高速网络中入侵检测的抽样方法*

宁 卓[1,2+]，龚 俭[1,2]，顾文杰[1,2]

[1](东南大学 计算机科学与工程学院,江苏 南京 210096)
[2](江苏省计算机网络技术重点实验室,江苏 南京 210096)

A Sampling Method for IDS in High Bandwidth Network*

Zhuo Ning[1,2+], Jian Gong[1,2], Wenjie Gu[1,2]

[1](School of Computer Science and Technology, Southeast University, Nanjing 210096, China)
[2](Jiangsu Provincial Key Laboratory of Computer Network Technology, Nanjing 210096, China)
+ Corresponding author: Phn: +86-25-83794000-304, E-mail: zhning@njnet.edu.cn, http://njnet.edu.cn
Received 2009-02-18; Accepted 2009-00-00

Abstract: It is well known that Intrusion Detection System (IDS) does not scale well with Gigabit links. Unlike the other solutions that try to increase the performance of IDS by the distributed architecture, we develop a novel sampling method IDSampling for backbone intrusion detection which is adaptive with the consumption of the memory bottleneck. With the help of the heuristic messages, such as the entropy of the single-packet flow and the flow length, IDSampling applies the simple sampling strategy based on the entropy of the single-packet flow when large-scale anomaly occurs, or another complicated one instructed by the feedback of the rear detection results by default. The results of experiment show that comparing with the other two overwhelming sampling methods, the Random Packet Sampling and the Random Flow Sampling, the number of attack packets sampled by IDSampling is higher one order of magnitude than that of the former two in large-scale anomaly case, so IDSampling makes IDS remain effective when it is overloaded for IDSampling can guarantee the detection accuracy of the trend information of the large-scale anomaly.

Key words: intrusion detection System; Sampling; Bloom Filter; sample entropy

摘 要: 目前入侵检测系统（IDS）在万兆网络上的应用存在着性能和精度不平衡问题。与传统采用分布式 IDS 集群方式来解决性能瓶颈的“开源”方法不同，本文提出了一个面向骨干网入侵检测，以内存瓶颈消耗量为测度的动态自适应抽样方法 IDSampling。通过分析攻击流量的流长和熵聚类信息特征指导抽样,过滤掉攻击可疑性低的报文，采取“节流”方法解决上述不平衡问题。在大规模异常发生时采用基于单报文属性熵的单一抽样策略，其他情况下采用带反馈指导的混合抽样策略，试图用尽可能小的检测代价来取得同样的检测效果。实验结果表明①IDSampling 可以大幅减低 IDS 处理输入，同时保证对骨干网大规模攻击趋势性信息的检测精度；②相较于通常的报文抽样和随机流抽样方法，IDSampling 凭借流长、熵聚类信息和后期检测结果等启发式信息的指导，其抽取攻击报文的准确性高于前两种方法，尤其是在大规模高强度攻击情况下 IDSampling 抽中攻击报文的数目甚至高于其他两种方法一个数量级。

*本文得到国家科技支撑计划No.2008BAH37B04和国家重点基础研究发展计划973计划No.2009CB320505资助;

作者简介: 宁卓(1975－),女,广西玉林人,博士生,主要研究领域为网络安全检测;龚俭(1957－),男,博士,教授,博士生导师,主要研究领域为网络安全,网络行为学。顾文杰(1984－),男,江苏宜兴人,硕士生,主要研究领域为网络行为学。

关键词: 入侵检测、单报文属性熵、抽样、Bloom Filter
中图法分类号: TP393 **文献标识码:** A

滥用入侵检测系统（IDS）由于误报率低且警报结论明确在实践中一直都占据主流位置。但是其检测原理导致它在高速网络（>1G+bps）中的应用普遍存在着性能-精度失衡问题。当规则集大而全时 IDS 分析准确性高，随之而来的是计算量大，当报文到达的速度大于 IDS 检测速度时丢包率会上升。此时被丢弃报文中的攻击 100%被漏报，而丢包引起的信息缺失使得信息不完整会间接导致误报率上升。如果采用小规则集，以牺牲分析准确性为代价，提高 IDS 数据处理的速度。此时虽然减小了丢包率，但是网络流量小时资源反而闲置，且牺牲了精度后由攻击的漏检带来的危害不好衡量。而按照网络技术发展的趋势，一方面，网络服务的多样化使得 IDS 的规则库越来越大；另一方面，网络链路传输速度激增，导致 IDS 单位时间的处理压力越来越大，上述性能-精度不平衡状态随技术发展反而日益加剧。

现有对缓解性能-精度失衡问题有贡献的研究主要分为两类：一类通过增加 IDS 的处理能力来缓解性能-精度失衡问题，我们称之为“开源”方法；另一类通过减少 IDS 的输入来解决问题，相对于前者，我们称之为“节流”方法。其中“开源”方法又分为两类：①高速硬件算法研究[1-3]；②分布式结构研究[4,5]。硬件算法以并行计算优势来提高检测性能，而分布式解决方法采用负载均衡和流量分配技术，以多台 IDS 的“多对一”集群技术来解决性能问题。

硬件算法成本过高，而且无法适应 IDS 规则库变化频繁的特点，难以普及。分布式解决方法也存在缺陷。首先不同于其它应用，IDS 的负载均衡技术要求不但追求均匀分配处理负载，而且更要保证相关的报文要分配到同一检测节点，否则会损坏攻击的上下文联系。所以分布式 IDS 的负载均衡算法都只能保证分配到每个检测节点上的流量负载在长时间粒度下宏观上的平衡，而短时间粒度下微观上的平衡无法保证。而且网络流量的重尾特性导致了大多数的流量负载实际上只由少数长流/超长流传输,所以即使负载均衡算法能够保证完美的 Hash 均匀性，大多数的流量负载还是会分到某个/少数单检测节点。在单检测节点中仍然存在报文到达速度大于检测速度的情况，也就是说少数 IDS 检测节点注定分到大部分的检测任务，很容易处于过载状态，而其他的检测节点则无法分担过载 IDS 的检测任务，处于闲置状态。尤其在某些典型的攻击场景下（同源同宿的 DDOS 攻击等），所有攻击流量会被分到同一 IDS 检测节点，丢包会导致无法检测这些关键攻击，分布式结构在最应该发挥作用的时候反而无用。其次，IDS 集群会增加管理、升级和维护的复杂性。

