基于抽样测量的高速网络实时异常检测模型*

程 光, 龚 俭, 丁 伟

(东南大学 计算机科学与工程系,江苏 南京 210096)

E-mail: gcheng@njnet.edu.cn

http://www.njnet.edu.cn

摘要: 实时异常检测是目前网络安全的研究热点。文章基于大规模网络流量的统计特征，寻找能够评价网络行为的稳定测度并建立抽样测量模型。基于中心极限理论和假设检验理论，建立网络流量异常行为实时检测模型。最后定义 ICMP 请求报文和应答报文之间比率的网络行为测度,并实现对 CERNET 网络 ICMP 扫描攻击的实时检测。文章提出的方法和思路对其它网络安全检测研究具有一定的指导意义。

关 键 词: 抽样测量;测度;异常检测;滑动窗口;高速网络

中图法分类号: TP393 文献标识码: A

随着 Internet 的普及，联网计算机的数量迅速增加，网络入侵问题也随之突出，因此专门面向网络入侵检测的网络安全监测系统越来越受到关注。异常检测关键是通过对网络流量正常行为的描述来分析和发现网络或系统中可能出现的异常行为，并向管理员提出警告，或主动做出反应。网络的异常行为通常表现为通过流量的异常，例如由特定的攻击程序或蠕虫爆发所引起的突发流量行为，这种流量异常行为的特点是发作突然，先兆特征未知或比较隐蔽，因此实时监测与响应是防范这类攻击的重要手段。

异常检测的核心问题是如何实现流量正常行为的描述，检测的实时性，获得信息的全面性和反应的灵敏性，因而使系统设计和实现难度较大，所以面向网络的实时安全监测系统是目前研究的一个热点。实时异常检测的前提是能够实时测量，对大规模高速网络流量进行异常检测首先要面临高速流量荷载问题，由于测量、分析和存储等计算机资源的限制，无法实现全网络流量的实时检测，因此，抽样测量技术成为高速网络流量测量的研究重点。

文章首先提出从大规模网络流量数据统计分析中寻找能够描述其正常行为的稳定测度标准。研究 IP 报头不同字段统计的随机性，提出并建立基于报文标识的抽样掩码实时测量模型。由于网络用户行为随时间变化而变化，流量正常行为也具有时间性，系统维护一个历史窗口描述正常行为，实现网络行为的实时更新。同时维护一个描述网络当前网络行为的窗口，通过中心极限理论和正态分布假设检验实现对当前流量行为的异常检验。文章最后定义 ICMP 请求报文和应答报文之间比率的网络行为测度，实现对 CERNET 网络流量 ICMP 扫描攻击的实时检测。

1 异常检测测度

异常检测方法主要有：统计异常检测法[1]、基于机器学习的异常检测方法[2]、基于数据挖掘的异常检测法[3]和基于神经网络的异常检测法[4]等。在异常检测中统计模型中常用的测量测度包括审计事件的数量、间隔时间、资源消耗等。Denning[1]提出了用于异常检测的 5 种统计模型：(1) 操作模型：该模型假设异常可通

* 收稿日期: 2002-02-25; 修改日期: 2002-04-26

基金项目: 国家自然科学基金资助项目(90104031); 国家 863 高科技发展计划资助项目(2001AA112060)

作者简介: 程光(1973－),男,安徽黄山人,博士生,主要研究领域为网络行为学;龚俭(1957－),男,上海人,博士,教授,博士生导师,主要研究领域为网络安全.丁伟(1962－),女,江苏南京人,博士,教授,主要研究领域为计算机网络应用

过测量结果和指标相比较得到，指标可以根据经验或一段时间的统计平均得到。(2) 方差：计算参数的方差，设定其置信区间，当测量值超出了置信区间的范围时表明可能存在异常。(3) 多元模型：操作模型的扩展，通过同时分析多个参数实现检测。(4) 马尔可夫过程模型：将每种类型事件定义为系统状态，用状态转移矩阵来表示状态的变化，若对应于发生事件的状态转移矩阵概率较小，则该事件可能是异常事件。(5) 时间序列模型：将测度按时间排序，如一新事件在该时间发生的概率较低，则该事件可能是异常事件。

