一种面向流量异常检测的概率流抽样方法

董书琴* 张 斌
(信息工程大学 郑州 450001)
(河南省信息安全重点实验室 郑州 450001)

摘 要：针对基于概率抽样的网络流量异常检测数据集构造过程中无法同时兼顾大、小流抽样需求及未区分flash crowd与流量攻击等问题，该文提出一种面向流量异常检测的概率流抽样方法。在对数据流按目的、源IP地址进行分类的基础上，将每类数据流抽样率定义为其目的、源IP地址抽样率的最大值，并在抽样过程中对数据流抽样数目向上取整，保证每类数据流至少被抽样一次，使抽样得到的数据集可有效反映原始流量在大、小流和源、目的IP地址方面的分布性。采用源IP地址熵刻画异常流源IP地址分散度，并基于源IP地址熵阈值设计攻击流抽样算法，降低由flash crowd引起的非攻击异常流抽样概率。仿真结果表明，该方法能同时满足大、小流抽样需求，具有较强的异常流抽样能力，可抽样到所有与异常流相关的可疑源、目的IP地址，并能在抽样过程中过滤非攻击异常流。

关键词：网络流量；异常检测；流抽样；概率抽样
中图分类号：TP393 **文献标识码**：A **文章编号**：1009-5896(2019)06-1450-08
DOI: 10.11999/JEIT180631

A Probabilistic Flow Sampling Method for Traffic Anomaly Detection

DONG Shuqin ZHANG Bin
(*Information Engineering University*, *Zhengzhou* 450001, *China*)
(*Henan Key Laboratory of Information Security*, *Zhengzhou* 450001, *China*)

Abstract: For problems of not meeting the demand of sampling both large flows and small flows at the same time, and not distinguishing flash crowd from traffic attacks in building network traffic anomaly detection datasets based on probabilistic sampling methods, a probabilistic flow sampling method for traffic anomaly detection is proposed. On the basis of the classification of network data flows according to their destination and source IP addresses, the sampling probability for each class of data flows is set as the maximum of its destination and source IP address's sampling probability, and the number of sampled data flows is ceiled to ensure that each class of data flows is sampled at least once, so that the sampled dataset can reflect the distributions of large, small flows and source, destination IP addresses in original traffics. Then, the source IP address entropy is used to characterize the source IP dispersion of anomaly flows, and the attack flow sampling algorithm is designed based on the threshold of the source IP address entropy, which reduces the sampling probability of non-attack anomaly flows caused by flash crowd. The simulation results show that the proposed method can satisfy the sampling requirements of both large flows and small flows, it has a high anomaly flows sampling ability, can sample all the suspicious sources and destination IP addresses related to anomaly flows, and can effectively filter the non-attack anomaly flows.

Key words: Network traffic; Anomaly detection; Flow sampling; Probabilistic sampling

1 引言

网络流量异常检测是发现网络攻击的重要手段，常用的流量异常检测方法有熵理论[1]和神经网络[2]等，当前研究通常集中于对现有方法的改进，实验中用到的数据集大部分采用全采集测量方式通过在线监测骨干网流量获得。然而随着网络速度的迅速提升和网络流量的指数级增长，采用全采集测

收稿日期：2018-06-28；改回日期：2019-01-15；网络出版：2019-01-28
*通信作者： 董书琴 dongshuqin377@126.com
基金项目：河南省基础与前沿技术研究计划基金(142300413201)，信息工程大学新兴科研方向培育基金(2016604703)
Foundation Items: The Foundation and Frontier Technology Research Project of Henan Province (142300413201), The New Research Direction Cultivation Fund of Information Engineering University (2016604703)

量方式对骨干网流量进行实时在线监测变得越来越困难。为解决上述问题，网络流量抽样技术已成为国内外学者研究的重点。

网络流量抽样技术主要分为包抽样和流抽样两种，其中流抽样技术能有效体现不同数据包间的关联性，抽样得到的数据集可较好地反映网络流量原始特征信息[3]，是抽样技术研究的热点。在流抽样技术中，概率抽样方法具有系统开销小、数据流处理速度快等特点，更适用于对实时性要求较高的在线流量监测领域。

依据关注的数据流大小不同，常用的概率抽样方法主要分为两类：第1类是大流抽样技术，如智能抽样方法[4]、抽样保持方法[5]等，该类方法中每条数据流的抽样概率与其包含的字节数成正相关性，因此包含字节数较多的大流抽样概率高，而包含字节数较少的小流抽样概率较低，抽样过程中易损失大量以小流形式存在的异常流量信息，不适用于流量异常检测领域。第2类是小流抽样技术，如选择性抽样方法[6]、优化神经网络方法[7]、异常流自适应抽样算法[8]等，该类方法中每条数据流的抽样概率与其包含的字节数成负相关性，因此小流抽样概率高，抽样得到的数据集可有效保留与异常流量相关的小流信息，但其对大流的抽样概率较低，抽样过程中会产生比较大的信息损失。综上，现有根据数据流大小设定其抽样概率的方法，难以同时满足大流和小流抽样需求，导致抽样得到的数据集无法完整反映真实网络数据流组成情况。

