于其需要统计的属性较多，发现异常的规则也比较复杂，若将其用在主干网上，会由于性能的问题而无法实现全部的规则。

下面首先定义一些本文用到的概念：

- 标志报文：TCP 首部码元字段中某位或某些位置 1 的 TCP 报文。
- SYN+ACK 报文：SYN 与 ACK 标志位同时置 1 的 TCP 报文，即 TCP 建立连接时的第二个报文。

TCP[7]是 Internet 中最基本最重要的传输层协议，其首部中定义的 6 个标志位[7]标识出了此 TCP 报文的类型。由 RFC793 中对建立和释放 TCP 连接的定义可知，SYN、SYN＋ACK、FIN、RST 报文与建立和释放 TCP 连接密切相关，它们在一定程度上体现了网络上流量的特性。但是由于 SYN 报文时常代表着扫描[9]与攻击[10]，短期的观察不能得到其的一般分布，将其引入分析往往会造成不正确的结果，因此本文不重点考虑 SYN 报文，只考虑 SYN＋ACK、FIN、RST 三者的关系。

本文通过对 CERNET 江苏省网主干的 TCP 标志报文数量的统计分析，发现如果分析其中的 SYN＋ACK、FIN、RST 标志报文数量之间的比值，在无异常的情况下，这个比值是相对稳定的，因此可以将这个比值作为指标来衡量出网络中某标志报文数量的变化，以便快速发现问题。

文章第二部分首先列出了对 CERNET 江苏省网主干观察 TCP 标志报文数量的统计结果，分析 RST 和 SYN+ACK 以及 FIN 报文数量的比值，说明在一般情况下（没有扫描攻击等异常的发生），比值分布是稳定的，并给出了比值的范围。第三部分以此比值的范围作为指标，对比网络的正常情况与异常情况，表明发现异常行为的可能性。第四部分详细分析文章第三部分中提及的异常，分析其发生的原因及异常的类型。

2、对 CERNET 主干 TCP 标志报文的观察

实验的数据采集自 CERNET 江苏省省网主干。主干网字节流速约 600Mbps，报文流速约 0.18Mpps，全天报文数量约 15G 个，观察时间为 2004 年 4 月 17 日全天。为了找出各标志报文的数量随时间的统计特性，以及其间比值的分布特性。首先统计出每个时间粒度内（5m）各标志报文的数量，分析其随时间分布的曲线；同时计算出此时间内报文数量的比值，分析其随时间分布的曲线。

为了获得各标志报文数量的比值随时间的分布，首先观察一下其各自随时间的分布情

况。图 1 中的三条曲线分别是 SYN+ACK、FIN、RST 报文的数量随时间分布的曲线。横坐标表示在当天的几点钟，纵坐标表示对应的标志报文的数量。

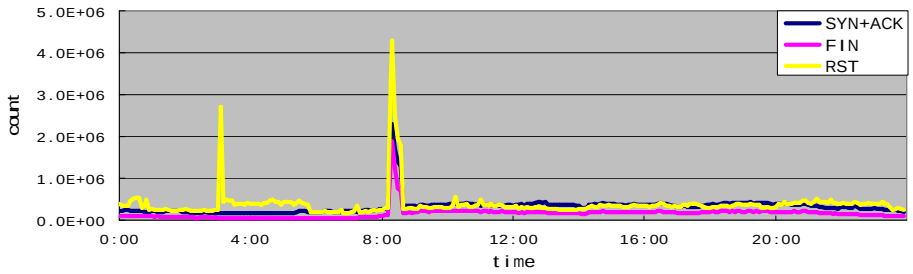

图 1 各标志报文数量随时间的分布曲线

从图 1 可见，各种标志报文的数量基本是同步增减的，这是网络正常的必要条件。但是图中有两个非常明显的尖峰值得引起注意：一个在凌晨三点左右，只是 RST 报文数量的突发；另一个在上午 8 点以后，表现为所有标志报文的数量全部增长。对于这两个尖峰，文章的第三部分会对其进行详细的分析。而对于其它较小的尖峰，本文不作讨论。

