基于 IP 报文 Identification 标识的网络异常流量检测

周明中　龚俭　丁伟　程光

（东南大学计算机系华东北地区网络中心　江苏南京 210096）

摘　要：由于相当一部分异常流量由于采用了特殊的生成机制而在结构上有别于遵循基本网络协议的正常流量，本文提出了一种基于 IP 报文 Identification 标识字段分布识别网络中异常流量的方法。通过 CERNET 网络不同时段的 IP 报文检测结果证明了该方法的准确性。

关键词：Identification 标识；二项分布；异常流量检测

Abnormal Network Traffic Detection based on Identification Mark of IP Packet

ZHOU Mingzhong, GONG Jian, DING Wei, CHWNG Guang

（Northeast Network Center, Department of Computer Science, Southeast Univ., Jiangsu, Nanjing 210096 China）

Abstract *This paper proposes a new method of abnormal network traffic based on the distribution of IP packets' Identification because many of abnormal network traffics are generated by special mechanisms, which are different from the ordinary traffics created on the basic network protocols. The correctness of this method is proved by the results of IP packets detect with different time on CERNET.*

Keywords *Identification mark; Binomial distribution; Abnormal traffic detection*

1　引言

随着 Internet 的发展，网络流量急剧增长，由于网络的发展具有一定的规律性，可以通过对网络协议的分析和网络流量的预测定义网络流量的正常行为，当观测所得的流量行为偏离正常时，对网络流量的进一步分析可能发现异常的原因。目前基于网络主干和边界的异常流量检测研究主要集中在流矩阵，报文分析等方面，但由于基于网络的探测，入侵和攻击行为也变得越来越普遍和复杂，这些方法在测量规模和粒度上不能达到很好平衡。本文提出了一种基于 IP 报文 Identification 标识字段分布的异常流量识别方法，可以以较小的代价有效地识别网络中的流量异常，适用于主干网络和边界网络，并通过实验验证了其可行性。

2　问题提出

目前 Internet 绝大部分使用 TCP/IP 协议簇进行网络传输，而 IP 协议是其中最重要也是最基本的协议。位于应用层的协议通过将服务内容切割成分片（fragment）的形式传递给 TCP 层协议，在 TCP 层加上相应的头部信息又传递给 IP 层。由于在接受端需要对分片进行重组获得完整的服务内容，而网络的延时、拥塞和报文本身的传输方式都可能导致分片的乱序，所以需要对 IP 报文进行标识。在 IP 协议[1]中 Identification 字段用于标识该报文而区别于来自相同源宿地址对使用同一个协议的其他报文。由于该字段被定义为 16bit 长，也就是说它所能表示最大数目为 2^{16} 即 65536。为保证服务的正常，网络中必须确保来自同一 IP 地址使用相同协议的报文应当有其唯一的 Identification 标识（该标识值位于 0-65535 之间）。

在文献[1]中并没有给出 Identification 标识的具体取值方式，但由于规定了其取值范围，采用不同取值方式的主机所选取的初始 Identification 应当是一个位于 0-65535 之间的随机数且取值相互独立，而大部分

服务被分解为若干个 IP 报文进行传送，在每个主机中都维护一个计数器（Counter），每发送一个 IP 报文该计数器加 1。可以做出以下假设：

假 设 1 在较大规模网络中，从宏观的角度分析 Identification 标识的取值是近似均匀分布的。

经过大量实验证明，在绝大多数情况下，Identification 标识是近似均匀分布的，有关实验的验证将在下节中具体介绍。在假设 1 的基础上，根据 Identification 标识的选取方式可以做出相关结论如下：

引 理 1 在较大规模的网络中用于正常服务的 IP 报文的 Identification 标识是近似服从参数为 n，p 的二项分布，其中 n 为 65536，p 为 0.5。

证明：由于每个源主机所发送的 Identification 标识是随机或者采用一定的机制选取的，而每个源主机选取的方式是相互独立的。设每个 IP 报文所对应的 Identification 为随机取以下值之一：X_1＝0，X_2＝1，…，X_n＝n－1（n 为 65536），它们的取值服从同一（0-1）分布，其分布率为：

$$P\{X_i = 0\} = 1 - p = 0.5, \qquad P\{X_i = 1\} = p = 0.5$$

则所有 IP 报文的 Identification 标识的分布为：

$$X = a1X_1 + a_2X_2 + \mathbf{L}\,\mathbf{L} + a_nX_n)$$

其中 a_1，a_2 ….. a_n 是取特定 Identification 值的 IP 报文数。由假设 1 可以得出以下结论：

a_1＝a_2＝….. ＝a_n=a

则

$$X = a(X_1 + X_2 + \mathbf{L}\,\mathbf{L} + X_n)$$

已知 X_1 + X_2 + …… + X_n 服从二项分布，那么比较容易证明 X 是近似服从参数为 n，p 的二项分布。根据二项分布的特性，很容易推出——任一随机变量的数学期望是 $\frac{65535-0}{2}$，方差为 $\frac{65535-0}{4}$。但是由于网络中还存在相当一部分异常 IP 报文，它们的主要来源之一是人为构造的攻击报文。这些异常 IP 报文和正常 IP 报文共同构成了网络中存在的 IP 报文。