为此，文献[9]提出一种新的概率抽样方法，在对所有数据流按目的IP地址进行分类的基础上，将每类数据流的抽样率定义为其目的IP抽样率，并通过调节数据流抽样率来控制该类数据流中大、小流抽样概率，从而使得访问某一目的IP地址域的大、小流都能被有效抽样，同时提高了受攻击目的IP地址的异常流抽样概率。然而由于网络中异常流量的多样性及源、目的IP地址分布的不均衡性，导致采用该方法抽样得到的数据集在应用于流量异常检测时，仍存在以下不足：该方法通常仅能对与目的IP地址相关的异常流保持较高的抽样概率，而对与源IP地址相关的异常流抽样概率较低，导致抽样得到的数据集无法反映与源IP地址相关的异常流量信息，进而影响后续流量异常检测精度。

此外，为有效提高网络安全态势感知的准确性，流量异常检测模型需要对flash crowd和网络攻击进行区分[10–12]，但由于网络环境的复杂性，在异常检测阶段区分flash crowd和网络攻击存在误报率高、耗时长等问题。采用抽样技术在流量异常检测数据集构造过程中降低flash crowd流量抽样概率，是解决上述问题的一种有效方法。但现有方法在抽样过程中通常未区分flash crowd与流量攻击两种情况产生的异常数据流，导致抽样得到的数据集中同时包含两种异常流量，进而加大了后续流量异常检测的误报率。

针对上述两个问题，结合目的、源IP地址抽样率，设计异常流抽样算法，提高异常流抽样的全面性。然后，采用源IP地址熵对异常流的源IP地址分散度进行刻画，并通过设置攻击流源IP地址熵阈值，在抽样过程中降低由flash crowd产生的非攻击异常流抽样概率。最后，通过仿真实验验证所提方法的有效性。

2　概率流抽样方法

2.1　基本思想

主要采取以下思想对概率流抽样方法进行设计：

(1)定义数据流抽样率为其目的、源IP地址抽样率的最大值。首先将具有相同目的、源IP地址的数据流划分为一类，然后将每类数据流的抽样率定义为其目的、源IP地址抽样率的最大值，从而保证与目的、源IP地址相关的异常流量都能以最大概率被抽样；

(2)设定攻击流源IP地址熵阈值。采用源IP地址熵刻画异常流源IP地址分散度，并通过设定攻击流源IP地址熵阈值，在抽样过程中降低源IP地址熵大于所设阈值的非攻击异常流抽样概率。

2.2　异常流抽样算法

假设在理想网络环境中，流量异常都是由流量攻击引起的。在这种情况下，定义异常流抽样函数，提高网络数据流中的异常流抽样概率，并给出异常流抽样流程。

2.2.1　抽样函数

异常流抽样函数主要由目的IP地址接收字节总数和源IP地址发送字节总数两个数据流特征参数确定。下面从目的IP地址出发，对研究过程中用到的相关概念进行定义，并给出异常流抽样函数。

定义1　定义FS表示抽样前给定抽样数据集的数据流样本空间

$$\mathrm{FS}=\{\mathrm{fl}_1,\mathrm{fl}_2,\cdots,\mathrm{fl}_n,\cdots,\mathrm{fl}_N\} \tag{1}$$

其中，N为正整数，表示抽样前给定抽样数据集中包含的数据流总数；fl_n表示FS中的第n条数据流，$1\le n\le N$。

定义2　定义dIPS表示抽样前FS的目的IP地址空间

$$\mathrm{dIPS}=\{\mathrm{dIP}_1,\mathrm{dIP}_2,\cdots,\mathrm{dIP}_m,\cdots,\mathrm{dIP}_M\} \tag{2}$$

其中，M为正整数，表示FS中数据流的唯一目的IP地址总数；dIP_m表示dIPS中的第m个目的IP地址，$1 \le m \le M$。

对于dIPS中的元素dIP_m而言，其数据流空间和源IP地址空间定义如下：

定义3 定义$\mathrm{fS}(\mathrm{dIP}_m)$表示抽样前FS中$\mathrm{dIP}_m$对应的数据流空间

$$\mathrm{fS}(\mathrm{dIP}_m)=\{f_1\cdot\mathrm{dIP}_m, f_2\cdot\mathrm{dIP}_m,\cdots, f_l\cdot\mathrm{dIP}_m,\cdots, f_L\cdot\mathrm{dIP}_m\} \tag{3}$$

其中，L为正整数，表示FS中流向dIP_m的数据流总数；$f_l\cdot\mathrm{dIP}_m$表示$\mathrm{fS}(\mathrm{dIP}_m)$中的第$l$条数据流，$1 \le l \le L$; $\mathrm{fS}(\mathrm{dIP}_m) \subset \mathrm{FS}$，且$\mathrm{FS}=\bigcup_{m=1}^{M}\mathrm{fS}(\mathrm{dIP}_m)$。

定义4 定义$\mathrm{sIPS}(\mathrm{dIP}_m)$表示抽样前$\mathrm{fS}(\mathrm{dIP}_m)$中数据流的源IP地址空间

$$\mathrm{sIPS}(\mathrm{dIP}_m)=\{\mathrm{dIP}_m\cdot\mathrm{sIP}_1, \mathrm{dIP}_m\cdot\mathrm{sIP}_2,\cdots, \mathrm{dIP}_m\cdot\mathrm{sIP}_k,\cdots, \mathrm{dIP}_m\cdot\mathrm{sIP}_K\} \tag{4}$$