进而为了考察标志报文数量之间比值的分布，图 2 将 RST 报文与 SYN+ACK 以及 FIN 报文数量的比值随时间的分布做成曲线。选取 RST 报文数量作分子的原因有两点：1、从图 1 可见其数量较其他标志报文数量多，这样可以保证小数点前存在有效位数，便于观察比较；2、图 1 中三点钟的尖峰是 RST 数量的突增，将其作为分子有利于下文说明问题的方便。

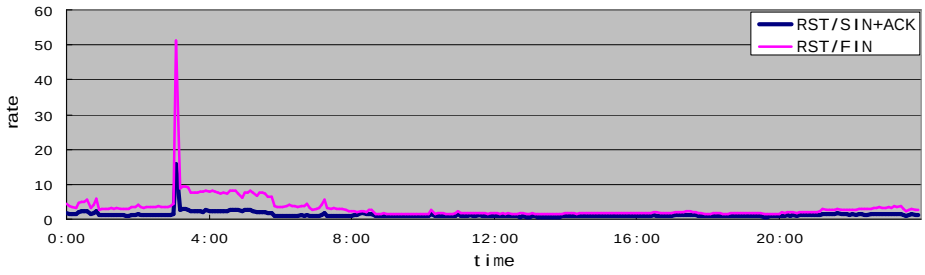

图 2 RST 与 FIN、SYN+ACK 报文数量比值随时间分布曲线

由图 2 可见，若除去三点钟的尖峰，RST 与其它两种报文数量比值的分布基本上是稳定的，而且两条曲线极为相似。这便意味着，SYN+ACK 与 FIN 是同步增减的，即二者的比值应该是稳定的。其结果见图 3 所示。

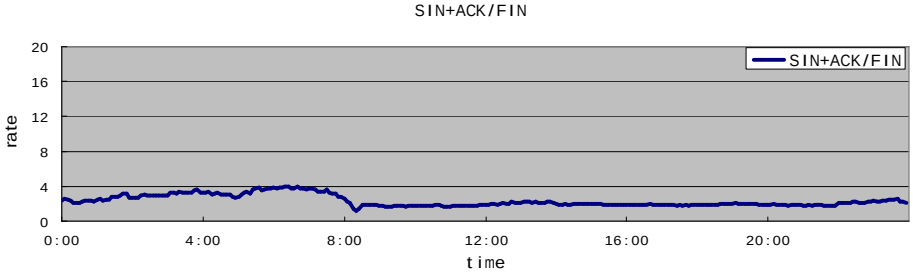

图 3 SYN+ACK 与 FIN 报文数量的比值随时间的分布曲线

图 3 中的曲线比图 2 中的曲线体现了更好的稳定性。如果我们暂时将图 2 中的尖峰作为

一种异常（下文中可以看到，其的确为一异常），将其忽略。那么图 2 和图 3 说明了：在正常情况下，网络上各种标志报文数量之间的比值是比较稳定的。

更重要的是，图 2 还能给出其在一般情况下的界限：RST/SIN+ACK∈[0，4]，RST/FIN∈[0，8]。于是可以将此范围作为指标，能够快速的检测在某种报文数量的突增或突减。例如，如果 RST/SYN+ACK 与 RST/FIN 同时增加，并且增加的倍数相近或者 SYN+ACK/FIN 没有大的变化，一般可以说 RST 报文数量增加了。如果 RST/SIN+ACK 减小，但 RST/FIN 没有变化，可以认为是 SIN+ACK 报文数量增加了。更进一步的可参看 SIN+ACK 与 RST 和 FIN 的比值得出结论。

由上述实验及论证可知：在一般情况下，网络上的各种标志报文数量的比值是相对稳定的。于是当某种报文的数量突然增加时，网络上就可能出现了异常情况。以上述的简便方法来监测网络，可以及时的发现某些异常，给后继工作提供方便。

3、对正常与异常情况的分析

文章第 2 部分指出，如果将一般情况下标志报文数量的比值范围作为一个指标，可以监测出网络上某种报文数量的突变，以快速地监测网络的异常行为。在下文中，我们取文章第 2 部分实验得出的比值范围作为监测异常的指标，即正常情况下，RST/SIN+ACK∈[0，4]，RST/FIN∈[0，8]。对图 1 中八点及三点的尖峰进行分析，以检验本文所提出的模型能否正确的识别出正常与异常行为。