“节流”方法目前还处于探索阶段,工作主要集中在采用采样方法进行异常检测的研究上[6-9]。[6~7]分别提出了采用抽样数据在线检测扫描的快速方法。[8]通过试验考察在不同抽样率、不同检测参数下几种经典异常检测算法的检测性能变化，从而得出随机流抽样方法较之随机报文抽样方法更适合异常检测的结论。[9]基于主成分分析方法提出了一种基于熵的骨干网经典异常检测方法，[10]通过试验证明基于熵的异常检测方法对抽样带来的数据丢失有良好的容忍度，即使在很低的抽样比下采用报文抽样算法仍能检测到 worm 攻击。但是至今抽样方法能否胜任入侵检测的需求仍然是个开放性问题。

上述工作都是在 IDS 中直接使用目前流量测量中的经典抽样方法,即报文抽样或流抽样方法[8]。目前尚未见专门针对 IDS 的抽样技术的讨论。这是因为传统的 IDS 对一个攻击报文生成一个警报，从这种单报文攻击的观点来看，抽样意味着信息的丢失，必然导致 IDS 检测率的降低，所以很难接受抽样技术。但是在骨干网高速入侵检测背景下采用抽样技术却可以在检测精度和检测性能上找到合适的折中点。首先，从骨干网高速入侵检测的需求上看,它的检测优势在于能提供对大规模攻击的入侵检测能力和更多攻击的宏观趋势性信息。对于大规模攻击局域网 IDS 通常由于视野不够而只看得到部分攻击轨迹，只有检测整个攻击场景范围流量的骨干网 IDS 才能重构整个攻击场景。而且由于检测范围广、信息量大，骨干网 IDS 的统计能提供更多宏观趋势性信息。比如出现最多的攻击类型 TopN 排行以及攻击源、宿地址分布和地理位置分布等等统计信息以及这些信息长时间粒度下的变化趋势比较分析结论。因此骨干网入侵检测对抽样会丢失攻击信息并不敏感，只要不影响到大规模攻击检测和趋势性分析都可以接受。其次，目前主流的大规模攻击包括 Probe、DDos、蠕虫和僵尸程序等，它们在传播和入侵过程中通常都具有多对一的冗余特性，因而适合采用抽样技术。再次，一

个有着明确目标的大规模攻击通常具备若干步攻击步骤，它们具有明确的上下文联系，包含若干个对话，有的攻击持续时间甚至长达数小时乃至数天，其攻击能力不只局限于攻击主机，甚至可以阻塞受害主机所在的网段链路。在这样一个复合攻击中丢失了其中某些攻击报文并非无可救药，只要整个复合攻击被抽样的攻击报文保持了足够的信息，丢失的攻击并不影响对整个复合攻击的意图识别，也可以采取补偿措施来弥补抽样损失。最后，因为入侵协同技术不成熟以及它行政实施上的困难，如果能够在骨干网上一个监测点上采取抽样技术，根据某种启发信息抽取攻击“嫌疑”高的报文，在大幅减低 IDS 的处理压力的同时保证对大规模攻击和攻击趋势性信息的检测精度，对骨干网入侵检测是很有实用意义的事情。

本文通过分析骨干网 IDS 的检测特点，考察攻击的流量特征提出了一个面向骨干网高速入侵检测的抽样方法 IDSampling。一改过去非交互式的检测流程，以攻击流量的单报文属性熵特征和流长特征为启发式信息抽样，以后期检测模块的反馈结果指导前期抽样，在感知大规模异常发生时采用基于单报文属性熵的单一抽样策略，其他情况下采用带反馈指导的混合抽样策略，动态自适应地选择攻击“嫌疑”度高的报文，丢弃“嫌疑”度低的报文。实验结果表明 IDSampling 能在 IDS 过载时有效减少 IDS 的输入，相同抽样率下较之其他两种随机流/报文抽样算法能更准确地抽样到攻击报文，尤其是在大规模高强度攻击情况下 IDSampling 抽取到的攻击报文数目甚至高于其他两种方法一个数量级，保证对大规模攻击趋势性信息的检测精度，达到用最小的检测代价换取最大安全的目的。

本文其他章节安排如下。第一节通过分析攻击流量的特点提出了针对不同攻击类型的抽样策略；第二节详细介绍 IDSampling 的两种抽样策略的实现方法和细节以及如何对攻击流打标记等。第三节是算法复杂度分析。第四节介绍实验结果及分析结论。最后我们总结全文并提出了将来值得进一步探讨的问题。

1 IDSampling 的基本思想

与网络行为学中抽样的目标不同，面向 IDS 的抽样并不关心如何对流量无偏地抽样使得原始流量的某些关键特征能从抽样样本中还原出来，比如流长分布、流到达率等。设到达报文组成的报文空间为 X，攻击报文空间为 A，p 为抽样率，抽样后得到的抽样报文空间为 S，有 $A \subseteq X$，$S \subseteq X$，则面向 IDS 的抽样目标是最大化 $Num=(A \cap S)$。因此对有不同“嫌疑度”的报文以不同概率抽样，从直觉看更为合理，而这种不等概率抽样实际上意味着流量选择。这隐含了两个任务：其一是挑选具有攻击特征的报文，其二是在满足任务一的要求下兼顾 IDS 的处理瓶颈压力，丢弃对瓶颈资源压力大的流量。

1.1 动态自适应抽样率

在抓取到网络报文之后，IDS 开始处理此报文之前，采用动态自适应的抽样技术过滤掉一部分报文，目的是在 IDS 的处理能力赶不上处理需求的时候丢弃部分攻击嫌疑小的流量，使 IDS 能够继续有效工作。因此 IDSampling 的抽样率受限于 IDS 的处理能力。而且 IDS 的处理能力和处理需求都在动态变化，因此抽样率也应随之变化。一般地，IDS 的处理能力无非受两种瓶颈因素制约，即处理速度和存储空间。存储空间限制同时相关于处理速度和报文到达速度，因此在使用动态自适应抽样方法解决存储空间限制问题的同时可以兼顾处理速度问题。这意味著一旦存储空间溢出就应该采用相应的抽样策略将输入报文量控制在一个固定的水平。

抽样率 P=IDS 的报文处理速率/网络报文输入速率 （1）

其中 IDS 的处理速率等于接收缓冲区环形队列报文读取的速率，网络报文输入速率等于报文入队列的速率（pps）。我们将时间轴划分为定长的时间片，每个时间片首按式(1) 动态自适应地确定当前时间片抽样率。