其中，K为正整数，表示$\mathrm{fS}(\mathrm{dIP}_m)$中数据流的不同源IP地址总数；$\mathrm{dIP}_m\cdot\mathrm{sIP}_k$表示$\mathrm{sIPS}(\mathrm{dIP}_m)$中的第$k$个源IP地址，$1 \le k \le K$。

对于$\mathrm{sIPS}(\mathrm{dIP}_m)$中的元素$\mathrm{dIP}_m\cdot\mathrm{sIP}_k$而言，其数据流空间定义如下：

定义5 定义$\mathrm{fS}(\mathrm{dIP}_m\cdot\mathrm{sIP}_k)$表示抽样前在FS中$\mathrm{dIP}_m\cdot\mathrm{sIP}_k$对应的数据流空间

$$\begin{aligned}\mathrm{fS}(\mathrm{dIP}_m\cdot\mathrm{sIP}_k)=\{&f_1\cdot\mathrm{dIP}_m\cdot\mathrm{sIP}_k, f_2\cdot\mathrm{dIP}_m\cdot\mathrm{sIP}_k,\cdots,\\ &f_p\cdot\mathrm{dIP}_m\cdot\mathrm{sIP}_k,\cdots, f_P\cdot\mathrm{dIP}_m\cdot\mathrm{sIP}_k\}\end{aligned} \tag{5}$$

其中，P为正整数，表示FS中由$\mathrm{dIP}_m\cdot\mathrm{sIP}_k$流出的数据流总数；$f_p\cdot\mathrm{dIP}_m\cdot\mathrm{sIP}_k$表示$\mathrm{fS}(\mathrm{dIP}_m\cdot\mathrm{sIP}_k)$中的第$p$条数据流，$1 \le p \le P$; $\mathrm{fS}(\mathrm{dIP}_m\cdot\mathrm{sIP}_k) \subset \mathrm{FS}$。

对于FS中源IP地址为$\mathrm{dIP}_m\cdot\mathrm{sIP}_k$，目的IP地址为$\mathrm{dIP}_m$的数据流而言，定义其数据流空间如下：

定义6 定义$f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}$表示抽样前FS中源IP地址为$\mathrm{dIP}_m\cdot\mathrm{sIP}_k$、目的IP地址为$\mathrm{dIP}_m$的数据流空间

$$f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m} \subset \mathrm{fS}(\mathrm{dIP}_m)\cap\mathrm{fS}(\mathrm{dIP}_m\cdot\mathrm{sIP}_k) \tag{6}$$

$f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}$是概率流抽样方法的抽样对象，为有效提高异常流抽样概率，结合目的IP地址抽样率[9]及其对应的源IP地址抽样率，给出其抽样函数如式(7)

$$p(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m})=\mathrm{Max}\{p(\mathrm{dIP}_m), p(\mathrm{dIP}_m\cdot\mathrm{sIP}_k)\} \tag{7}$$

其中，

$$p(\mathrm{dIP}_m)=\begin{cases} b, & \mathrm{S\_dIP}_m < t \\ \dfrac{\mathrm{S\_dIP}_m}{\mathrm{S\_tl}}, & \mathrm{S\_dIP}_m \ge t \end{cases} \tag{8}$$

$$p(\mathrm{dIP}_m\cdot\mathrm{sIP}_k)=\frac{\mathrm{S\_dIP}_m\cdot\mathrm{sIP}_k}{\mathrm{S\_tl}} \tag{9}$$

式(8)、式(9)中，t表示抽样阈值；b为一个较小的常数，$0 < b \le 1$; $\mathrm{S_dIP}_m$表示dIP_m接收字节总数，$\mathrm{S_dIP}_m\cdot\mathrm{sIP}_k$表示$\mathrm{dIP}_m\cdot\mathrm{sIP}_k$发送字节总数，S_tl表示样本空间字节总数，具体定义如下：

定义7 定义$\mathrm{S_dIP}_m$表示抽样前dIP_m接收字节总数

$$\mathrm{S\_dIP}_m=\sum_{l=1}^{L}(\mathrm{oc}(f_l\cdot\mathrm{dIP}_m)) \tag{10}$$

其中，$\mathrm{oc}(f_l\cdot\mathrm{dIP}_m)$表示$\mathrm{fS}(\mathrm{dIP}_m)$中第$l$条数据流包含的字节数。

定义8 定义$\mathrm{S_dIP}_m\cdot\mathrm{sIP}_k$表示抽样前$\mathrm{dIP}_m\cdot\mathrm{sIP}_k$发送字节总数

$$\mathrm{S\_dIP}_m\cdot\mathrm{sIP}_k=\sum_{p=1}^{P}\mathrm{oc}(f_p\cdot\mathrm{dIP}_m\cdot\mathrm{sIP}_k) \tag{11}$$

其中，$\mathrm{oc}(f_p\cdot\mathrm{dIP}_m\cdot\mathrm{sIP}_k)$表示$\mathrm{fS}(\mathrm{dIP}_m\cdot\mathrm{sIP}_k)$中第$p$条数据流包含的字节数。

定义9 定义S_tl表示抽样前样本空间FS中包含的字节总数

$$\mathrm{S\_tl}=\sum_{m=1}^{M}\mathrm{S\_dIP}_m=\sum_{n=1}^{N}\mathrm{oc}(\mathrm{fl}_n) \tag{12}$$

