高速网络访问超点实时检测算法精度分析

于海青，丁伟，徐杰
（东南大学网络空间安全学院，江苏 南京 211189）

摘　要：访问超点的实时获取有利于管理者更好地掌控网络。在高速网络条件下，访问超点检测的难点在于大流量给计算带来的压力。基于 40 Gbit/s 带宽的网络环境，从精度分析角度对比了流记录统计算法和基于估值原理的估算类超点检测算法。流记录统计算法采用基于重尾分布的阈值模型加以改进，估算类超点检测算法采用 SRLA 在 GPU 环境下实现，研究结果表明，估算类超点检测算法明显具有更高的检测精度和更大的改进空间，比流记录统计算法更适合被用于高速网络超点实时检测。

关键词：访问超点；网络管理；估算精度；高速网络
中图分类号：TP302
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2020117

Accuracy analysis on access hyper-point real-time detection algorithms in high-speed network

YU Haiqing, DING Wei, XU Jie
School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China

Abstract: Real-time access hyper-point helps managers better control the network. Under high-speed network conditions, the difficulty in accessing hyper-point detection lies in the pressure exerted by large traffic on computing. Based on the 40 Gbit/s bandwidth network environment, from the perspective of accuracy analysis, the flow record statistics algorithm and the estimation-based superpoint detection algorithm were compared. The flow record statistical algorithm was improved by the threshold model based on heavy-tailed distribution. The estimation-based superpoint detection algorithm was implemented in the GPU environment by SRLA. The research results show that the estimation-based hyper-point detection algorithm which has obviously higher accuracy and more room for improvement is more suitable for real-time detection of hyper-point in high-speed network than the flow record statistical algorithm.

Key words: access hyper-point, network management, estimation accuracy, high speed network

收稿日期：2019-10-07；修回日期：2020-03-18
基金项目：国家重点研发计划基金资助项目（No.2018YFB1800200）
Foundation Item: The National Key Research and Development Project (No.2018YFB1800200)

1　引言

互联网的高速发展给网络管理和安全防护带来巨大的挑战。对大规模的网络进行有效管理，保证其安全运行是一个世界性的难题。由于体系结构的局限性，病毒、木马、黑客攻击等各种恶意行为在互联网中普遍存在。它们一方面威胁网络用户自身的信息安全，另一方面也给网络整体的可用性带来影响[1]。面对复杂的网络环境，主干网层级的监测与防护是其中最重要和最基础的一环[2]。对网络里的一些核心主机给予更多关注，是提高网络管理效率的一个思路[3]。互联网中访问超点就是这样一类核心主机[4]，一般认为，访问超点指的是在一段时间内与远大于平均值数量的其他主机之间有通信的网络节点。访问超点在网络里扮演着重要角色，如服务器、代理、扫描器[5]，被 DDoS（distributed service of denial，分布式拒绝式服务）攻击的主机等。对访问超点的检测，在网络安全和网络管理领域里有着至关重要的作用[6]。

检测访问超点最直观的做法是精准统计，然而在高速网络环境下，全流量的巨大压力使得检测不能满足实时性的要求。一般有两种算法能够有效解决检测算法实时性的问题：压缩数据源，用路由器提供的流记录代替全流量进行统计；基于数学领域的估值原理的估算算法。基于流记录的统计算法以网络测量器聚合而成的流记录作为数据源，该数据源通过抽样以及聚合流量的方式大幅度地压缩原始流量，极大地减轻存储空间和计算处理上的压力。基于估值原理的估算方法是轻量级的，凭借固定有限的内存空间就能在线性时间内估算得到结果，但是如果单纯只使用估算方法，现有的大多超点检测算法在普通的硬件条件下仍然不能被应用于访问超点的实时检测，如 CSE（compact spread estimator，紧缩型连接度估值器）[7]算法和 CBF（counter Bloom filter，布隆过滤计数器）[8]算法为了达到实时性的要求，采用了以估算为主、报文抽样为辅的方法；VBF（vector Bloom filter，布隆过滤向量）算法[9]和 DCDS（double connection degree sketch，双向连接度速写）算法[10]虽然记录 IP 对的过程在线性时间内就能完成，并且以时间和空间复杂度很小的方式间接存储了潜在访问超点，但是恢复潜在访问超点的过程非常耗时，这两个算法的实时性受制于此。最新的研究表明， SRLA（sliding rough and linear algorithm，滑动线性模糊估值算法）在 GPU（graphics processing unit，图形处理器）硬件平台的支持下能够在高速网络环境中达到面向全流量的实时性要求[11]。