假设正常流量行为特征是 $s(t)$，异常流量特征是 $n(t)$，故实际测量到的流量行为特征是 $y(t)=s(t)+n(t)$。文章使用基于统计的方法来实现异常检测，确定网络流量历史行为测度框架检测当前的入侵活动。历史行为测度框架是由一组测度统计值构成，在系统运行时，异常检测系统统计测量当前流量行为测度框架，并同历史行为测度框架相比较，同时更新历史行为测度框架，当两种测度框架出现明显的偏离即认为出现了异常行为，并可进一步引发入侵检测分析。因此，为了实现统计异常检验，需要先给出下面假设：

假设 1．异常流量行为 $n(t)$ 是一种偶然行为，$n(t)$ 和 $s(t)$ 具有明显差异。

根据假设 1，$n(t)$ 可以通过统计数学工具分离并加以识别。例如扫描，作为白噪声的扫描现象总是存在，但当扫描数量出现异常时，可能预示着蠕虫或 DDOS 的爆发。假设 1 也可能会存在一些不适应的情况：1）如果异常流量行为 $n(t)$ 不是一种偶然行为，那么统计分析过程中，异常行为将会作为正常行为；2）如果 n（t）和 s（t）没有明显差异，异常行为特征 n（t）被正常行为 s（t）所掩盖，不能被检测。这两类问题会导致误判和漏判的发生。

文章的异常检测系统运行在 CERNET 2.5Gbps 高速主干网络环境中，主要考虑大规模网络流量异常行为的检测；而 Denning [1] 定义的测度框架主要针对局域网和主机，不适合高速主干异常检测的研究，因此需要定义新的测度框架。文章主要研究会产生异常流量的攻击行为，如端口扫描、flood 型 DOS 攻击和蠕虫[5]等。

由于网络中多数路由器对外封锁 ICMP 操作，因此 ICMP 的请求报文比应答报文多数倍。通过统计发现，正常情况下，两者比值在 10 以内，一旦发生 ICMP 扫描攻击，两者比值甚至高达接近 1000。图 1 显示了 CERNET 网络某日内受到的 ICMP 扫描攻击的观测结果，可以看出，ICMP 请求报文和应答报文的比值基本稳定在 10 以下，但其中出现 2 次大规模 ICMP 扫描攻击使得两者之间比值剧增。

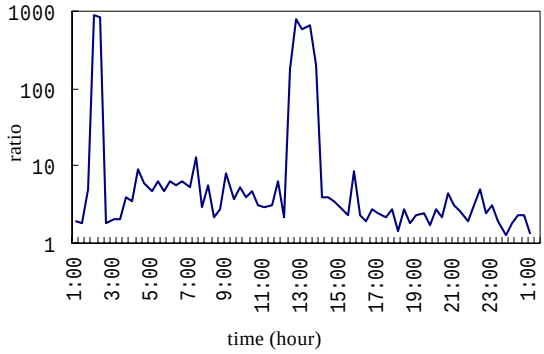


Fig.1 ratio between ICMP request traffic and reply traffic

图 1 ICMP 请求流量和应答流量的比率

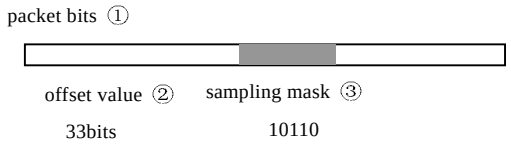
正常流量的报文平均长度是一个较为稳定的数据，而扫描攻击会产生大量的短报文，所以一旦出现大量扫描攻击发生，平均报文长度会有显著变化，因此报文平均长度可以作为一种测度。Flood 型 DOS 攻击和扫描攻击有类似之处，就是在攻击时一般会产生大量特定流量，这些特定流量会改变网络流量总体的统计性能。通过对大量 CERNET 流量统计研究发现，除了平均长度测度以外，还有一些网络流量统计特性如：TCP 平均报文长度、UDP 平均报文长度、TCP 占总流量比重、UDP 占总流量比重、以及各种 TCP 和 UDP 应用流量占总流量的比率等在大规模网络中均有较为稳定的统计值，当出现异常行为时，这些统计量的值将会明显的变化。因此可以选用这些统计量作为异常检测的测度指标，并建立统计模型以检验流量异常行为的发生。