其中，$\mathrm{oc}(\mathrm{fl}_n)$表示FS中第$n$条数据流包含的字节数。

2.2.2 异常流抽样流程

在概率流抽样方法基本思想的基础上，结合异常流抽样函数，设计异常流抽样流程如图1所示。每类数据流经抽样后得到的数据流数目等于其包含的流数与抽样率的乘积，对于抽样数目为小数的，最终抽样量向上取整，保证每类数据流都至少被抽样一条，从而使抽样得到的数据集可有效反映原始流量的分布性。

该抽样流程具体描述如表1所示。

2.3 攻击流抽样算法

真实网络环境中，流量异常并非全部由流量攻击引起，flash crowd同样可引起网络流量异常，且这种异常通常与拒绝服务攻击具有相似的表象。两种情况下，目的IP地址接收的字节总数都比较大(通常不小于t)，从而导致两种情况下流向受攻击目的IP地址的数据流都具有较高的抽样概率，使得抽样得到的数据集中既包含flash crowd数据流，也

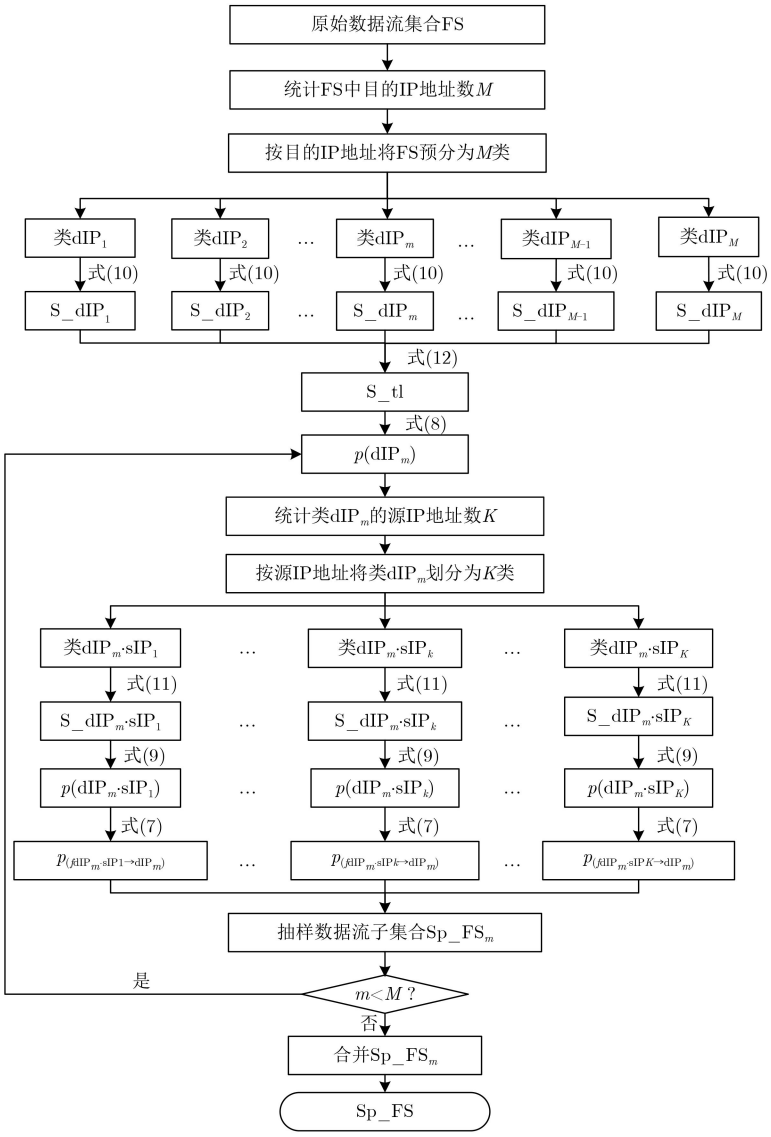

图 1 异常流抽样流程

表 1　抽样流程具体描述

输入：原始数据流集合FS

输出：抽样后的数据流集合Sp_FS

步骤1　统计FS中目的IP地址数M，根据目的IP地址将FS分成M类，即：类dIP$_1$，类dIP$_2$，…，类dIP$_m$，…，类dIP$_M$；

步骤2　分别统计M类的数据流空间大小L，依次采用式(10)计算其接收字节总数S_dIP$_1$, S_dIP$_2$, …, S_dIP$_m$, …, S_dIP$_M$，并采用式(12)计算样本空间字节总数S_tl；

步骤3　对于类dIP$_m$而言，采用式(8)计算目的IP地址dIP$_m$的抽样率p(dIP$_m$)；

步骤4　统计类dIP$_m$的唯一源IP地址数K，并将类dIP$_m$进一步划分为K类，即：类dIP$_m\cdot$sIP$_1$, …, 类dIP$_m\cdot$sIP$_k$, …，类dIP$_m\cdot$sIP$_K$；

步骤5　分别统计K类的数据流空间大小P，采用式(11)计算各类发送字节总数S_dIP$_m\cdot$sIP$_1$, …, S_dIP$_m\cdot$sIP$_k$, …, S_dIP$_m\cdot$sIP$_K$，并采用式(9)计算dIP$_m$对应的K个源IP地址抽样率p(dIP$_m\cdot$sIP$_1$), …, p(dIP$_m\cdot$sIP$_k$), …, p(dIP$_m\cdot$sIP$_K$)；