引 理 2 在网络中实际观测到的 IP 报文 Identification 标识的分布应当是正常 IP 报文分布和异常 IP 报文分布的叠加。

因此，可以通过区分这两部分分布，有效地识别网络中可能存在的流量异常，为其他检测方法（如报文分析）提供预警，从而为网络行为分析、入侵检测等提供必要的依据。

3 基于标识字段的异常检测方法

本文基于 IP 报文 Identification 标识的角度将 IP 报文分为正常 IP 报文和异常 IP 报文，整个网络的流量也是由其分别对应的正常流量和异常流量构成的。从较大规模网络（主干网络和部分局域网）的角度分析，正常 IP 报文 Identification 标识的分布规律由引理 1 可知近似为二项分布，而异常报文的分布具有随机性和多样性，也就导致了其的不可预测性。但是根据引理 2，可以通过绘制网络流量曲线并从中分离出正常流量，就可以得到目前网络中异常流量的曲线，然后对这些流量的进一步分析就可以获得或部分获得流量异常的原因。

将获取的所有 IP 报文不同 Identification 标识数量的非空有限集合定义为 P， 则根据 IP 协议的定义可

知：$P=\{p(i)\mid i\in(0,\mathbf{L},65535)\}$，其中隶属于正常 IP 报文的集合定义为 $P_a=\{p_a(j)\mid j\in(0,\mathbf{L},65535)\}$，异常 IP 报文的集合为 P_b，$P=P_a \cup P_b$。

根据引理 1 可知，在集合 P_a 中 j 应服从参数为（n_a，0.5）的二项分布，则集合中的元素 p_a（x）的值应当基本等于其均值 $\frac{n_a}{65536}$。采用报文总数 *n* 乘以一个预定义的比例 *r* 来估计正常 IP 报文的数量，由于在非极端情况下，正常 IP 报文数量占报文总数量的绝大多数，所以 r 值的估计偏差比较小，还可以根据网络当前的状况动态地调整 r 的取值。这样就可以从 IP 报文集合里分离出正常 IP 报文集合，从而获得异常 IP 报文结合，然后根据异常 IP 报文集合中元素的分布状况给进一步分析提供依据。

对异常 IP 报文的分析，主要通过定义一个阈值（$F_{threshold}$）来将可能存在的网络异常流量从其他原因所引起的噪声区别出来，这个阈值可以设定初始值然后根据识别结果动态修正。

基于 IP 报文 Identification 标识的异常流量发现算法

FindAbnormalTraffic（P, n, r, $F_{threshold}$）

Begin

$n_a = n \times r$

$$n_{mean}=\frac{n_a}{65536}$$

$$P_a=\{p_a(j)=n_{mean}\mid j\in(0,\mathbf{L},65535)\}$$

$P_b = P - P_a$

for *value*$\in P_b$ and *value*$> n_{mean}\times F_{threshold}$

add value to $P_{abnormal}$

Return $\{P_{abnormal}\}$

end

通过大量的实验证明，在一般情况下，异常 IP 报文的 Identification 集中在 0 附近，有时还出现个别标识的报文数量偏移平均值较远。

4 流量行为的实例分析

针对 CERNET 主干网络的长期观测结果显示，在大多数情况下，Identification 标识的分布是十分均匀的，具有某个特定 Identification 值的 IP 报文数量均在基于标识的平均报文数量附近。从整个分布曲线来看，IP 报文数是围绕平均值作平稳的小幅振动，这也证实了第 2 节所提出的假设 1。

观测结果还发现，在所有观测时段中，Identification 标识为 0 的 IP 报文数量远远高于平均值，这与文献[2]中实验观察结果相一致，这不符合正常 IP 报文均匀分布这个论断，所以在标识为 0 的 IP 报文中有可能大量存在非正常报文。对 Identification 标识为 0 的 IP 报文进行分析发现，有大量相同源宿 IP 和相同端口的 IP 报文在短时间内重复出现，所以导致了标识为 0 的 IP 报文数量大大超过平均值，在剔除了这些异

常报文之后，所剩的 IP 报文数量非常接近于平均值。由此可见，标识为 0 的 IP 报文数量异常的主要原因是因为大量存在这些异常 IP 报文。对实验观测所得的其他标识的 IP 报文数量异常进行进一步分析，发现结果与此相类似。在实验中还发现相当数量的来自同一源 IP 和源端口对应不同宿 IP 的相同宿端口的 IP 报文，这是典型的扫描攻击的表现。