3．1、正常情况的流量突发

首先考虑 8 点钟的尖峰。从图 2 及图 3 可见，八点以后各标志报文数量的比值并没有变化，所以唯一的解释是所有的标志报文数量都增加了，而且增加了相同的比例，即网络上的正常流有个突增，才会造成图 1 中所出现的尖峰。表 1 列出各时段报文数量的对比情况。

表 1 8 点之后半小时与其余时间报文数量对比情况

报文类型	SYN	SYN+ACK	FIN	RST
全 天 数 量	889,029,537	93,486,184	43,785,199	104,870,336
半小时的均值	18,521,449	1,947,629	912,192	2,184,799
7：30~8：00 的数量	89,929,932	1,443,765	564,933	1,332,867
8：12~8：42 的数量	123,913,329	8,839,403	3,691,841	8,522,347
较均值的倍数	6.7	4.5	4.1	3.9
较前半小时的倍数	1.4	6.1	6.5	6.4

表 1 的数据是对全天标志报文段的数量对半小时取的平均、8：00 之前的半小时以及从 8：12～8：42 这半个小时内标志报文数量的对比情况。其中 SYN+ACK 报文及 FIN 报文的数量及其倍数可以证明：在 8：12～8：42 这段时间之内，网络流的数量较全天的平均增加了 4 倍左右，较其前半个小时增加了 6 倍左右。

进一步，下文从端口使用情况的分布说明此次突发完全是属于正常情况。表 2 是全天端口使用情况的前十位以及从 8：12~8：42 这半小时内端口使用情况的前十位以及较均值的倍数（“—”表示此端口倍数对说明问题无意义，忽略）：

表 2 全天与 8 点之后半小时端口使用情况的对比

TOP 10	全天半小时的均值	8：12 ~ 8：42

	端口号	数量	端口号	数量	较均值的倍数
1	80	1,417,310	80	5,472,171	3.9
2	21	171,934	21	2,020,227	11.8
3	25	48,139	22	1,027,934	45.8
4	8080	26,988	25	72,778	1.5
5	22	22,446	8080	21,047	0.80
6	443	16,732	443	15,842	0.95
7	554	12,589	554	14,447	1.15
8	5354	8,280	1274	11,318	—
9	65350	8,011	1456	11,087	—
10	6881	6,071	1345	10,960	—

由表 2 可见：排位在前的端口都是一些常用端口，两组数据相比，其顺序基本没有改变，只是在数量上有很大的增长，如 HTTP 端口（80）为均值的 3.9 倍，FTP 控制端口（21）为均值的 11.8 倍，邮件端口（25）为均值的 1.6 倍。而 SSH 端口（22）则高于均值 45.8 倍。由于怀疑 22 端口存在着异常，对其组流，发现确实存在这么多的完整流。这一结论也可以从表 3 中各个标志报文的数量统计中近似得出。

表 3　　8 点之后半小时 22 端口的报文数量

Type of packet	SYN	SYN+ACK	FIN	RST	Data
Count	3,466,850	1,027,934	1,032,025	1,907,909	3,391,111

SYN+ACK 和 FIN 的数量以及其相近的程度都表明，8 点之后的半小时内 22 端口的流的确有了明显的增加。这完全可以解释为人的行为对网络流量的影响。因为这段时间正是大家开始上班的时间，利用网络开始一天的工作，例如上网查资料、收发邮件、远程登录等。最好的证明就是 80、21、25、22 端口这些常用端口的流量上升，而排位在 5 名之后的一些并不常用的端口流量并没有太大的变化。