步骤6　依次采用式(7)计算类dIP$_m$中由不同源IP地址流向dIP$_m$的数据流抽样率$p(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_1\to\mathrm{dIP}_m})$, …, $p(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m})$, …, $p(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_K\to\mathrm{dIP}_m})$，并输出抽样数据流子集合Sp_FS$_m=\bigcup_{k=1}^{K}\lceil f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}p(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m})\rceil$；

步骤7　如果$m<M$，则返回步骤3；如果$m=M$，则迭代终止，输出最终抽样后的数据流集合Sp_FS$=\bigcup_{m=1}^{M}$ Sp_FS$_m$。

包含拒绝服务攻击数据流，容易增大后续流量异常检测的误报率。

针对上述问题，在分析flash crowd与流量攻击两种情况产生的异常流区别的基础上，引入熵的概念对异常流源IP地址分散度进行刻画，并定义攻击流源IP地址熵阈值，有效降低由flash crowd产生的异常流抽样概率，然后重新定义此种情况下dIP_m抽样率及$f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}$抽样函数，进而给出攻击流抽样流程。

2.3.1 源IP地址熵

Flash crowd和流量攻击产生的异常流通常都表现为“多对一”的模式，即异常流中包含的源IP地址数较多，而目的IP地址数较少。但两者也有明显的区别，flash crowd通常是由位于多处的正常用户发起，异常流的源IP地址通常比较分散；流量攻击通常由某些恶意用户发起，异常流的源IP地址大部分相对集中。为此，引入源IP地址熵对异常流的源IP地址分散度进行刻画。

定义10 定义$H(\mathrm{dIP}_m)$表示目的IP地址为dIP_m的异常流源IP地址熵

$$H(\mathrm{dIP}_m)=-\sum_{k=1}^{K}P(\mathrm{dIP}_m\cdot\mathrm{sIP}_k)\cdot\lg\left(P(\mathrm{dIP}_m\cdot\mathrm{sIP}_k)\right)\tag{13}$$

其中，

$$P(\mathrm{dIP}_m\cdot\mathrm{sIP}_k)=\mathrm{card}\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}\right)/\mathrm{card}\left(\mathrm{fS}\left(\mathrm{dIP}_m\right)\right)\tag{14}$$

式中，$\mathrm{card}\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}\right)$, $\mathrm{card}\left(\mathrm{fS}\left(\mathrm{dIP}_m\right)\right)$分别表示两个集合的基数，即两个集合中元素的个数。

根据熵的定义可知，$H(\mathrm{dIP}_m)$取值越小，表明目的IP地址为dIP_m的异常流中源IP地址分布越集中，该数据流为攻击流的可能性就越大；反之，$H(\mathrm{dIP}_m)$取值越大，表明目的IP地址为dIP_m的异常流中源IP地址分布越分散，该数据流为flash crowd流的可能性越大[12]。

2.3.2 攻击流抽样流程

首先设定攻击流$H(\mathrm{dIP}_m)$阈值为T，然后通过调整T的大小，降低flash crowd产生的异常流抽样概率。定义Q表示dIP_m对应的数据流是否为攻击流的情况，当其为攻击流时取值为1，否则取值为0，即

$$Q=\begin{cases}1, H(\mathrm{dIP}_m)<T\\ 0, H(\mathrm{dIP}_m)\geq T\end{cases}\tag{15}$$

此时，dIP_m抽样率可重新定义为

$$p(\mathrm{dIP}_m)^*=\begin{cases}b, & \mathrm{S\_dIP}_m<t\\ \dfrac{\mathrm{S\_dIP}_m}{\mathrm{S\_tl}}Q, & \mathrm{S\_dIP}_m\geq t\end{cases}\tag{16}$$

此种情况下，$f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}$的抽样函数为

$$p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}\right)^*=\mathrm{Max}\left\{p(\mathrm{dIP}_m)^*, p(\mathrm{dIP}_m\cdot\mathrm{sIP}_k)\right\}\tag{17}$$

在上述定义的基础上，只需对表1中步骤3与步骤6进行改进，并保留其余算法步骤，即可得到攻击流抽样算法。改进后的步骤3和步骤6表述如下：

步骤3 对于类dIP_m而言，采用式(13)—式(16)计算目的IP地址dIP_m的抽样率$p(\mathrm{dIP}_m)^*$；

步骤6 依次采用式(17)计算类dIP_m中由不同源IP地址流向dIP_m的数据流抽样率$p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_1\to\mathrm{dIP}_m}\right)^*$, …, $p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}\right)^*$, …, $p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_K\to\mathrm{dIP}_m}\right)^*$，并输出抽样数据流子集合$\mathrm{Sp_FS}_{nt}=\bigcup_{k=1}^{K}\left[f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}\cdot p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}\right)^*\right]$；

同时，只需将图1中的“式(8)”替换为“式(13)—式(16)”，“$p(\mathrm{dIP}_m)$”替换为“$p(\mathrm{dIP}_m)^*$”；“式(7)”替换为“式(17)”，“$p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_1\to\mathrm{dIP}_m}\right)$, …, $p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}\right)$, …, $p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_K\to\mathrm{dIP}_m}\right)$”替换为“$p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_1\to\mathrm{dIP}_m}\right)^*$, …, $p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_k\to\mathrm{dIP}_m}\right)^*$, …, $p\left(f_{\mathrm{dIP}_m\cdot\mathrm{sIP}_K\to\mathrm{dIP}_m}\right)^*$”即可得到攻击流抽样流程图。