本文对不同时段在 CERNET 主干网采集的 IP 报文进行分析（每个时段持续 10 分钟），分析所得报文分布曲线如图 1 所示。

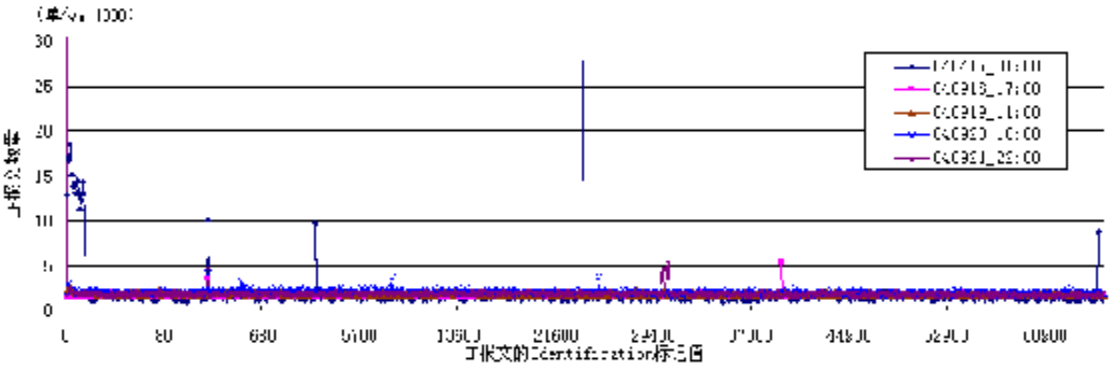

图 1 不同时段基于 Identification 标识的 IP 报文数量分布曲线

选取参数 r＝0.98， 获得各个时段对应正常报文集合 P_a，并从总体 IP 报文集合中去除正常 IP 报文集合获得异常报文数量分布曲线如图 2 所示。

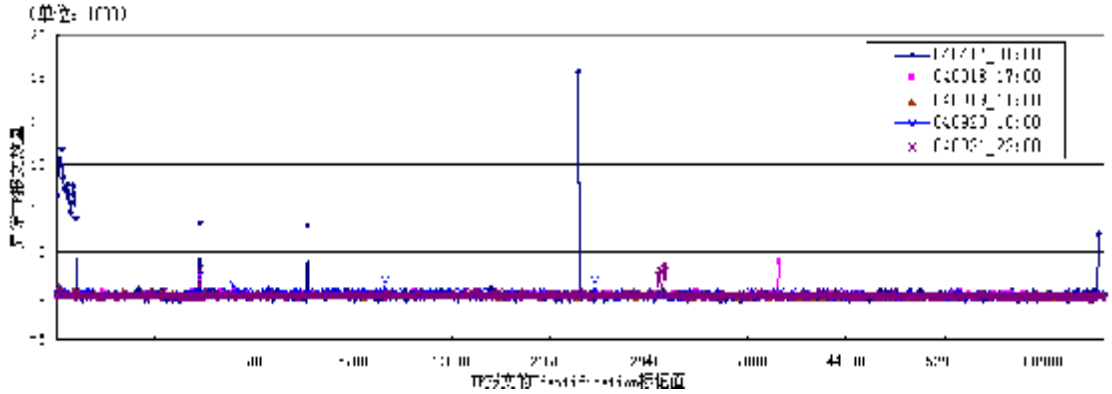

图 2 不同时段基于 Identification 标识的异常 IP 报文数量分布曲线

采用初始阈值 $F_{threshold}$＝1，除去了可能存在的网络其他噪声后，可以得到在 Identification 标识为若干特定值的 IP 报文中可能存在相当数量的异常报文，从而为进一步的报文分析提供预警。实验数据显示，在 2004 年 4 月 17 日 1：00 和 2004 年 9 月 21 日 22：00 在网络中存在一定的流量异常。

5 结论

本文提出了一种新型的基于 IP 报文 Identification 字段进行网络异常流量发现和分析方法，该方法的主要优点在于算法简单，所占用的系统资源较小，误报率低，可以较方便地嵌入到目前流量检测工具中和其他报文分析方法及工具结合使用等等。但是由于其观测的对象所限，该方法并不能有效地发现伪装成正常 IP 报文的异常流量，它必须和其他报文分析工具配合使用才能达到最佳的效果。

参考文献

[1] DARPA Internet Program Protocol Specification. Internet Protocol. Information Sciences Institute University

of Southern California. 1981【RFC791】.

[2] 程光。大规模高速 IP 网络流量抽样测量及行为分析研究，东南大学博士论文，2003 年 1 月：41-44。

[3] 程光，龚俭，丁伟。基于抽样测量的高速网络实时异常检测模型，软件学报，2003.14（Vol.14, No.3）：594-599。

[4] 邹柏贤。一种网络异常实时检测方法，计算机学报，2003 年 8 月（Vol. 26, No. 8）：940-947。

[5] 高艳，管晓宏，孙国基等。基于实时击键序列的主机入侵检测，计算机学报，2004 年 3 月，（Vol.27，No. 3）：396-401。

[6] 徐永红，杨云等。基于权重包标记策略的 IP 跟踪技术研究， 计算机学报，2003 年 11 月，（Vol. 27，No.11）：1598-1603。

【作者简介】

周明中，东南大学计算机系博士研究生，主要研究方向：网络安全和网络行为学。

龚　俭，东南大学计算机系教授，博导，主要研究方向：大规模网络的入侵检测，网络行为学，计算机体系结构。

丁　伟，东南大学计算机系教授，博导，主要研究方向：网络行为学，计算机体系结构。

程　光，东南大学计算机系博士，讲师，主要研究方向：网络行为学。