以上论证充分证明了：即使网络上的流量增加了，但只要都是正常的流，比值分析的方法仍然不会将其误认为是异常情况，不会出现误判。

3．2、某种标志报文数量突发的异常情况

图 1 中凌晨 3 点的尖峰表明：在那个时刻，RST 报文数量有了一个很大的增长。图 2 中的比值分布以及下表中比值数据的对比也证明了这一点。

表 4　　3：00~3：15 之间 RST 与其它报文数量的比值与均值相较的情况

Type	Average of day	In 3:00~3:15	**Times**
RST/SYN	0.0218	0.1067	**4.9**
RST/SYN+ACK	1.2707	9.2298	**7.3**
RST/FIN	3.2619	29.9499	**9.2**

表 4 中倍数（Times）一列的数据表明：在 3：00~3：15 这段时间里，相较其它的标志报文，网络上 RST 报文的数量有了很大的增长。若取文章第二部分中提及的比值（RST/SIN+ACK∈[0，4]，RST/FIN∈[0，8]）作为衡量的指标，这里我们可以将其预警为一种异常。在下一部分中我们将通过几组数据的对比证明我们对此异常的假设，并揭示出

3：00~3：15 之间网络上发生的是何种异常。

4、扫描异常的实例分析

表 5 是全天 RST 报文端口使用情况对 15 分钟取平均之后的排序结果，以及与 3：00~3：15 之间 RST 报文端口的对比，只取代表性的前五位。表中最后一行是 RST 报文数量的对比。

表 5　　3：00~3：15 之间 RST 报文端口使用情况与全天均值的对比

TOP 5	全天 15 分钟的均值		3：00 ~ 3：15		
	端口号	数量	端口号	数量	较均值的倍数
1	80	565,730	**21**	**2,421,062**	**23**
2	21	105,235	80	483,182	0.85
3	40000	53,400	445	115,531	2.5
4	445	45,487	135	101,092	2.5
5	135	40,564	40000	38,844	0.73
TOTAL		1,155,930		3,414,399	2.96

表 5 中最后一行的 RST 报文数量的对比表明：在 3：00~3：15 这 15 分钟内 RST 报文段的数量为均值的 3 倍。而加粗的数据表明：21 端口 RST 报文段的数量又占这 15 分钟 RST 总数量的 71%（2,421,062/3,414,399），为其均值的 23 倍。已经可以初步看出 21 端口的异常。为了进一步分析问题，将源宿端口的使用情况分开考虑，宿端口的情况较均值没有明显的变化，在此不予列出。

表 6　　3：00~3：15 之间 RST 报文源端口使用情况与全天均值的对比

TOP 5	全天 15 分钟的均值		3：00 ~ 3：15		
	端口号	数量	端口号	数量	较均值的倍数
1	80	203,659	**21**	**2,411,171**	**35.5**
2	**21**	**67,887**	80	253,953	1.2
3	40000	53,397	40000	38,844	0.73
4	25	10,252	33225	37,773	—
5	6881	9,856	2248	29,614	—

表 6 中加粗数据的对比更突出了 21 端口的异常。若将表 6 中 3 点源端口排序的结果与表 5 中 3 点端口排序结果（不区分源宿端口）相比较，可以看出 3 点钟 21 端口 RST 报文有 99.6%（2,411,171/2,421,062）是从 21 端口发出的，即从 3：00 到 3：15 的这段时间内 RST 报文有 70%(2,411,171/3,414,399)是从 21 端口发出的。我们将 RST 报文数量的均值（1,155,930）加上从 21 端口发出的 RST 报文的数量（2,411,171），其结果（3,567,101）近似的等于 3：00~3：15 之间 RST 报文的数量（3,414,399）。再根据其余端口的使用情况较均值并无显著变化，可以确定：这超出的数量是由 21 端口发出的 RST 引起的。也就是说，这次突发的 RST 报文完全是由 21 端口发出的。

进一步，这些源端口为 21 的 RST 报文是由哪些地址所发出的？表 7 将源端口为 21 的 RST 报文段的源 IP 列出，其中我们将四字节的 IP 地址视为无符号长整型：