由于篇幅所限，此处不再给出详细的攻击流抽样流程图和步骤描述。

3 仿真实验

通过选取CAIDA traces 2013[13], CAIDA traces 2018[14]正常流量数据集和DARPA 1999[15]第2周恶意流量数据集的子集，基于MATLAB R2012a平台，对数据流特征参数进行验证，并对概率流抽样方法进行仿真分析。

3.1 数据流特征参数验证

随机选取DARPA 1999数据集和CAIDA traces 2018数据集的子集，对恶意流量和正常流量的目的IP地址接收字节总数和源IP地址发送字节总数两个数据流特征参数进行分析。

将DARPA 1999恶意流量和CAIDA traces 2018正常流量数据集中的数据流按目的IP地址进行合并分类，然后分别随机选取150个目的IP地址，统计各个目的IP地址接收字节总数的lg对数，如图2所示。

由图2可知，相比CAIDA正常流量数据集而言，DARPA恶意流量数据集中目的IP地址接收字节总数维持在更高水平，表明存在恶意流量攻击时，目的IP地址接收的字节总数较多。因此，从IP

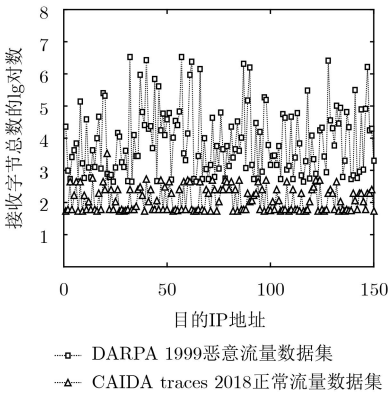

图 2 目的IP地址流量对数分布特征

地址接收字节总数出发设计流抽样方法，可有效提高恶意流量抽样概率。

将DARPA 1999恶意流量和CAIDA traces 2018正常流量数据集中的数据流按源IP地址进行合并分类，并分别随机选取150个源IP地址，统计各个源IP地址发送字节总数的lg对数，如图3所示。

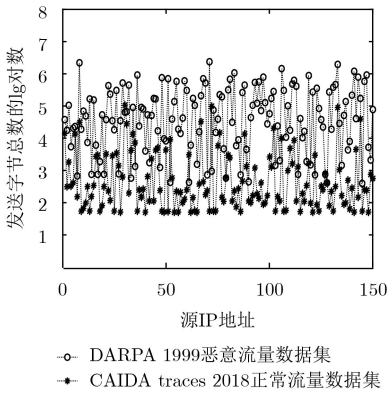

图 3 源IP地址流量对数分布特征

由图3可知，CAIDA正常流量和DARPA恶意流量数据集中源IP地址发送字节总数大都维持在较高水平。相对CAIDA正常流量数据集而言，DARPA恶意流量数据集中源IP地址发送字节总数更多。因此，从源IP地址发送字节总数出发设计流抽样方法，既可提高恶意流量抽样概率，又能保证正常流量抽样概率。

综上可知，结合目的IP地址接收字节总数和源IP地址发送字节总数两个特征参数，设计流抽样方法，可兼顾正常流量和恶意流量抽样概率，从而有效反映原始流量分布特征。

3.2 概率流抽样方法仿真分析

按恶意流量与正常流量1∶5的比例[9]，通过随机选取DARPA 1999, CAIDA traces 2013和CAIDA traces 2018数据集的子集构造数据集Dataset1和Dataset2，每个数据集中包含4000条恶意数据流和20000条正常数据流。其中，Dataset1由DARPA 1999和CAIDA traces 2013的子集构成，Dataset2由DARPA 1999和CAIDA traces 2018的子集构成。仿真过程中，b对结果影响较小，通过分析Dataset1和Dataset2流量分布特征设定b=1.0×10^{-5}；然后参照文献[12]中攻击流的最大熵，设定攻击流$H(\mathrm{dIP}_m)$的阈值T=0.91。对于对抽样结果影响较大的S_dIP$_m$抽样阈值t，则在异常流抽样算法分析过程中通过仿真实验确定。

3.2.1 异常流抽样算法分析

通过将所提异常流抽样算法与经典的小流抽样方法(选择性抽样方法[6])、大流抽样方法(智能抽样方法[4])及文献[9]方法进行对比，分析所提异常流抽样算法性能。首先选取数据集Dataset1，分析抽样阈值$t\in[1,10^6]$时，4种方法抽样得到的异常流数量，仿真结果如图4所示。

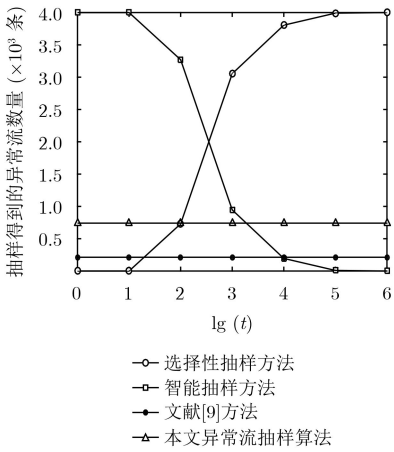

图 4 抽样阈值对不同抽样方法的影响

由图4可知，选择性抽样方法抽样得到的异常流数量随抽样阈值的增大而增大，这是因为随着抽样阈值的增大，小异常流的数量不断增加，而选择性抽样方法仅对小异常流具有较强的抽样能力；智能抽样方法抽样得到的异常流数量随抽样阈值的增大而减小，这是因为随着抽样阈值的增大，大异常流的数量不断减少，而智能抽样方法仅对大异常流具有较强的抽样能力。但上述两种方法无法同时满足大、小流抽样需求。文献[9]方法和本文异常流抽样算法对异常流的抽样能力保持不变，能同时满足大异常流和小异常流抽样需求，且所提异常流抽样算法具有比文献[9]方法更强的异常流抽样能力。

