第42卷 第2期 西 安 交 通 大 学 学 报 Vol.42 №2
2008年2月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Feb. 2008

一种用于异常检测的网络流量抽样方法

潘乔，裴昌幸，朱畅华
（西安电子科技大学综合业务网国家重点实验室，710071，西安）

摘要：为了减小抽样数据对网络异常检测的影响，提出了一种新的可变抽样率的网络流量抽样方法.通过利用哈希模式匹配算法，将到达的数据报文按流标识分类并记录下该报文在流中的位置，然后根据报文所属流的位置顺序减函数来设置不同的报文抽样概率.实验结果表明，所提方法增加了短流报文的抽样概率，解决了由于随机报文抽样方法偏向于长流抽样而导致的网络异常丢弃的问题，从而提高了异常检测的正确性.

关键词：网络流量；可变抽样；随机报文；异常检测
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2008)02-0175-04

Novel Traffic Sampling Method for Anomaly Detection

PAN Qiao, PEI Changxing, ZHU Changhua
(State Key Laboratory of Integrated Services Networks, Xidian University, Xi′an 710071, China)

Abstract: In order to reduce the impact of sampled traffic on network anomaly detecting, a novel method with variable sampling rates in traffic sampling is proposed. By using the hash pattern matching algorithm, the incoming packets are classified by flow's ID and the packet's positions in the flow are recorded. Then, various sampling rates are specified according to the decreasing order function of the flow that the incoming packet belongs to. Experiment results show that the method increases the sampling rates to small flows, and resolves the problem that a great many network anomalies are discarded by the random packet sampling that has a bias towards large flows, and that the accuracy of anomaly detecting is improved.

Keywords: network traffic; variable sampling; random packet; anomaly detecting

网络异常检测的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，确定新的网络威胁，并向管理员提出警告或主动做出反应.网络的异常行为主要表现为网络流量异常，例如多源的分布式拒绝服务攻击(DDoS)或蠕虫爆发所引起的突发流量行为，其特点是发作突然、先兆特征不明显或者比较隐蔽，因此网络流量的正确测量是网络异常检测的关键和前提.

随着大规模、高速互联网的发展，流量全采集测量技术已经无法继续进行.近来，网络流量的抽样数据已作为异常检测的数据源[1-4]，但是抽样测量是一种近似的测量方法，每个抽样数据的流量相对于总体流量的分布必然存有偏差，这将大大影响异常检测的结果[5].

本文通过分析随机报文抽样数据对异常检测的影响，提出了一种新的可变抽样率的网络流量抽样方法.该方法可减小抽样数据对于异常检测的影响，提高异常检测的正确性.

1 随机报文抽样数据对异常检测的影响分析

目前，在网络流量抽样测量中，使用最为广泛的方法主要是基于IP流的随机报文抽样方法[6]，它的

收稿日期：2007-06-27. 作者简介：潘乔(1977—)，男，博士生；裴昌幸(联系人)，男，教授，博士生导师. 基金项目：国家自然科学基金资助项目(60572147，60132030)；陕西省科技攻关资助项目(2006K04-G33).

特点是简单、可行、易于操作.

设一个长度为 l 的 IP 流,其随机数据报文抽样概率为 p,即在以长度为 l 的 IP 流中,每一个数据报文被抽到的概率为 p,抽不到的概率为 $1-p$,则流中有 k 个数据报文被抽到的概率为 $\boldsymbol{P}_r(X=k)$,其中 X 表示 IP 流中被抽到的报文数.根据随机报文抽样的性质,X 应服从参数为 l、p 的二项式分布,所以

$$\boldsymbol{P}_r(X=k)=\begin{bmatrix} l \\ k \end{bmatrix} p^k(1-p)^{l-k} \qquad (1)$$