表 7　　3：00~3：15 源端口为 21 的 RST 报文段源地址的统计

IPAddress	count	IPAddress	count	IPAddress	count	IPAddress	Count
325xxxxx	1	3263xxxxx	1	3356xxxxx	1	5557xxxxx	1
8239xxxxx	1	23898xxxxx	1	26230xxxxx	1	33044xxxxx	1
42045xxxxx	1	799xxxxx	2	14894xxxxx	2	16359xxxxx	2
6456xxxxx	3	8390xxxxx	3	19134xxxxx	3	24580xxxxx	3
4019xxxxx	4	38304xxxxx	4	4278xxxxx	5	6029xxxxx	5
404xxxxx	6	2565xxxxx	6	16715xxxxx	6	1693xxxxx	7
3076xxxxx	7	1762xxxxx	9	3323xxxxx	9	1776xxxxx	12
4110xxxxx	13	20178xxxxx	18	1594xxxxx	34	2768xxxxx	51
3104xxxxx	52	1426xxxxx	53	2389xxxxx	74	415xxxxx	235
10229xxxxx	308	27897xxxxx	310	23559xxxxx	461	31749xxxxx	906
38639xxxxx	1,205	267xxxxx	1,308	17300xxxxx	5,106	3733xxxxx	14,169
33953xxxxx	**2,386,761**						

问题进一步明朗了，IP 地址为 33953xxxxx 的机器从 21 端口发出的 RST 报文占源端口为 21 的 RST 报文段数量的 99%（2,386,761/2,411,171）。源地址如此的集中，但经过分析，此源地址发出的 RST 报文的宿地址非常的均匀。考虑到 RST 报文集中在某一源地址的 21 端口，可以初步认为其是一个异常情况。而其异常具体有两种可能：DOS 攻击和利用 21 端口的扫描。

在 DOS 攻击中存在一种利用 RST 位的 IP 欺骗 DOS 攻击[8]。如果是发生了这种情况的话，在从地址 33953xxxxx 的 21 端口回送 RST 报文之前，应该收到从别的地址发来的至少是等量的 RST 报文。而如果是利用 21 端口扫描的话，那么在地址 33953xxxxx 的 21 端口发出 RST 报文之后，应该收到别的地址返回的 RST 报文。其数量理论上应该是等量的，但两种情况的存在会大大减少返回的 RST 报文的数量：一是扫描到了不存在的地址或网段，这时路由器会回送给源主机 ICMP 报文；二是虽然地址存在，但由于该地址的防火墙能识别此种扫描，将其屏蔽掉。

虽然 CERNET 为了管理的方便，管理员将 ICMP 功能关闭，致使无法从 ICMP 报文判断异常情况，但要区分这两种异常很简单，就是要看看在地址 33953xxxxx 发出 RST 报文的前后，有没有到该地址的报文，特别是 RST 报文，以及其数量。因为如果是 DOS 攻击，到该地址的 RST 报文应该在该地址发出 RST 之前；而如果是扫描，那么到该地址的 RST 报文应该在该地址发出 RST 之后。

对 2：15~4：00 这段时间内地址为 33953xxxxx 的报文进行组流，并分析到达该地址的报文的情况。结果是：在 RST 爆发之前，到达地址 33953xxxxx 的报文数为 0，而在地址 33953xxxxx 发出 RST 报文之后，有为数不多的从各个地址回应的 RST 报文。这也可以从图 4 中看出。

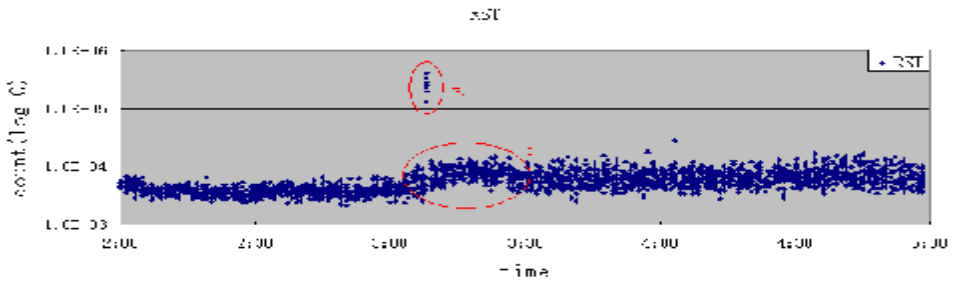

图 4　　2：00~5：00RST 报文段数量的分布

从图 4 中可以看出，紧接着 3 点 10 分（a 点）的 RST 突发之后，RST 报文的数量还是较均值有一个从上涨到下降的过程（b 点），虽然在图中看不出来其宿地址为何，但组流结果已经证明，这些增多的 RST 报文确实很大一部分是发往地址 33953xxxxx 的。虽然较从 33953xxxxx 发出的 RST 相比要少得多，但已经足可以否定我们对 DOS 攻击的假设，证明结论：在大约 3 点 10 分左右，IP 地址为 33953xxxxx 的机器利用了 21 端口进行了一次扫描。