2 实时抽样测量

抽样测量是使用抽样理论对总体中抽取的部分样本进行分析，估计出总体的相关特性。因此，当需要知道流量总体特性，而分析总体中的每个元素代价太昂贵或时间太长时，可以使用抽样技术。高速流量抽样测

量的目的是抽取流量子集来实现对总体流量信息的估计，抽样理论建立在抽样样本随机性基础上，样本的随机程度越大，对总体信息估计就越精确。针对大规模高速 IP 网络流量抽样测量，目前有二类抽样技术：一类是集中式抽样测量技术，如 RFC2330 [6] 定义的泊松抽样模型。抽样测量算法随机生成一个抽样事件，如以确定的计时器或计数器溢出作为激发抽样的事件，系统在报文到达之前就已经决定其是否被抽样，这种抽样方法需要时刻生成抽样事件。第二类是分布式抽样测量技术，抽样事件事先确定，在报文到达之前是不能确定其是否被抽样，只有当报文到达以后根据报文内容才能决定抽样与否，Cozzani I. [7] 和 Nick Duffield [8] 采用了这种抽样技术。其中随机性和效率是生成算法性能的两个关键指标。

文章所研究的是第二类抽样模型，根据 IP 报文中的标识字段的随机性，使用确定的掩码和标识字段的部分比特相匹配以实现报文抽样。如果抽样掩码和匹配比特串发生匹配，那么测量器将抽样该报文。这种匹配机制以比特为基础，使用一个内容随机的比特掩码比较每个报文中的指定比特的内容，比特掩码的偏移和长度决定测量体系结构的精度和可靠性，图 2 为文章使用的抽样掩码测量模型。理论上抽样比率是由抽样掩码比特长度决定，理论抽样比率 $ratio=1/2^n$，但实际抽样比率同被匹配的比特串有直接关系，被匹配比特串的位流熵信息效率越接近 1，则实际抽样比率越接近理论抽样比率。标识字段在统计上具有高随机性，和流量统计特性无关，且在传输过程中不发生变化，因此选用标识字段的部分比特作为抽样掩码匹配比特串。

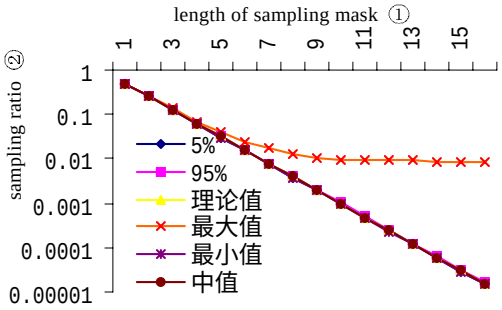


①报文比特串,②偏移值,③抽样掩码

Fig.2 sampling mask measurement model

图 2 抽样掩码测量模型

图 3 表明抽样掩码长度 n 和抽样比率之间的关系。n 比特掩码位串的理论抽样比率 $ratio_t=1/2^n$，共有 2^n 种可能位串形式，故对应有 2^n 种抽样比率。图中的 5%、95%、最大值、最小值、中值分别是指 2^n 种抽样比率中分别位于 5%、95%、100%、最小值、50%分位点处的抽样比率。图 3 可知，中值、95%、5%、最小值的抽样比率曲线和理论抽样比率曲线完全重合，只有最大值的抽样比率曲线和理论抽样曲线偏离较远。对实测数据分析发现由于 n 个比特中取全 0 的概率远高于取其它值的概率，故全 0 的掩码抽样概率远高于理论抽样比率值，其中可能原因是由于大多应用程序报文的标识字段从 0 开始赋值。因此在设置抽样掩码时，尽量不要采用掩码全 0。图 3 证明标识字段的 1～16 位作为掩码匹配字段具有非常良好的随机性能。



①抽样掩码长度,②抽样比率

Fig.3 the relation between length of mask and sampling ratio

图 3 掩码长度和抽样比率的关系

3 实时抽样测量

使用基于标识字段的实时抽样模型，从报文中提取流量的特性，实时计算和更新测度统计值。类似地，可以计算流量中的不同测度统计值，如果测度超出某一规定的阀值，则就认为出现异常行为。

3.1 流量异常检测

中心极限定理认为不管研究的统计总体服从什么分布，样本平均值的分布接近一个正态分布，正态分布的均值等于总体分布的均值，标准偏差等于总体分布的标准偏差除以样本大小的平方根。