对于一个 IP 流抽样,如果在 l 个报文中只要有一个报文被抽到,那么就认为这个流被抽样.所以,一个数据报文不被抽到的概率为 $P'_r=\boldsymbol{P}(X=0)=(1-p)^l$,也就是说长度为 l 的流至少有一个数据报文被抽样的概率为 $1-P'_r=1-(1-p)^l$.当 $p=1/10$ 时,流被抽样的概率与流大小的关系如图 1 所示.从中可看出,流越长,被抽样的概率就越大.所以,在固定抽样概率的情况下,长流的抽样概率大,易于抽样,抽样概率几乎是 100%,而短流的抽样率则很低.但是,网络异常流通常由突发的一系列短流组成,且大部分仅为单个报文,如 MS SQL 发生蠕虫攻击[7],则受感染的主机将向随机选出的目的 IP 地址发出一个序列攻击报文.因为,基于 IP 流的随机报文抽样方法主要抽取长流,无法获取足够的异常短流,当网络发生异常时,就出现了抽样数据不可见的情况,给异常检测带来了错误的数据源.

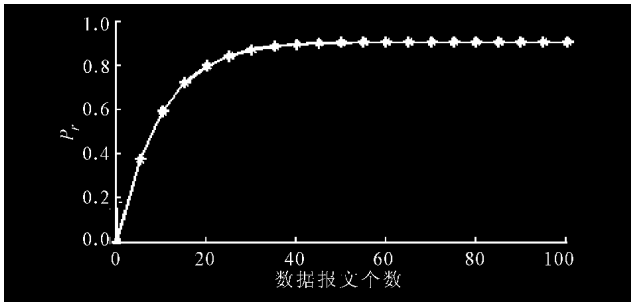

图 1 流抽样概率与流大小的关系

2 变抽样率的网络流量抽样方法

根据数据流的特点,可将 IP 流到达的过程看成簇生点的过程,即由主过程(流到达过程)和以主过程的每一点为中心产生的从属过程(数据报文到达过程)组成.那么,基于流的数据报文到达过程如图 2 所示.从中看出,只要 IP 流中的第 1 个数据报文到达了,就可以看作该流已经到达了,其后到达的数据报文只是主过程中的从属过程,它不影响主过程(流到达过程)的分布特征.由于随机报文抽样方法的抽样概率只考虑了从属过程,即到达的数据报文与 IP 流的大小及其先后顺序无关,所以均可以固定概率 $p=1/N$ 随机抽样.长 IP 流包含的数据报文比较多,因此在抽样过程中自然偏向于采集长流,忽略短流.

f_a、f_b、f_c、f_d 代表 4 种 IP 流;括号中的数字表示到达流报文的顺序

图 2 基于流的数据报文到达过程

可变抽样概率的网络流抽样方法(简称本文方法)的基本思想是,在设置从属过程的数据报文抽样概率的同时,应考虑到主过程(流过程)IP 流的性质,并根据每一个到达报文的 IP 流大小和报文到达的先后顺序来设置可变的报文抽样概率,这样既减小了长 IP 流的报文抽样概率,也提高了短 IP 流的报文抽样概率,从而消除了随机报文抽样方法存在的缺陷.

判断到达报文 IP 流的大小是一个难点问题.一个 IP 流可能包含一个或者多个报文,而报文的到达过程是一个随机过程.一个到达的报文,一般不会事先知道其后是否有同属一个流的报文将要到达,所以很难预知所属流的大小.解决此问题,本文采用了逆向分析的方法.

对于刚到达的报文,先不考虑其所属流的大小,而是判断该报文在其所属流中的位置,然后再将该报文抽样概率定义为以其位置参数为自变量的减函数,这样就保证了 IP 流中先到达的报文抽样概率最大.当一个短流到达时,它的报文抽样概率大于长流中的大部分报文的抽样概率,从而提高了随机报文抽样方法中短流的抽样概率.以图 2 为例,对于 IP 流 f_a,它有 4 个报文 $f_a(1)$、$f_a(2)$、$f_a(3)$、$f_a(4)$,若按照报文到达的先后顺序设置抽样概率,则有 $p_{f_a(1)}>p_{f_a(2)}>p_{f_a(3)}>p_{f_a(4)}$,这是一个位置的减函数.对于 IP 流 f_c,只有 1 个报文 $f_c(1)$,其抽样概率为 $p_{f_c(1)}=p_{f_a(1)}$.如果采用随机报文抽样方法,令抽样概率 $p=1/N$,那么在短流 f_c 和长流 f_a 中所有报文的抽样概率都是 $1/N$,因此短流 f_c 的抽样判决必然会受其附近长流 f_a 报文 $f_a(2)$、$f_a(3)$的影响.在可变抽样率的方法中,$p_{f_c(1)}>p_{f_a(2)}>p_{f_a(3)}$,从总体 IP 流抽样的观点来看,短流的相对抽样概率将得以提高.