1.2 不同流量特征下的抽样策略

挑选攻击报文首先要考虑的问题就是攻击具有什么特征。IDS 中攻击的特征主要分为报文头特征和报文负载特征。报文头特征是指流的五元组信息以及一些标志位信息等，而报文负载特征主要是一些字符串特征，用于标示某种特定攻击的语义特征。由于高速线速处理的要求，IDSampling 优先使用报文头特征，而且通过利用后期 IDS 检测结果反馈指导的方式间接使用了报文负载特征。报文头特征主要刻画流量特征，通常包括流长、流速和流到达率等[11]。其中流速和流到达率两个测度因其受网络状况影响较大，不能体现攻击的固有

特征，因而我们选择流长作为测度来讨论攻击流的特点和相应的抽样策略。

一般情况下网络中的攻击报文的百分比不高，大多不到 0.05%，而当大规模异常攻击发生时网络中的攻击报文的百分比含量会变得很高，如 DDoS 攻击甚至能产生大于 1Gbps 的攻击流量[12]。一般状况下含攻击少，所以丢掉攻击报文的概率也低，但是一旦错过少数攻击报文则抽样就毫无作用可言。大规模攻击发生时流量中攻击报文含量变大，所以丢掉攻击报文的概率也变高，但是可以利用大规模攻击多对一重复特征指导抽样，更易捕获攻击报文。所以大规模异常是否发生采取的抽样策略相应有很大不同，分别讨论如下。

大规模的异常攻击通常会导致攻击流量激增，如 DDOS 和蠕虫爆发时网络中的报文数、字节数和流数目都大幅增加，它的另一个特点是攻击报文重复性高，通常具有某种多对一模式。此时为了最大限度地利用有限的 IDS 处理能力，最明智的抽样策略就是将有限的抽样机会尽量分配给“最异常”的报文，提高异常流的流内报文抽样比。如何有效衡量流量的异常度呢？考察目前网络中广泛存在的 5 种恶意大规模异常攻击[9]：水平扫描、垂直扫描、DDos、worm 和僵尸网络，这些攻击要么由于对宿主机主机相关信息缺失，要么由于对宿主机的可用性探测在不同的特定时候都会产生大量的单向报文，规模越大、强度越高、单向报文的数量就越多。骨干网单向 IP 流的流长分布研究[11]表明单向流量的流长决大多数都不长，主要集中在 1~100 以内，其中流长小于 10 的流占到了 95%以上，而流长为 1 的单报文流又占了其中的大部分，在其研究的 4 个骨干网 trace 中都超过了 50%。可见正确反映单报文流的聚类特征已经可以反映大部分有用的异常聚类语义信息。我们在 2.4 节中提出了根据单报文流多对一特性衡量流量异常度的方法，在大规模攻击发生时尽力抽取“最异常”的流，有益于集中有限资源发现更多的大规模异常流量。

相较于以数量取胜的大规模异常攻击，还有很多攻击只关注于机器本身存在的漏洞和弱点，比如说信息窃取类的攻击。这类攻击不再具备大规模异常攻击所具有的某种相同模式的大量重复报文，对于此类攻击我们通过分析一个典型攻击的一般过程入手发现它的一般规律。一个典型攻击分为七步：①隐藏自身、②收集攻击目标信息、③弱点挖掘、④获得系统控制权、⑤隐藏攻击行为、⑥实施攻击和⑦清除攻击痕迹。开始时攻击者需要收集信息以决定攻击目标和手段，从攻击步骤上来说②~③步攻击离不开扫描（扫描 IP 地址、端口和漏洞），因此此时攻击信息的载体是短流，攻击信息更多体现在短流的聚类特性中。攻击的后期④~⑦才以长流为攻击载体。为了达到最大攻击的可能性，黑客工具通常在攻击的关键步骤④和⑥时会采取很多同义重复，反复探测的操作。比如著名的 sadmind 就会同时尝试三种指针的溢出攻击以达到获取根目录特权的目的，攻击的冗余提高了抽样技术的可用性。此时的抽样策略是：高抽样比抽取短流，保证在攻击的开始阶段最大程度地检测到攻击，标记这些有“嫌疑”的攻击流，对它们后继报文采用 100%的抽样率全部抽样，而对于没有被标记的流，其可疑性较小可以降低其抽样率。由于正常情况下网络中的攻击含量很低，不超过万分之五，因此标记方法可以大幅降低 IDS 的处理负载。

IDSampling 的第二个任务是丢弃对瓶颈资源压力大的流量。上述一般情况下的抽样策略实际上已经兼顾到了减低 IDS 压力的任务。这是因为流长的重尾特性[11]决定了短流数量虽多但产生的流量负载却小，长流数量虽少却承担了大部分网络流量负载，大量的骨干网 trace 长时间统计数据表明近 80%的短流实际上只承担了不到 20%的报文负载，而不到 20%的长流却承担了 80%以上的报文负载[11]。因此高抽样比抽取短流并不显著增加 IDS 的检测负担，而标记攻击流的方法大幅降低了抽取到的长流数量，在保证精确度条件下大幅降低了 IDS 的处理报文数。另一个进一步降低 IDS 压力的方法是丢弃前期未检测到攻击的超长流。这是因为大量研究表明网络中存在一些数量极少的超长流，它们的数量虽少，由于持续时间长，报文多，实际上担负了大部分的网络负载[11]。IDSampling 没有必要对这些超长流抽样的原因有二。首先，研究发现这些超长流大多数用于 P2P 的流媒体传输[13]，而利用多媒体的攻击技术因其技术复杂目前还很少，多媒体流中包含攻击的概率接近于零。其次，从资源利用率的角度来说一个攻击者很少有如此多的资源和耐性等到双发交互流长变为超长流后才发起攻击，大多数情况下攻击者会直奔主题，在攻击流长还是短流时就发动攻击。而 IDSampling 高抽样率抽样短流报文，在攻击的前期阶段保证检测到攻击，并依靠打标记来保证后继攻击流的检测。因此不再抽样未被标记为攻击流的超长流对 IDS 来说既不影响检测率又可以大幅减低处理负载。

综上所述，缺省状态下 IDSampling 抽样方法采取带反馈指导的混合抽样方法，其抽样策略是：①高抽样

比抽取短流，保证在攻击的开始阶段检测到攻击，标记有“嫌疑”的攻击流，对它们加大后继报文的抽样率，对于没有被标记的长流，降低其抽样率。②不再抽样未被标记的超长流。当大规模异常发生时 IDSampling 抽样方法采取基于属性熵的单一抽样方法，其抽样策略是③抽取“最异常”的流以提高异常流的流内抽样率。