无论是以抽样聚合流为分析数据源的算法，还是估值算法，共同存在的问题就是检测精度的损失。流记录统计算法的检测精度依赖于抽样比和 IP 对端数的估计方法。显然，抽样比越大，检测精度越低。不少研究[12-13]都已经证实，网络流报文数具有“重尾”特性，大部分网络流的报文数量都非常少，那么将流记录统计的 IP 对端数直接作为原始的 IP 对端数或者将其除以抽样比进行补偿都没有考虑到这种分布特性，在抽样时会造成大量“短”流的丢失，导致获得的 IP 对端数与真实值产生很大的偏差，从而大大降低超点检测算法的性能。而基于估值原理的估算算法以随机理论和统计学原理为基础，会不可避免地产生估计偏差。为此，本文将基于实测数据对这两类算法进行分析，尝试在相同条件下给出精度水平，并以此说明哪类算法更适合被用于高速网络环境下的超点实时检测。

2　定义与相关工作

2.1　流记录统计算法

流是在特定的时限内具有共同特征规范的数据分组，常规的流特征规范有五元组（源地址、宿地址、源端口、宿端口、协议号）。流记录是网

络测量器将具有相同流特征规范的报文集合聚合而成的通信摘要信息[14]，一般包括源宿地址、源宿端口号、协议、ICMP 的信息类型和编码、TCP 标志位、流开始时间和结束时间、报文数、字节数、路由器转发信息等。由于处理器能力、缓存容量、网络带宽等硬件资源的限制，测量高速链路中传输的所有报文信息代价过大，现有的网络测量设备大多采用报文抽样的方法进行流量测量，然后对抽样报文采用统计推断方法进行总体统计特性的估计[15]。

流记录统计算法使用的是统计的方法，只需在测度时间窗口末端读取流记录，对各个 IP 地址分别统计相关流的数量。由于采用了报文抽样，由此得到的流数并非与 IP 地址相关流数的真实值，一般会选用一种流数补偿策略纠正以上统计的直接结果，再将纠正后的数值作为 IP 地址对端数。对端数超过指定阈值的 IP 地址被判为访问超点。

2.2 基于估值原理的估算算法

这里先介绍最常用的线性估值器[16]的原理。令 Q 表示一个包含 n 个元素的集合（其中不同元素可以是重复内容），$|Q|$表示 Q 的基数，将 Q 中元素去重以后得到的集合为 Q'，Q 的基数也即 Q' 中的元素个数。以 1 个包含 m 个 bit 位的 Bitmap 作为估算载体，起始将 Bitmap 中所有 bit 位都设为 0。逐一扫描 Q 中的元素将元素信息记录到 Bitmap。记录方式如下：对于每一个元素，取其元素 ID 作为参数通过同一个哈希函数随机映射到 Bitmap 中某一个 bit 位的位置上，将该 bit 位设为 1。在扫描完所有元素以后，统计 Bitmap 中 bit 位为 0 的数量（记为 N），根据式（1）可以估算出 Q 的基数。

$$|Q| = -m\ln\frac{N}{m} \qquad (1)$$

除此之外，还可以使用联合计数器的方法。该方法以一个包含 T 个简单计数器的计数器池作为估算载体，通过 K 个哈希函数给 Q'中的每一个元素指定分配 K 个计数器。逐一扫描 Q 中的元素将元素信息记录到计数器池。记录方式如下：对于每一个元素，将其对应的 K 个计数器累加 1。在扫描完所有元素以后，取这 K 个计数器中的最小值作为单个 IP 地址估计的基数。

现有的大多数估算类的超点检测算法都以线性估值器或联合计数器为基础完成 IP 地址对端数的估算。可根据估值载体的元素类型将这些算法划分成两类：共享计数器，如 CBF[8]算法、vHLL[17]算法等；共享估值器，如 DCDS[10]算法、VBF [9]算法、CSE[7]算法、SRLA 等。

本文重点介绍 SRLA。SRLA 解决了滑动窗口下超点实时检测中增量式更新和估算时间过长的技术难点，是首个可以在滑动时间窗口下实时检测访问超点的算法[11]。SRLA[11]大致处理流程如下。