为了便于描述本文方法，先作如下说明.

(1)在报文集合 $A=\{a_1,a_2,\cdots,a_i,\cdots\}$ 中，a_j 表示在时刻 j 到达的报文，$j=1,2,\cdots,i,\cdots$.

(2)报文 a_i 的流标识关键字定义为 $K_{\text{flowID}}(a_i)$.

(3)引入关联计数器，以记录报文在所属 IP 流中的位置，并定义地址空间 r 对应的关联计数器为 $C(r)$，其值为 $n\in\{1,2,\cdots\}$.

(4)存储器 R_1、R_2、R_3 分别存储未抽样的报文 IP 流记录、抽样后输出的报文 IP 流记录和关联计数器的值.

(5)流的超时时间定为 $T=64$ s(参照文献[8]的比较结果)，这样可节约存储器空间.

可变抽样概率的网络流量抽样方法的步骤描述如下.

(1)利用哈希模式匹配法，将新到达的 a_i 按流标识快速进行分类，并记录下 a_i 在 IP 流中的位置.

哈希模式匹配法主要是，先将模式串和本文子串分别用哈希函数转换成哈希值，在模式匹配时只要找出相同的哈希值即可. 这样，便用数值比较法代替了位比较法，进而提高了运算效率.

本文首先计算到达 a_i 的流标识哈希值，再将其映射为流记录存储器空间上的地址，并在该地址空间存放 a_i 所属流的记录. 当下一个报文到达时，只需计算报文流标识的哈希值，通过查看该值对应的地址空间的流记录，将新到达的 a_i 归类，否则新建一个流记录.

例如，当一个 a_i 到达测量节点时，利用定义的哈希函数 H 计算 $K_{\text{flowID}}(a_i)$ 对应 R_1 的地址 $r=H(K_{\text{flowID}}(a_i))$，然后判断该值在对应的地址空间上是否有数据. 如果没有数据，说明存储空间上不存在 a_i 所属的 IP 流，应在地址 r 的空间上新建该流，记录流信息，并令其对应的关联计数器 $C(r)=1$；如果有数据，说明 a_i 所属的 IP 流已经存在，只需将该地址空间 r 对应的关联计数器 $C(r)$ 加上 1.

(2)根据关联计数器 $C(r)$ 设置 a_i 的抽样概率 $p(n)$，并以概率 $p(n)$ 抽样 a_i. 如果要对 a_i 进行抽样，在 R_2 上应建立 a_i 的流记录；反之，则丢弃 a_i. 另一方面，在 R_2 中实时检查每一个流是否超时或者是否收到了一个具有流结束标志的报文[9]，以确定该流是否结束. 如果该流结束，将流记录和流大小的信息传出，并送入采集分析器，此时抽样完成.

3　实验结果的比较分析

实验数据是 2006 年 10 月 21 日在某电信公司特级骨干网上测得的，前 1 h 未抽样且网络流量数据仅含有数据报头信息，即包含 7 010、324、645 个报文，合计为 84、745、636 个 IP 流，流超时时间 $T=64$ s.

本文的重点是讨论抽样数据对异常检测的影响，为了解决被动测量的数据不可能全面包含网络异常的缺陷，应将该数据作为背景流量进行重放，并加入异常检测中常见的网络攻击，如业务量异常. 这种异常主要是由非正常的流量负载引起的网络流量突变，如蠕虫攻击、拒绝服务攻击(DoS)等.

本文模拟了流量数据中单源 DoS 对单个目标地址的攻击，多源 DDoS 对单个目标地址的攻击及其对随机选出的目的 IP 地址发出的一系列报文的蠕虫攻击. 这 3 种业务异常共出现 30 次，以便于评价不同抽样数据的异常检测结果.

在本文方法中，哈希函数采用 Bob 函数[9]，它可以满足高速网络的测量要求，其输入是以字节为单位的可变长度的比特值，输出是一个具有足够随机化的 32 b 的数值.