2 IDSampling 的抽样方法设计

IDSampling 以流量特征作为指导来实施不同的抽样策略以期最大限度地抽取到攻击报文。为了进行流量特性统计将时间轴分为若干等长的时间片，由于流量特征的自相似性和长相关性，因此使用上一时间片的流量特征结果来指导当前时间片内的抽样。由于每个时间段内的抽样过程都是一样的，因此后文讨论的抽样都是在一个时间片范围内。2.1 节给出了 IDSampling 的整体算法结构；2.2 节介绍大规模异常发生时采取的基于单报文属性熵的单一抽样策略；2.3 节介绍其他情况下采用的带攻击反馈指导的混合抽样策略；2.4 节介绍如何有效标记有“嫌疑”的攻击流量，即长流反馈和短流聚类策略。

2.1 基本算法结构

IDSampling 由两个并行的模块组成：抽样模块 Sampling 和标记模块 Feedback。Sampling 负责对每个报文分情况抽样，而 Feedback 负责从 IDS 的检测模块接收后继检测结果并标记攻击流，标记方法将在 2.4 节中详细讨论。其基本结构如图 1 所示，抽样方法如下，其中我们采用[9]的方法判断某个时间片内网络流量是否发生了大规模异常，采用多级 Bloom Filter 方法实时判断流长[14]，其技术细节不再赘述。

①对于时间片内每一个到达的报文 X，其 flow ID 为 F，首先通过多级 Bloom Filter 以计算报文所属流的长度 Len_F，如果 Len_F 大于 1，则将 F 从单报文流表中删除，否则在单报文流表中为 F 新建表项；

②每个时间片开始时按 1.1 节中的式(1)统计出本时间片采用的抽样率 P，判断上一时间片内是否发生了大规模异常[9]。若是转③，否则转④；

③当前已经发生大规模流量异常，采取 2.2 节所示的基于属性熵的单一抽样策略。

④采取 2.3 节所示的带反馈指导的混合抽样策略。

2.2 基于单报文属性熵的单一抽样方法

大规模异常攻击发生时的抽样策略是尽力抽取“最异常”的流，因此首要问题是如何标示流量的异常度，我们引入单报文属性熵刻画了攻击流量的多对一聚类特性，并提出了按照这种多对一聚类特性标识不同流量异常度的方法。

定义 1　单报文属性熵　设时间片中所有流长为 1 的单报文流为 S，特征属性向量 $feature_i$ 可以看作是样本 S 在特征属性上的一个划分，定义如下：$feature_i = \{(x_i, n_i), i=1,2, \ldots N\}$，意思即样本 S 中 $feature_i$ 共有 N 个特征值 x_i，每个特征值出现了 n_i 次。则样本单报文属性熵定义为：$H(Feature_i) = -\sum_{i=1}^{N} (\frac{n_i}{|S|}) \log_2 (\frac{n_i}{|S|})$，其中|S|是时间片样本中观察到的所有报文的总数。

单报文属性熵的值的变化范围在$(0, \log_2 N)$。当测度值取值为 0 时，$feature_i$ 的分布达到最大集聚，即所有报文的 $feature_i$ 值都是一样的。测度值取值为 $\log_2 N$ 时，$feature_i$ 的分布最发散，其每种可能值的取值次数相等 $n_1=n_2=\ldots=n_n$。采用单报文属性熵我们可以方便地统计某个时间片中攻击报文属性的汇聚和发散趋势。同时由于属性熵和 N 的大小有关，所以它的值同时可以反映在一个时间片内字节数、报文数和流数等和数量相关的攻击流量变化。表 1 罗列了四种异常攻击发生时四种单报文属性熵（源地址属性熵 H(srcIp)、源端口属性熵 H(srcPort)、宿地址属性熵 H(dstIp)和宿端口属性熵 H(dstport)）的变化趋势，根据定义这些变化是显然的，而 Botnet 由于在其形成过程中可以利用上述四种方式中的任一种作为传播方法，因此它的单报文属性熵变化囊括了表 1。其中“↓”表示变小，“↑”表示变大，“-”表示不确定。如表所示每一种攻击都会引起至少两个单报文属性熵的相对变化，此时最小的那个属性熵即刻画了“多对一”聚类模式中的“一”，也即是说最小属性熵刻画了汇聚特性最强的属性。此属性熵的 topN 值标示了攻击发生时出现频率最高、“多对一”汇聚特性

最强的异常流量。

表 1　大规模攻击发生时单报文属性熵变化对应表

Abnormal Lable	Defination	H(srcIp)	H(srcPort)	H(dstIp)	H(dstPort)
Port Scan	Probes to many destination ports on a small set of destination address	-	-	↓	↑
Network Scan	Probes to many destination addresses on a small set of destination ports	-	-	↑	↓
DOS/DDos	Denial of Service Attack(distributed or single-source)	↑	-	↓	-
Worms	Scanning by worms for vulnerable hosts(special case of Network Scan)	-	-	↓ ↑	↑ ↓

为了刻画最小单报文属性熵的 topN 值代表的“最异常”流量，我们相应地定义了 top-N 流。为了判定 top-N 流是否使得属性熵变得足够异常，首先引入其期望偏差作为属性熵异常度判断。

定义 2 期望偏差 设 X 代表一个单报文属性熵，设网络无大规模异常攻击时其数学期望和标准差分别用 E(X)、d_X 表示，则称 $z_X = \frac{|X - \mathrm{E}(X)|}{d_X}$ 为 X 的期望偏差。期望偏差 z_X 刻画了 X 偏离期望值多少个标准差。

上述 E(X)、d_X 和 z_X 都可以通过机器学习得到。通过训练为单报文属性熵的期望偏差设定一异常阀值 $z_{X_{threshhold}}$，当 $z_X \geq z_{X_{\text{threshhold}}}$ 时，认为 X 较之正常状况下发生了足够大的异常偏差。

定义 3 top-N 流 设 $\text{feature}_i = \{(x_i, n_i), i=1,2, \ldots R\}$ 是时间片内四个单报文属性熵中熵值最低的属性，且 n_i 按出现次数单调递减，即 $n_i > n_{i+1}$，则 top-N 流是 feature_i 值为 x_i，且对应的 (x_i, n_i) 使得式(2)成立的 N 个流。

$$z_{\mathrm{H}(\text{feature}_i)} > z_{\mathrm{H}(\text{feature}_i)\,\text{threshhold}}, \qquad (2)$$