如测量平均吞吐量，平均吞吐量可以用流量随机抽样样本的平均吞吐量来估计。随着不停地测量报文，网络流量的平均吞吐量也随着不断更新。最简化的形式可以表示为：

$$T_{n+1} \leftarrow T_n + t_{n+1} \qquad (1)$$

其中，T_n是前 n 个单位时间内平均抽样吞吐量的累加和（$T_0 = 0$），t_{n+1}是第 $n+1$ 个单位时间内平均抽样吞吐量，使用 T_{n+1}代替 T_n是存储信息，因此新的平均抽样吞吐量 $\overline{T}_{n+1}$为

$$\overline{T}_{n+1} = \frac{T_{n+1}}{n+1} \tag{2}$$

标准方差是测量数据的偏差，如果数据离平均值近，则置信区间较窄，对于 $n+1$ 个数据值，样本标准差对总体标准差的无偏估计定义为：

$$s_{n+1} = \sqrt{\frac{1}{n}\left[\sum_{i=1}^{n+1}\left(t_i - \overline{T}_{n+1}\right)^2\right]} = \sqrt{\frac{\sum_{i=1}^{n+1} t_i^2 - (n+1)\overline{T}_{n+1}^2}{n}} \tag{3}$$

对于每个测度，系统只要维护三个值：样本均值 $\overline{T}_n$、样本累加和 T_n、样本平方累加和 ϖ_n。标准差能通过计算：设 $\varpi_n = \sum_{i=1}^{n} t_i^2$， $\omega_0 = 0$， $\varpi_{n+1} \leftarrow \omega_n + t_{n+1}^2$，则

$$s_{n+1} = \sqrt{\frac{\varpi_{n+1} - T_{n+1}\overline{T}_{n+1}}{n}} \tag{4}$$

样本均值和标准差能为流量特性的总体均值构造一个置信区间（如：平均吞吐量）。样本均值的标准偏差 $s_{\overline{X}_n} = \frac{s_n}{\sqrt{n}}$。如果从相同的流量中重复选择样本，并计算每个样本的均值，则这个统计量表明期望的变化量。中心极限定理称对于大于 n 的样本，其均值服从均值等于流量总体的均值，标准差为 $s_{\overline{T}_n}$ 的正态分布。因此可以构造总体均值的置信区间 μ 为：

$$\overline{T}_n - z\frac{s_n}{\sqrt{n}} \le \mu \le \overline{T}_n + z\frac{s_n}{\sqrt{n}} \tag{5}$$

z 是一个标准正态分布的分位数。样本中元素数目 n 越大，样本均值的偏差越小，其总体均值的偏差越小。如果当前时间范围内测度值满足（5）式要求，说明当前流量行为正常，系统将更新队列数据，如果当前时间范围测度不满足（5）式要求，则说明当前流量行为异常，不更新历史队列统计数据记录，报告异常错误。

3.2 流量行为实时更新

即使对于稳定的网络行为测度，随着时间的推移，由于用户行为和网络环境的变化，其正常行为也会发生变化，因此描述正常行为的稳定测度只能称为一段时间内的准稳定测度。通常使用某一段历史时期的流量数据描述网络流量正常行为，而不是使用所有的历史测量数据，因而需要维护一个流量滑动窗口抛弃旧样本。下面给出假设 2。

假设 2. 历史窗口中流量行为接近正常行为，异常行为对历史窗口总体流量测度行为作用可以忽略不计。

根据假设 1，异常行为具有偶然性，因此当选择的历史窗口较大时，假设 2 应当是合理的。为了维护一个具有固定大小的滑动窗口队列，需要在窗队列头抛弃旧流量数据，在队列尾增加新到的数据。文章考虑使用基于单位时间尺度的滑动窗口模型，通过比较当前时间单位内抽样流量和前 n 个时间单位内的样本流量来分析在历史窗口时期观测到的持续 t_h 时间的流量行为是否和当前时期流量有相同的行为，如时间单位为 1 分钟，而历史窗口为 1 天。增加异常测试频率会增加观测数据的内存空间和系统计算时间，因此需要根据计算资源的有效性和异常分析的详细程度决定测试的粒度。