为了比较随机报文抽样方法、本文方法，可采用文献[2]的小波分析法来探测本地时间序列方差的突变情况，以判断业务量异常的检测结果. 由于这 2 种抽样方法的参数有所不同，所以应将抽样得到的报文占原始总报文的比例作为它们共同的参数. 实验中的业务量异常检测结果见表 1，其中随机报文抽样方法的抽样概率分别为 1、0.1、0.01 和 0.001. 若抽样概率为 1，则表示将原始未抽样的数据作为输入的检测结果(见表 1 第 1 行).

由表 1 可以看出，随着抽样概率的提高和抽样报文所占比例的减小，随机抽样方法的业务量异常检测性能大幅下降，而本文方法的检测结果与原始数据基本相符. 主要原因在于，这些异常通常是由短流引起的，而本文方法是基于 IP 流大小设置抽样概率的，当抽样报文的比例下降时，抽样得到的长流报

表 1　2 种抽样方法的业务量异常检测结果

抽样报文所占比例/%	随机报文抽样概率	异常检测数	
		随机抽样法	可变抽样概率法
100.0	1.000	29	29
10.0	0.100	25	29
1.0	0.010	10	28
0.1	0.001	3	28

文比例将比短流情况下降得快，这并不影响短流抽样，长流和短流所占的比例基本不变，所以检测结果几乎不受抽样概率的影响. 随机抽样方法的抽样概率是针对报文设计的，短流其本身所含的报文数比较少，所以随着抽样报文比例的下降，短流被抽到的机会就更小，从而影响到异常检测的结果.

4 结 论

通过分析随机报文抽样方法对异常检测结果的影响，提出了一种可变抽样概率的网络流量抽样方法. 该方法根据到达报文所属流的大小来设置不同的报文抽样概率，由此减小了长流报文的抽样概率，提高了短流报文的相对抽样概率. 通过实验分析、比较可变抽样概率的网络流量抽样方法、随机报文抽样方法对于业务量异常的检测结果，可以看出利用前者可减少抽样数据对于异常检测的影响，检测结果更加接近真实情况.

参考文献：

[1] ZSEBY T, FOKUS F, MOLINA M, et al. Sampling and filtering techniques for IP packet selection [EB/OL]. [2007-01-20]. http://www.ietf.org/html.charters/psamp-charter.html, 2007.

[2] PAUL B, JEFFERY K, DAVID P, et al. A signal analysis of network traffic anomalies [C]// Proceedings of ACM SIGCOMM Internet Measurement Workshop. New York: ACM Press, 2002:71-82.

[3] AVINASH S, TAO Ye, SUPRATIK B. Connectionless port scan detection on the backbone [C]// IEEE International Performance Computing and Communications Conference. Pisataway, USA: IEEE, 2006:567-576.

[4] CHOI B Y, PARK J, ZHANG Zhili. Adaptive random sampling for traffic load measurement [C]// Proceedings of IEEE International Conference on Communications. Piscataway, USA: IEEE, 2003:1552-1556.

[5] MA Jianning, CHUAH Chen-Nee, ASHWIN S, et al. Is sampled data sufficient for anomaly detection [C]// Proceedings of the 6th ACM SIGCOMM on Internet Measurement. New York: ACM Press, 2006:165-176.

[6] CISCO Systems Inc. Random sampled NetFlow [EB/OL]. [2007-01-20]. http://www.cisco.com/en/US/products/ps6566/products_feature_guide 09186a008-0796a49.html.

[7] BEVERLY R. MS-SQL slammer/sapphire traffic analysis [EB/OL]. [2007-01-20]. http://momo.lcs.mit.edu/slammer/.

[8] APISDOR I J, CLAFFY K, THOMPSON K, et al. OC3MON: flexible, affordable, high performance statistics collection [C]// Proceeding of the 7th Annual Conference of the Internet Society. Kuala Lumpur, Malaysia: The Internet Society, 1997:97-112.

[9] LUND D N, THOURUP C M. Estimating flow distributions from sampled flow statistics [J]. IEEE/ACM Transactions on Networking, 2005, 13(5):933-946.

（编辑 苗凌）