其中 $\mathrm{H}(\text{feature}_i) = -\sum_{i=1}^{N}(\frac{n_i}{S})\log_2(\frac{n_i}{S})$，$z_{\mathrm{H}(\text{feature}_i)\,\text{threshhold}}$ 是 H(feature$_i$)的异常阀值。当大规模异常发生时熵值最小的属性的 top-N 值刻画了此时汇聚程度最大，出现频率最高的“最异常”N 个流，其定义如下：

2.3　基于单报文属性熵的单一抽样方法

此时要完成抽样到“最异常”流的任务，IDSampling 只需要以抽样率 P 采用报文抽样方法抽取 top-N 流中的报文即可。我们将这种策略称为基于属性熵的单一抽样策略，步骤如下：

①统计上一时间片中四个单报文属性熵中熵值最低的属性 feature$_i$，

②按式(2)求取 feature$_i$ 的 top-N 流，

③对于每个到达报文 X，如果 X 属于 top-N 流，则以抽样率 P 采用报文抽样的方式抽样，否则丢弃。

2.4　带反馈指导的混合抽样方法

混合抽样策略中将所有流分为三类：流长小于等于 10 的短流，流长大于 1000 的超长流，其它全部为长流。这样设定是基于流长小于 10 的流在单向报文数量上的优势，以及单向报文对攻击的指示意义[11]，将超长流的流长阀值定为 1000 是因为从资源利用率的角度来说一个攻击者很少有如此多的资源和耐性等到双方交互了 1000 个报文后还没有发起攻击，且大多数服务在 1000 个报文内都可以完成交互。

按 IDSampling 抽样思想不再抽样超长流，采样率 $P_{superlong}=0$。P_{short} 和 P_{long} 由下式决定：

$P_{short} + P_{long} = P$　　（3）

$P_{short} / P_{long} = W_{short} / W_{long}$　　（4）

其中 W_{short} 和 W_{long} 分别表示一个时间片中短流和长流的权重，用短流或长流占所有流中的比重来衡量。

W_{short}=短流个数/总的流数目，W_{long}=长流个数/总的流数目。P_{short} 称为短流基准抽样率，P_{long} 称为长流基准抽样率。抽样方法描述如下，如图 1 所示。

①按式（3~4）统计计算上一时间片的 P_{short} 和 P_{long}，作为本时间片内的抽样率；

②对于每个到达报文 X，如果 X 属于被标记的攻击流或在黑名单中则 100%抽样 X。否则 X 的抽样率取决于其流长特征 Len_F，采用报文抽样方法对短流报文以高抽样比 P_{short} 抽取报文；对长流以低抽样比 P_{long} 抽取报文；对超长流不再抽样。

其中涉及的攻击流的标记方法在 2.4 中详述。

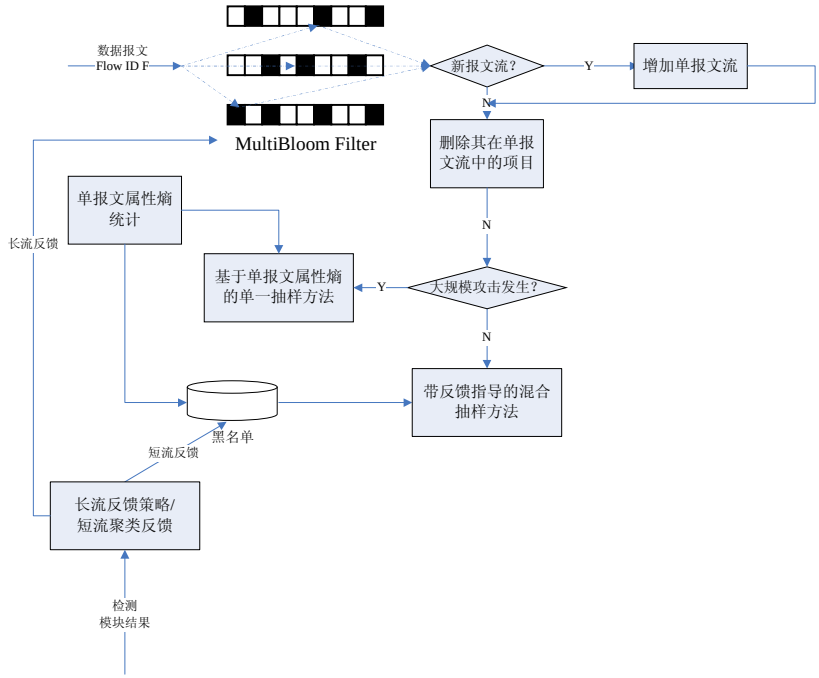

图 1 IDSampling 的基本结构

2.5 基于Bloom Filter的长流标记反馈方法和基于属性熵的短流聚类方法

由于抽样策略能利用的攻击特征仅仅限于报头特征，损失了大部分攻击的语义信息，当然很难准确地抽样。后期检测模块会对报文负载进一步进行语义特征检测因而攻击结论明确。所以我们设计反馈策略，利用后期检测的结果指导抽样策略，提高已发现前期攻击的攻击流的抽样率。不同流长的反馈策略也不相同，对长流采用基于 Bloom Filter 的长流标记反馈法，而短流采用提取流量聚类特性加入黑名单的反馈方法。

为了标记方法的简单有效我们将长流标记与流长的判断方法结合在一起，如图 1 所示。当报文 X 在后继检测中被认定为攻击报文时，计算其四元组（源 IP、宿 IP、源端口、宿端口）在各级 Bloom Filter 中的 hash 值，返回这些 hash 值给 IDSampling，并标记这些 hash 值标记的 Bloom Filter 空间（返黑）。由于 X 的后继攻击报文 Y 与 X 具有相同的 hash 值，当报文 Y 通过多级 Bloom Filter 以判断其所属流长度时会发现 Y 所 hash 到的每级 Bloom Filter 中的空间都被标记，此时可以确定 Y 是被 X 标记过的攻击流的后继攻击，应以 100%的抽样率抽样 Y。

基于 Bloom Filter 的四元组 hash 方法不适合应用于短流。这是因为短流持续时间短，等不到反馈值指导抽样该短流就结束了。而且短流数量多，采用 hash 反馈不但导致检测模块和抽样模块间的通讯量过大，而且频繁标记有限的 Bloom Filter 空间会增大攻击流判断的错误肯定率。短流的攻击语义信息多由聚类信息体现，考虑到效率问题，只存储的单报文流的聚类信息不但减少了存储量还获得没有误差的启发式信息，因此我们沿用 2.1 节中的单报文信息熵的聚类方法，在每个时间片头将按式(2)计算出的 top-N 加入黑名单中并定期更