由图4可知，当抽样阈值$t>1.0\times10^4$时，选择性抽样方法和智能抽样方法的性能趋于稳定，因此设定抽样阈值t=1.0×10^4，并基于Dataset1和Dataset2两个数据集，测试4种方法抽样得到的数

据集中数据流保留情况、总体字节保留率和可疑IP地址保留情况，仿真结果如表2、表3和表4所示。

表 2 不同抽样方法数据流保留情况(条)

抽样方法	Dataset1		Dataset2	
	异常流	大流	异常流	大流
选择性抽样方法	3805	0	3826	0
智能抽样方法	195	1933	174	1294
文献[9]方法	213	747	226	565
本文异常流抽样算法	741	1063	745	1091

表 3 不同抽样方法总体字节保留率(%)

抽样方法	Dataset1	Dataset2
选择性抽样方法	10.8	21.8
智能抽样方法	89.2	78.2
文献[9]方法	43.8	43.4
本文异常流抽样算法	57.9	86.2

表 4 不同抽样方法可疑IP地址保留情况(个)

抽样方法	Dataset1		Dataset2	
	源IP	目的IP	源IP	目的IP
原始流量	215	212	206	217
选择性抽样方法	209	211	199	216
智能抽样方法	65	28	67	31
文献[9]方法	41	212	43	217
本文异常流抽样算法	215	212	206	217

由表2可知，同等条件下，选择性抽样方法异常流抽样能力最强，但由于其只关注小流抽样需求，基本上抽样不到原始数据集中的大流，导致抽样得到的数据集信息损失比较严重，不适用于构造流量异常检测数据集。其他3种方法均能在抽样异常流的同时，保留网络中的大流信息，不会造成严重的信息损失，可应用于流量异常检测数据集的构造，并且所提异常流抽样算法与文献[9]方法及智能抽样方法相比，具有更高的异常流抽样能力，平均抽样得到的异常流数量分别提高了近2.4倍和3.0倍，可以抽样更多的异常流。

由表3可知，同等条件下，本文异常流抽样算法具有与只关注大流的智能抽样方法相当的平均总体字节保留率(差值仅在10.0%左右)，且与文献[9]方法及选择性抽样方法相比，具有更高的总体字节保留率，其平均总体字节保留率分别提高了近65.3%和3.4倍，表明所提异常流抽样算法可在抽样过程中保留较多的真实网络业务流量信息，有效反映原始网络流量特征。

由表4可知，相比选择性抽样方法、智能抽样方法和文献[9]方法而言，本文异常流抽样算法能抽样得到所有可疑目的IP地址和源IP地址，抽样得到的数据集中异常流信息更加全面。

按照Dataset2构建方式，随机选取DARPA 1999和CAIDA traces 2018数据集的子集，构造10个不同的实验数据集，并基于构建的数据集测试不同方法在抽样过程中的数据流处理时间，结果如图5所示。

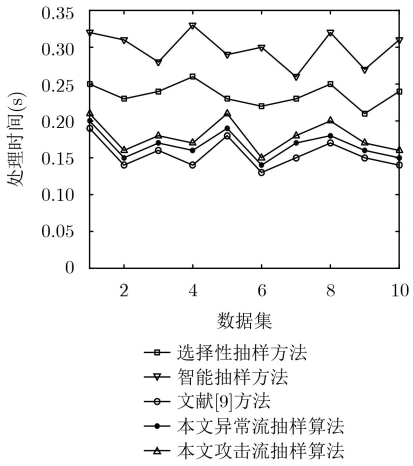

图 5 不同抽样方法数据流处理时间

由图5可知，所提异常流抽样算法处理速度优于选择性抽样和智能抽样方法。相比文献[9]方法，所提算法平均数据流处理时间增加了7.7%，但由表2可知，所提算法的异常流抽样能力比文献[9]方法提高了近2.4倍，故所提算法处理时间的增加是可以接受的。

3.2.2 攻击流抽样算法分析

将所提攻击流抽样算法与选择性抽样方法、智能抽样方法、文献[9]方法及本文异常流抽样算法进行对比，测试5种方法的非攻击异常流过滤情况。仿真结果表明，由于在抽样过程中未区分flash crowd和流量攻击，选择性抽样方法、智能抽样方法、文献[9]方法和所提异常流抽样算法无法过滤非攻击异常流，其非攻击异常流过滤量均为0。而所提攻击流抽样算法，在抽样过程中通过设定攻击流$H(\mathrm{dIP}_m)$阈值，降低了由flash crowd产生的非攻击异常流抽样概率，使得其在抽样过程中可有效过滤非攻击异常流，在Dataset1和Dataset2上分别可过滤271条和268条非攻击异常流，从而验证了所提攻击流抽样算法的有效性。