设历史窗口持续时间 n 个时间单位，在每个时间单位中维护一个三元组（T_i,ϖ_i,n_i），其中 n_i 表示在第 i

时间单位内的观测样本数，T_i 表示第 i 子时间区间内的 n_i 个观测值之和， ω_i 表示表示第 i 时间单位的 n_i 个观测值的平方和，设历史窗口各累加和为： $sT_h=\sum_{i=1}^{n}T_i$ ， $s\varpi_h=\sum_{i=1}^{n}\varpi_i$ ， $sn_h=\sum_{i=1}^{n}n_i$ 。*head* 指向窗口队列头单位时间区间（初值为 1），*tail* 指向窗口队列尾时间单位，因此新单位时间后各累加和更新如下：

$$\begin{aligned} sT_h &\leftarrow sT_h - T_{head} + T_{head+n} \\ s\varpi_h &\leftarrow s\varpi_h - \varpi_{head} + \varpi_{head+n} \\ sn_h &\leftarrow sn_h - n_{head} + n_{head+n} \end{aligned} \qquad (6)$$

同时滑动 *head* 和 *tail* 指针，$head \leftarrow head+1, tail \leftarrow tail+1$。历史窗口新的均值和标准差表示为：

$$\overline{sT_h} = \frac{sT_h}{sn_h}， \quad ss_h = \sqrt{\frac{s\varpi_h - sT_h\overline{sT_h}}{sn_h - 1}} \qquad (7)$$

如果单位时间内所考虑的测度简单，不需要考虑三元组的情况，可以直接在单位时间内维护该测度值，使用（6）式和（7）式更新历史窗口记录。需要说明的是，如果当前的流量行为已经检测到是异常了，将不更新历史窗口的数据记录。

4 流量实例分析

将 ICMP 请求报文与应答报文比值 *IR* 作为监测流量网络受 ICMP DOS 攻击的测度。设测量流量时间粒度为 T，第 i 单位时间内抽样到的 ICMP 请求报文数为 Q_i，收到的 ICMP 应答报文数为 R_i，因此第 i 时间段的 ICMP 请求报文和应答报文的比率 $IR_i=Q_i/R_i$，对于每个时间单位只需要保留的数据 IR_i。在测量流量时间粒度 T、历史窗口大小 *HL* 选择上，需要从系统资源消耗和异常检测的灵敏程度两方面考虑。选择时间粒度小，判断异常出现的实时性好，但是需要消耗更多的系统资源用于检测和更新。对于历史窗口大小 *HL* 选择上，一般是选择的窗口越大越好，因为异常行为是一些偶然现象，而正常行为是普遍现象，所以窗口越大，其体现出来的行为越接近正常现象，不会因为某些偶然现象而干扰历史行为规律。但历史窗口选择过大会使得需要更多的内存空间，同时适应正常行为变化性也较差。通过对实际网络流量实验，在使用 *IR* 测度进行异常分析时选用的时间粒度为 1 分钟，而历史窗口大小为 60 分钟。

图 1 是 CERNET 网络某日内受到的 ICMP 扫描攻击，在 1：42 时，历史窗口平均 IR 值是 3.6，方差 4.533，窗口大小为 60。因此总体平均 *IR* 值的 99%置信区间为： $3.6-2.81(0.585)\le IR\le 3.6+2.81(0.585)$ ，即： $1.96\le IR\le 5.24$ 。当前时间内测度 *IR* 值是 2.4，该值落在历史窗口 *IR* 区间范围内，所以在 1：42 时流量行为正常。更新后的历史窗口的统计记录为 *IR=3.54*，方差为 4.561，由此 *IR* 的 99%置信区间为 $1.89\le IR\le 5.19$ 。在 1：43，当前一分钟的 *IR* 平均值为 875，远大于历史窗口中 *IR* 的 99%置信区间范围，可知在 1：43 开始网络受到 ICMP 扫描攻击，报警并继续检测流量行为，但历史窗口中的 *IR* 流量统计数据并不更新，异常行为流量一直持续到 2：18 结束。在该日 12：15 又检测到发生大规模 ICMP 扫描攻击，到 13：48 结束。从图 1 可以清楚看出该日的两次大规模扫描攻击。从该例可以看出模型具有很高的实时检测性能。