新黑名单，对每个报文进行黑名单检查即可起到反馈攻击信息的作用。原理如图 1 所示。

3 算法的性能分析和误差估计

存储量 算法主要存储量来源于①多级 Bloom Filter 的存储空间；②单报文流存储空间；③黑名单的存储空间。设 b 为每一级 Bloom Filter 所含的 hash 空间数，d 为多级 Bloom Filter 的级数，n 为活动流数，则①的空间复杂度为 O(bd)；②的空间复杂度为 O(n)，实际上单报文流的数量变化较大，不发生大规模攻击时 O(n)；③的存储空间大小和黑名单更新策略有关，一般不超过 topN 的常数倍 O(c*topN)，其中 c 为常数，topN O(n)，N 值普遍较小，试验中 2.5Gbps 链路下 N 全部小于 1000。

处理速度 IDSampling 的处理复杂度可分为两部分：预处理复杂度和对每个报文的抽样复杂度。预处理复杂度影响抽样指导策略的更新效率，抽样复杂度影响报文实时处理效率。预处理负责统计单报文属性熵，其复杂度为 O(n)。另外还需要计算异常流量的 top-N，目前 top-N 的计算方法已经有很多快捷的方法来实现[15]，其复杂度可达到 O(nlogn)，即预处理复杂度为 O(n+nlogn) 。不同抽样策略的抽样复杂度不同：单一抽样策略中每个报文只需二分查找 topN 表并抽样，因此每个报文的处理复杂度为 $O(\log_2 N)$。而对于混合抽样策略每个报文要检查黑名单 $O(\log_2(c*N)$并实时判断其流长才可分情况抽样，总复杂度为 $O(\log_2(c*N)+1)$。

误差分析 IDSampling 利用长流标记反馈方法标记攻击流，其误差对最终的抽样检测率有直接影响。而打标记方法建立在多级 Bloom Filter 的流识别方法上，所以 IDSampling 继承了多级 Bloom Filter 的误差。多级 Bloom Filter 的一大优点是错误否定率为零，即不可能将攻击流误判为非攻击流而漏抽样。由于每级采用了独立的 hash 空间，多级 Bloom Filter 的错误肯定率随着级数的增加指数级减小，其错误肯定率为 $FPR_s \approx (\frac{n}{b})^k$，其中 $n << b$。每个时间片平均有 $n*FPR_s$ 个非攻击流会误判为攻击流而被抽样。

上述理论分析表明 IDSampling 算法速度快、存储要求低、标记误差小。对于 1Gbps 的链路，我们按平均每个报文包含 500 个字节估算，时间片为 1 分钟，则最大活动报文数 k 约等于 $1.2*10^7$，最大活动流数 n 不超过 k，一般情况下 n 的数量级为 10^6。为了保证精度 $n << b$，多级 Bloom Filter 采用 24 位 hash 函数。可知采用 4 级 Bloom Filter 要用 16M *4=64M 内存。此时 Bloom Filter 的错误肯定率 $FPR=(1.2*10^6/2^{24})^4=1.09*10^{-6}$。一个流的四元组属性占用 12B，则单报文属性熵统计空间最多不超过 12M。考虑处理速度时每个时间片花在单报文属性熵统计的时间不超过 0.1 秒（当 n 数量级为 10^6 时）[16]。即每个时间片中 IDSampling 最多滞后 0.1 秒得到属性熵的抽样指导。由于流量的长相关和自相似特性，这 0.1 秒内使用上一时间片属性熵的抽样指导是可以接受的。单一抽样策略要检查的 top-N 流和混合抽样策略要检查的黑名单都是流量聚类特性，而异常攻击的高重复率导致 top-N 中的 N 值普遍较小，即使 N 达到 10^3，每个报文比较次数也不超过 $\log_2 N$，约 10 次。因此 IDSampling 速度足以胜任 1Gbps 链路的线速处理。10Gbps 链路下上述 k、n 值增加 10 倍，其它各项指标都增加 10 倍，此时瓶颈存在于①4 级 Bloom Filter 的错误肯定率 $FPR=(1.2*10^7/2^{24})^4 \approx 30\%$，准确度已经不可接受；②即使是采用 hash 技术的多级 BloomFilter 也不能胜任实时计算如此高速链路的报文长度。③对如此多的报文无法实时进行如此多的 topN 比较，这三个瓶颈限制决定了 IDSampling 适于工作在 1~3Gbps 的链路下，但不适于更高速的链路。

4 实验结果及分析

如前所述，目前要实现对 10G+bps 骨干网环境下的高速入侵检测，主要有硬件和分布式系统两种方式。完全由硬件实现的高速入侵检测，如采用现场可编程门阵列（FPGA）技术的背板，由于对报文负载的匹配仍采用软件编程实现，因而速度提高有限，不可能在前端完成所有入侵检测需要的功能。因而更普遍的解决办法是采用分布式系统。采用 10Gbps 数据采集器（比如烽火 FH-SN-TC1000）将 OC-192 信道上的报文分为若干个 1Gbps 的流量数据输出，之所以要分成若干 1Gbps 流量是由数据处理设备以太网的硬件接口规格限制的。因此我们考察 IDSampling 在 1Gbps 链路下的抽样精度更具有普遍意义。

实验在1Gbps的链路上采用开源的Snort作为IDS，IDS实验机硬件配置为Pentium 4 Xeon 2.4G双核CPU，2G内存。为了量化IDSampling的抽样准确度，实验通过比较在不同种类和规模的攻击下三种抽样算法：随机流抽样算法（RamFlow）、随机报文抽样算法（RamPak）和IDSampling能抽取到的攻击报文数目来说明IDSampling对攻击的抽样准确度远优于其它两种经典抽样方法。