这样的结果进一步证明了：只要是由于网络上的某种异常而引发的某个标志报文数量的变化，用比值分析的方法都可以迅速的发现异常，提出预警。当然，利用这种比值分布的方法只能快速的预警，而如果需要知道究竟是何种异常，还需要后继工作的进一步研究。

5、结果与讨论

本文分析了网络上 TCP 流标志报文数量的统计特性，发现在正常情况下各种标志报文数量的比值是相对稳定的，可以将正常情况下的比值范围作为一个指标，以后的工作中可以用这个指标来快速预警某种标志报文数量发生突变的情况，以此来对网络的异常行为如扫描攻击、IP 欺骗 DOS 攻击等进行预测。因为这些异常都有一个共同的特点：当它们发生的时候，网络上某种标志报文的数量会发生很大的变化。为了发现问题的方便，这里建议对于想观察变化的标志报文，将其数量作为比值的分子，而将其余的标志报文作为分母。

更重要的是，这种方法可以以较小的代价检测出引言中提到的“脉冲蛇神”、“反射式 DDoS”等大规模、分布式的入侵与攻击。虽然这种类型的攻击较隐蔽，但在整个网络上的表现形式却是一致的：某种标志报文会大量的增长。只要能识别出这样的现象，后继的工作便可有目的的展开。

当然，这种方法的缺点只能快速的提出预警，如果需要知道究竟是何种异常，还需要后继工作的进一步研究。对于本文提出的比值监测异常的方法中，部分模型还有待将来完善，例如正常情况下比值的范围还需要对更大的时间粒度下的网络进行观察，以得出一个更为合适的经验值。对于这个经验值的自适应模型也是下一步的研究重点。在实用阶段，也可以不断的学习以完善模型。

参考文献：

[1] 匿名. 入侵检测技术综述. http://www.csai.cn/net/idsall.htm

[2] CNNS. 网络攻击机制和技术发展综述. http://it.rising.com.cn/newSite/Channels/Safety/SysSafety/Safe_Other/200302/27-100809862.htm

[3] Sandeep Kumar. Classification and Detection of Computer Intrusion[D]. [Doctor Thesis]. Indiana:Purdue University. Computer Science. 1995.

[4] 陆晟．基于规则的高速网络入侵监测．[Doctor Thesis]. 中国南京：东南大学计算机系．2003．

[5] 秋阳. 监测分布式拒绝服务. http://www.20cn.net/ns/hk/hacker/data/20040111171439.htm

[6] Marina Bykova, Shawn, Brett Tjaden. “Detecting Network Intrusions via a Statistical Analysis of Network Packet Characteristic”. School of Electrical Engineering & Computer Science, Ohio University. 2001

[7] RFC 793. Transmission Control Protocol.

[8] refdom. 防不胜防——DOS 攻击. http://www.xfocus.net/articles/200111/297.html

[9] 李海翔, 方睿, 李祥和, 芦康俊. 网络隐患扫描理论及其实践. http://www.nsc.org.cn/disp_article.asp?AE_ACID=457

[10] xundi. SYN Flood 攻击的基本原理及防御. http://www.xfocus.net/articles/200106/208.html

作者简介：

陈亮（1981-），男，江苏南京人，南京东南大学硕士生，主要研究方向为网络行为学。

彭艳兵（1974-），男，湖北洪湖人，东南大学计算机科学与工程系博士生，主要研究方向为网络行为学。

龚俭（1957—），男，上海人，工学博士，东南大学计算机系教授、博导，主要研究方向包括网络管理、网络安全、分布式系统等。