5 结论

对大规模高速网络流量异常行为进行实时检测会受到检测系统性能的约束而变得不可行，包括在检测数据的处理速度和存储容量这两方面，而使用抽样检测则可使这种检测成为可能。这种异常检测不需要使用传统方法测量所有的流量数据、记录所有的流量信息，为了能够实现实时需求，使用统计随机抽样报文流量代替测量所有流量信息。由于使用确定的掩码和标识字段的部分比特相匹配所获得的随机抽样样本能够很好地描述流量总体的统计属性，因此可以使用抽样样本的均值和方差估计流量总体的均值和方差。通过仔细选择统计抽样方案和使用合适的累加方案可以大大减少系统计算时间和内存需求空间。为了实现对抽样流量进行实时检测和更新，文章提出了基于滑动窗口的实时异常检测模型，使得异常行为检测能在系统资源可控制范

围内实现。文章提出的新的异常检测测度，使检测模型可以建立在异常行为实际问题上，检测模型可以根据检测测度的变化，实现对不同具体异常行为的检测，进一步提高系统检测能力。

网络流量行为的变化与网络异常行为之间存在内在联系，文章通过对 ICMP 请求报文和应答报文之间比率变化的检测分析表明这种内在联系可以通过特定的网络测量方法和数学分析方法来予以揭示，这种方法与传统的将所有的观测结果做事后的聚类分析方法相比在实时性和处理开销上都具有更多的优势。文章提出的方法和思路对其它网络安全检测研究也具有一定的指导意义。

References:

[1] Denning, D. E., An intrusion-detection model, IEEE Transaction on Software Engineering, 1987, SE-13: 222-232.

[2] Carla T. L., Brodley E., Temporal sequence learning and data reduction for anomaly detection, In: Reiter Med. Proceedings of the 5th conference on computer and communications security, New York: ACM, 1998. 150~158.

[3] Lee, W., Stolfo, S., and Kui Mok. A data mining framework for adaptive intrusion detection. In Proceedings of the 1999 IEEE Symposium on Security and Privacy, IEEE Press, 1999.

[4] Lunt, T. F., Tamaru, A., Gilham, F., A real-time intrusion detection expert system (IDES), Technical Report, Computer Science Laboratory, SRI International, Menlo Park, California, 1992.

[5] Gong Jian, Lu Sheng, Wang Qian, Introduction to Computer Network Security, Nan Jing: Southeast University Press, 2000. 203~236 (in Chinese).

[6] Paxson, V., Almes, G., Mahdavi, J., Mathis, M., Framework for IP Performance Metrics, IETF RFC 2330, 1998.

[7] Cozzani I.; Giordano S, A passive test and measurement system: traffic sampling for QoS evaluation, Global Telecommunications Conference, 1998. GLOBECOM 1998. The Bridge to Global Integration. IEEE, 1998, Volume: 2, 1236 –1241.

[8] Nick Duffield, Matthias Grossglauser, Trajectory Sampling for Direct Traffic Observation, Proceedings of ACM SIGCOMM 2000, Stockholm, Sweden, August 28 – September 1, 2000

附中文参考文献:

[5] 龚俭,陆晟,王倩.计算机网络安全导论. 东南大学出版社. 南京. 2000. 203~236

A Real-Time Anomaly Detection Model on Sampling Measurement in a High-Speed Network*

Cheng Guang, Gong Jian, Ding Wei

(*Department of Computer Science and Engineering, Southeast University, Nan Jian 210096, China*)

E-mail: gcheng@njnet.edu.cn

http://www.njnet.edu.cn

Abstract: Real-time anomaly detection is a highlighted topic of network security research in recent years. In the paper, based on statistics character of traffic in a large-scale network, the steady metrics that can estimated network behavior are found and a sampling measurement model is presented. According to the center limited theory and hypothesis test, a real-time detection model on anomaly behavior of network traffic is built. Finally, the network behavior metrics on the ratio between ICMP request packets and reply packets is defined and the ICMP scan attack in the CERNET network is monitored real timely. Method and idea of the model has some directed sense for other network security detection research.

Key words: sampling measurement; metrics; anomaly detection; smoothing window; high-speed

* Received Feb. 25, 2002; accepted Apr. 26, 2002

Supported by the National Natural Science Foundation of China under Grant No.90104031; the National High Technology Development 863 Program of China under Grant No.2001AA112060