实验主要分为三步：⑴生成背景流量。采集CERNET华东北地区网到江苏省网的某条2.5Gbps链路的真实流量报头信息（真实流量800Mbps~1.3Gbps），随机产生64B~1500B的报文负载，然后组合起来作为背景流量。⑵生成攻击流量。由于传统的DARPA公开测试数据集攻击数据量规模过小并不适合作为实验的数据源，我们改造DARPA公开测试数据集的部分数据，生成大规模攻击测试数据集。这一步工作最复杂，也分为三步：①选择攻击类型。实验中采用的攻击类型其编号和名称如表2所示，它们主要来自DARPA1999攻击测试数据集第4周和第5周的攻击测试数据集，根据攻击序号可在DARPA数据集中查找到该攻击细节说明。之所以选择它们首先是因为它们必须是Snort能检测到的攻击，其次它们涵盖了所有4种攻击类型，由于骨干网更加关心大规模攻击的检测，因此更多选择了不同规模的Probe和Dos攻击。②截取攻击数据。实验中按攻击编号在相应的tcpdump文件中按5元组特性（攻击时间段、源地址、源端口、宿地址、宿端口）来截取攻击数据，即在攻击时间段内从源地址、源端口到宿地址、宿端口的所有报文全都是攻击报文。③控制生成不同规模攻击。通过修改攻击数据的IP地址前缀，复制生成其他网段上相同的攻击数据，如此可控制生成原攻击n倍的攻击数据。即每秒钟的攻击强度和攻击规模都扩大到了原始DARPA数据集的n倍。如表2所示，我们将Portsweep由原来38个攻击者对1个C类地址中每个主机3个端口的扫描扩大了16倍，变为对16个C类地址的扫描，每秒产生16*38*3个扫描报文，其它攻击的规模和意义如此类推。生成攻击流量后，最后一步工作是⑶将上述生成的攻击流量与背景流量混合在一起。分别以400Mbps和800Mbps的速度向IDS播放混合数据，在IDS端采取不同的三种抽样方法动态抽样，然后通过比较攻击检测率A来衡量不同抽样方法在不同抽样率下对攻击报文的抽样能力，其中A=抽样后能检测到的攻击数/未抽样时检测到的攻击数。每种攻击持续时间约5分钟，实验结果如图2和图3所示，实验数据均取5次实验的平均值。实验中IDSampling的主要参数设定如下：$Z_{H(SrcIP)}$：2.6，$Z_{H(SrcPort)}$：1.7，$Z_{H(DstIP)_1}$：2.2，$Z_{H(DstPort)}$：1.4。

表2 实验攻击列表

攻击类型	攻击编号	攻击名称	攻击序号	攻击报文数目	攻击扩大倍数n
Probe	1	Portsweep	52.211313	16*254*38*3	16
	2	Ipsweep	45.192523	16*254*9	16
Dos	3	Syn-flood	44.164944	8*254*9*5	8
	4	Smurf	43.144547	8*254*9*2	8
	5	Mailbomb	42.155148	4*6667	4
	6	teardrop	44.082615	4*254*7	4
R2L	7	Snmpget	43.191217	268	1
	8	named	45.130542	39	1
U2R	9	loadmodule	45.165009	17	1

表2和表3比较了分别以400Mbps和800Mbps的速率下播放混合数据时三种抽样方法的抽样准确度。400Mbps、800Mbps两种速率下IDS的动态抽样率分别在[1/3，1/6]和[1/9，1/13]之间变化。首先，在任何抽样率对于所有9种攻击IDSampling能抽样到的攻击报文数都是最高的，尤其是对于前6种大规模攻击，IDSampling能抽样到的攻击报文数甚至高出流抽样和报文抽样一个数量级，比如800Mbps下IDSampling能抽样到369246个Portsweep报文，而RamFlow和RamPak分别只能抽样到61618和34283个Portsweep报文。这说明单一抽样策略准确地定位到了“最异常”的流量，因此大幅提升了对大规模异常攻击的抽样准确性。其次，对于那些持续时间较长的小规模攻击，比如Snmpget（268个攻击报文、持续时间近3分钟），400Mbps下IDSampling采用长流反馈技术标记攻击流，可以抽样到64.9%的攻击报文，远高于RamPak和RamFlow分别4.4%和0的检测率。而800Mbps下由于抽样率降低，IDSampling不能在第一时间片内将攻击流判为长流，因此也不能再受益于长流反馈技术，所以此时IDSampling的抽样率会锐减至5.2%。对于那些规模又小，

持续时间很短的攻击，如 named 和 loadmodule，三种算法能抽到的报文数都很少，IDSampling 因其对短流的抽样偏好，最初的几个攻击报文几乎总能被抽到，IDSampling 仍优于其余其他两种。

如图所示，RamPak 和 RamFlow 的抽样能力基本持平。理论上由于 RamFlow 平等地看待流，对大规模异常攻击来说能保留更多的攻击信息，因而抽样能力应该优于 RamPak。但是试验结果表明在高抽样率下[1/3~1/6]，RamFlow 的优势并不明显，有时甚至低于 RamPak。随着抽样率的降低，800Mbps 下 RamFlow 对大规模攻击的抽样检测率才渐渐优于 RamPak。另一方面，RamPak 抽样会偏向于超长流这一点在对 mailbomb 的抽样结果中表现得相当突出，两种速率下 RamPak 抽样数据的攻击检测率均大幅高于 RamFlow，分别达到 22.70%和 12.79%。

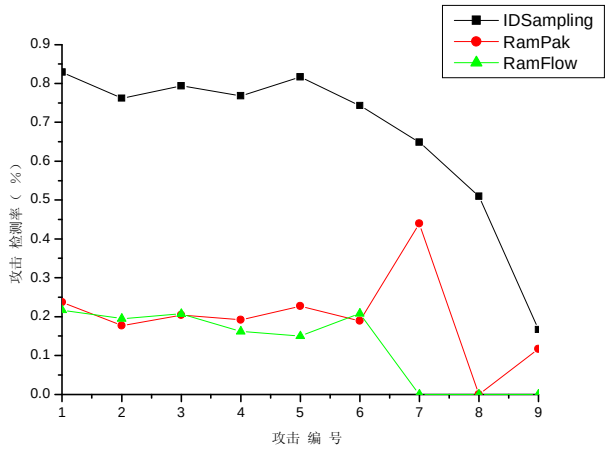

图 2 400Mbps 流速下三种才抽样算法的抽样精度比较图

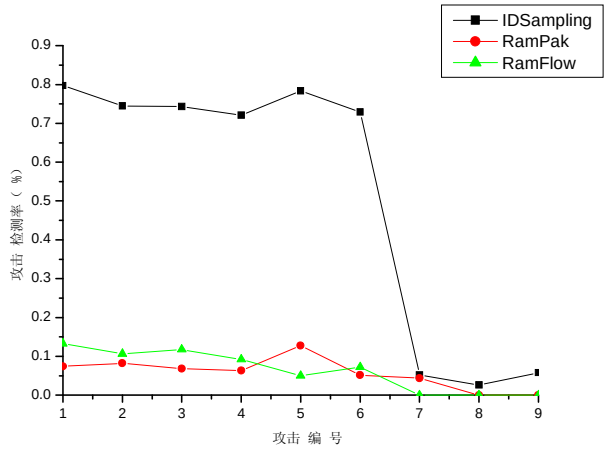

图 3 800Mbps 流速下三种才抽样算法的抽样精度比较图

IDSampling 的另一个优势在于 IDSampling 对抽样率衰减不敏感。当抽样率从 [1/3，1/6]降至[1/8，1/13]时 IDSampling 对 6 种大规模攻击的的抽样检测率降低得最多的也不到 4%，而 RamFlow 降低了 10%左右，RamPak 对抽样率降低最敏感，抽样检测率最多的降低了 14%。