在3.2.1节构造的10个数据集上，测试攻击流抽样算法的数据流处理时间，并与其他4种方法进行对比，结果如图5所示。由图5可知，所提攻击流抽

样算法处理速度优于选择性抽样和智能抽样方法。相比文献[9]方法，所提攻击流算法平均数据流处理时间增加了15.5%，相比异常流抽样算法，其平均数据流处理时间增加了7.2%，但所提攻击流抽样算法可有效过滤非攻击异常流，且其异常流抽样能力比文献[9]方法提高了近1.2倍，故所提攻击流抽样算法处理时间的增加是可以接受的。

综上可知，所提概率流抽样方法可同时满足大流和小流抽样需求，具有较强的异常流抽样能力，能抽样得到所有与异常流量相关的源IP地址和目的IP地址，并能有效过滤非攻击异常流。

4　结束语

本文提出一种面向流量异常检测的概率流抽样方法，该方法能同时满足大流和小流抽样需求，具有较高的异常流抽样能力，并可有效过滤非攻击异常流，为下一步采用所提方法对大规模网络流量数据集进行抽样，并基于抽样数据集开展流量异常检测研究奠定了基础。

参 考 文 献

[1] YANG Chen. Anomaly network traffic detection algorithm based on information entropy measurement under the cloud computing environment[J/OL]. https://doi.org/10.1007/s10586-018-1755-5, 2018.

[2] KWON D, KIM H, KIM J, *et al*. A survey of deep learning-based network anomaly detection[J/OL]. https://doi.org/10.1007/s10586-017-1117-8, 2017.

[3] 周爱平, 程光, 郭晓军. 高速网络流量测量方法[J]. 软件学报, 2014, 25(1): 135–153. doi: 10.13328/j.cnki.jos.004445.
ZHOU Aiping, CHENG Guang, and GUO Xiaojun. High-speed network traffic measurement method[J]. *Journal of Software*, 2014, 25(1): 135–153. doi: 10.13328/j.cnki.jos.004445.

[4] ANDROULIDAKIS G, CHATZIGIANNAKIS V, and PAPAVASSILIOU S. Network anomaly detection and classification via opportunistic sampling[J]. *IEEE Network*, 2009, 23(1): 6–12. doi: 10.1109/MNET.2009.4804318.

[5] ESTAN C and VARGHESE G. New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice[J]. *ACM Transactions on Computer Systems*, 2003, 21(3): 270–313. doi: 10.1145/859716.859719.

[6] ANDROULIDAKIS G and PAPAVASSILIOU S. Improving network anomaly detection via selective flow-based sampling[J]. *IET Communications*, 2008, 2(3): 399–409. doi: 10.1049/iet-com:20070231.

[7] JADIDI Z, MUTHUKKUMARASAMY V, SITHIRASENAN E, *et al*. Intelligent sampling using an optimized neural network[J]. *Journal of Networks*, 2016, 11(1): 16–27.

[8] 伊鹏, 钱坤, 黄万伟, 等. 基于抽样流长与完全抽样阈值的异常流自适应抽样算法[J]. 电子与信息学报, 2015, 37(7): 1606–1611. doi: 10.11999/JEIT141379.
YI Peng, QIAN Kun, HUANG Wanwei, *et al*. Adaptive flow sampling algorithm based on sampled packets and force sampling threshold S towards anomaly detection[J]. *Journal of Electronics & Information Technology*, 2015, 37(7): 1606–1611. doi: 10.11999/JEIT141379.

[9] JADIDI Z, MUTHUKKUMARASAMY V, SITHIRASENAN E, *et al*. A probabilistic sampling method for efficient flow-based analysis[J]. *Journal of Communications and Networks*, 2016, 18(5): 818–825. doi: 10.1109/JCN.2016.000110.

[10] BEHAL S, KUMAR K, and SACHDEVA M. Discriminating flash events from DDoS attacks: A comprehensive review[J]. *International Journal of Network Security*, 2017, 19(5): 734–741. doi: 10.6633/IJNS.201709.19(5).11.

[11] BEHAL S and KUMAR K. Detection of DDoS attacks and flash events using novel information theory metrics[J]. *Computer Networks*, 2017, 116: 96–110. doi: 10.1016/j.comnet.2017.02.015.

[12] 张斌, 刘自豪, 董书琴, 等. 基于偏二叉树SVM多分类算法的应用层DDoS检测方法[J]. 网络与信息安全学报, 2018, 4(3): 24–34. doi: 10.11959/j.issn.2096-109x.2018020.
ZHANG Bin, LIU Zihao, DONG Shuqin, *et al*. App-DDoS detection method using partial binary tree based SVM algorithm[J]. *Chinese Journal of Network and Information Security*, 2018, 4(3): 24–34. doi: 10.11959/j.issn.2096-109x.2018020.

[13] CAIDA. The CAIDA UCSD anonymized internet traces 2013[EB/OL]. http://www.caida.org/data/passive/passive_2013_dataset.xml, 2018.

[14] CAIDA. The CAIDA UCSD anonymized internet traces 2018[EB/OL]. http://www.caida.org/data/passive/passive_2018_dataset.xml, 2018.

[15] MIT Lincoln Lab. 1999 DARPA intrusion detection evaluation dataset[EB/OL]. https://www.ll.mit.edu/r-d/datasets, 2017.

董书琴：男，1990年生，博士生，研究方向为网络安全态势感知.

张　斌：男，1969年生，教授，博士生导师，研究方向为网络空间安全.