实验证明采用 IDSampling 使得极限处理能力只不到 100 Mbps 的开源入侵检测系统 Snort 能高效处理 1Gbps 的流量。对于大规模攻击它能够抽样到的攻击报文数目较之流抽样和报文抽样要高一个数量级，完全

可以保障大规模攻击的检测精度以及趋势性信息的准确性。而对于小规模攻击 IDSampling 只能尽力而为，虽然不能保证检测精度，但是其保证在攻击开始阶段对报文高抽样比检测，借助单报文流聚类信息和反馈标记方法提高了 IDSampling 对攻击报文的抽样准确度，其抽样攻击报文的能力仍旧高于其他两种抽样方法。

5 总结与展望

本文提出了一个面向骨干网入侵检测的动态自适应抽样算法 IDSampling。不同于普通流抽样算法，IDSampling 采用变化的抽样率，当大规模异常攻击发生时，以单一报文属性熵为指导抽样出现频率最多，“多对一”汇聚特征最明显的异常流量。缺省则按照攻击流量的流长特征，对不同长度的流采取了不同的抽样率，同一个流的不同时期抽样率也不同，最大限度地分配抽样机会给攻击嫌疑度大的报文。实现了用最小的检测代价获取最大安全的目的。其算法简单，速度快，存储量小。实验表明采用 IDSampling 抽样方法，极限处理能力只不到 100 Mbps 的开源入侵检测系统 Snort 能高效处理 1Gbps 的流量，保证对大规模攻击的检测精度及其攻击趋势性信息的准确性。这样借助诸如烽火 FH-SN-TC1000 等 10Gbps 的数据采集器，使用最多 10 个 Snort 就可以应付 10Gbps 的高速入侵检测，这个解决方案的性价比远高于迄今已知的各种 IDS 分布式结构[4,5]。

IDSampling 的不足之处在于缺乏相应的抽样补偿策略，因此不能得到更好的修正结果。因为 IDSampling 对于攻击特征的抽样偏好不能简单地乘以抽样率倒数来估计抽样前的攻击流/报文数目，这样会过于夸大攻击规模，所以开发相应的抽样补偿策略是我们的下一步工作。另外文章还存在很多可以继续深入的研究：如研究怎样从理论上保证检测到一个复合攻击所需要具备的信息量，当抽样率太低不足以提供足够的信息量时干脆放弃检测此攻击而将检测资源留给其他攻击检测以及研究攻击步骤间的互信息量等。这些研究可以进一步指导我们更有效地分配抽样机会、提高攻击检测率，获得更有效的面向更高速网络入侵监测的抽样方法。

References:

[1] H. Bos and Kaiming Huang, Towards Software-Based Signature Detection for Intrusion Prevention on the Network Card. RAID 2005,Vrije University, The Netherlands Xiamen University,China.

[2] Y. Cho and W. Mangione-Smith. Fast reconfiguring deep packet filter for 1+gigabit network, In IEEE Symposium on Field-Programmable Custom Com[putting Machines, (FCCM), NaPa, CA, April 2005.

[3] R.Fanklin, D.Caraver, and B. Hutchings. Assisting network intrusion detection with reconfigurable hardware. In Proceedings from filed Programmable Custom Computing Machines. 2002.

[4] Matthias Vallentin, Robin Sommer. The NIDS Cluster: Scalable, Stateful Network Intrusion Detection on Commodity Hardware[c], In: Proc. Symposium on Recent Advances in Intrusion Detection, LNCS vol.4637.p107-126,Queensland, Australia,Sep.2007.

[5] I.Charitakis, K.Anagnostakis, and E.Markatos, “An active traffic splitter architecture for intrusion detection,” in Proceedings of 11th IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS 2003), Orlando, October 2003, pp. 238–241.

[6] A.Sridharan, T.Ye, and S.Bhattacharyya. Connection Port Scan Detection on the Backbone. In malware Workshop held in conjunction with IPCC, Phoenix, Arizona, USA, April 2006.

[7] J.Jung, V. Paxson, Using Sequential Hypothesis Testing. In Proc. Of 2004 IEEE Symposium on Security and Privacy, Oakland, CA, USA, May 2004.

[8] J Mai, CN Chuah, A Sridharan, T Ye, H Zang, Is sampled data sufficient for anomaly detection? In Proc. of the 6th ACM SIGCOMM on Internet measurement, Brazil, 2006.

[9] A.Lakhina, M.Crovella, and C.Diot. Mining Anomalies Using Traffic Feature Distributions. In Proc. ACM SIGCOMM ’05, Philadelphia, PA, USA, Aug. 2005.

[10] D Brauckhoff, B Tellenbach, A Wagner, Impact of packet sampling on anomaly detection metrics**,** In Proc. ACM SIGCOMM’06 Rio de Janeriro, Brazil,2006.

[11] Mingzhou Zhou, Study of large-scale network IP flows behavior characteristics and measurement algorithms, PHD dissertation. School of Computer Science and Technology, 2006.10.

[12] “The test method of Spirent Firewall”, http://www.chinaunix.net.

[13] Subhabrata Sen and Jia Wang. Analyzing Peer-to-Peer Traffic Across Large Networks. IEEE/ACM Transactions on Networking, 2004.

[14] Estan C, Varghese G. New Directions in Traffic Measurement and Accounting[C].In: SIGCOMM2002, August 2003, Pages 270–313.

[15] WenJ, Anthony KHT. Mining top-N local outliers in large databases, In: ACM SIGLDD Int'l on knowledge Discovery and Data Mining, San Francisco, New York: ACM Press,2001.293-298.

[16] Feng Hu, Guoyin Wang, Analysis of the complexity of Quick sort for Two Dimension Table, Chinese Journal of Computers, 2007 Vol.30 No6 P.963-968.

附中文参考文献:

[11] 周明中.大规模网络 IP 流行为特性及其测量算法研究.博士论文,东南大学,2006.7