（1）对于到达的每一个报文 IP 地址对，使用哈希函数将其存入多个线性估值器中，同时使用轻量级的模糊估值器强力度地过滤出潜在的访问超点。

（2）在每一秒结束时对记录下的潜在访问超点估算其在以当前秒为末端的时间窗口内的对端数，将其中超过指定阈值的潜在访问超点判为访问超点进行输出。

（3）更新估值载体和潜在的访问超点集合。

3 研究条件与思路

3.1 定义

首先给出与本文研究相关的术语和定义。

定义 1 （观测流量 Traffic(W)）时间窗口 T 内，从数据源获取的原始流量。

定义 2 （分析流量 IPair(W)）基于观测流量 Traffic(W)生成的 IP 地址对集合。

定义 3 （IP 地址连接对端数）IPair(W)中，所有与 iip（或 oip）构成 IP 地址对的 oip（或 iip）的集合称为 iip（或 oip）的对端地址集合，其中

iip 表示源 IP 地址，oip 表示宿 IP 地址。

定义 4　（访问超点）IPair(W)中连接对端数超过指定阈值 θ 的 IP 地址称为访问超点。

定义 5　（访问超源）IPair(W)中连接对端数超过指定阈值 θ 的 iip 称为访问超源。

定义 6　（访问超宿）IPair(W)中连接对端数超过指定阈值 θ 的 oip 称为访问超宿。

本文的研究只针对访问超源（下文简称为超点）。时间窗口 T 的长度均为 5 min，也称为时间粒度。阈值 θ 的取值为 1 024。

3.2　研究条件

CERNet（China education and research network）南京主节点 IPv4 网有 153 个接入单位，接入带宽为 40 Gbit/s。在此环境下，不仅可以获取基于 NetFlow v5 格式的流记录数据（抽样比为 1/256），还可以从以此为基础搭建的基于网络探针技术的 Netview 流量采集系统中获取全流量的报文数据。Netview 流量采集系统有两大功能：（1）采集流经采集点的所有报文的前 100 byte；（2）对流经采集点的所有匹配既定规则的全报文信息进行采集存储。访问超点检测只关注源地址和宿地址，本文使用的是功能（1）。

3.3　整体思路

本文先基于网络流报文数的分布特性提出了基于重尾分布的阈值模型，试图纠正流记录统计算法的阈值，使其能够在既定抽样比的条件下实现整体性能最优。为了保证之后比对工作的合理性，将先通过实验验证该模型的正确性。在证实这一点以后，再将该阈值模型应用到流记录统计算法中，与基于估值原理的估算算法在同等条件下进行对比分析，最终得出结论。

3.4　实验数据

在 CERNET 南京主节点网络边界同步获取流记录数据和全流量数据，从 2019 年 8 月 5 日 15 时到 2019 年 8 月 8 日 15 时，共计 72 h。实验只针对 CERNet 被管网内到被管网外方向的访问超点，所以只处理了单向全流量。以 5 min 为时间窗口（时间粒度），原始分析数据有 12×72=864 个时间窗口。

为了对两种方法的结果进行精度分析，需要基于全流量数据使用精准统计算法获取标准答案，而精准统计算法的时间复杂度较大，经多次试验一个时间窗口的标准答案产生的全过程需要 8～10 min。为了保证所有算法的实时性，每 3 个时间粒度只能有 1 个有标准答案，共计 864/3=288 个标准答案。为此，只选取了这 288 个时间粒度的超点检测结果作为精度分析的原始数据。

3.5　精度评价标准

对于访问超点检测结果，本文使用 FPR（false positive rate，误判率）、FNR（false negative rate，漏判率）、FTR（false total rate，整体错误率）来评价检测方法的精度。计算 FPR 和 FNR、FTR 的方法如式：

$$\mathrm{FPR}=S^{+}/S \tag{2}$$

$$\mathrm{FNR}=S^{-}/S \tag{3}$$

$$\mathrm{FTR}=\mathrm{FPR}+\mathrm{FNR} \tag{4}$$

其中，S 为超点的真实个数，即用精准统计方法获取的超点总数。S^{+}表示被算法误判为超点的个数，S^{-}表示被算法遗漏的个数。

4　流记录统计算法的检测精度分析

4.1　基于重尾分布的阈值模型

如前所述，流记录统计算法的检测精度受制于对端数的补偿策略。为了进一步提升流记录统计方法的检测精度，本文根据网络流报文数的分布特性提出了一个基于重尾分布的阈值模型。

基于网络流报文数量的“重尾”特性，可以近似计算出流记录统计算法因抽样而产生的对端数损失率，将其统计得到的对端数除以对端数留存率（对端数留存率=1－对端数损失率）作为最终的对端数估值，以此来补偿流记录的对端数的损失。Pareto（帕拉托）分布是具有代表性的重尾分布，可以较好地反映网络流量的随机特性，所

以本文采用 Pareto 分布进行数学建模。

定理：若流记录统计算法采用的抽样比为 s，所有网络流中单个流报文数的最大值为 u，最小值为 b，流记录统计算法测得的对端数损失率 L 为 $1-\int_b^u \frac{kb^k}{i^{k+1}}(1-s^i)\mathrm{d}i$，其中，$k$ 为 Pareto 分布的参数。以下是证明过程。

设一个网络流包含 i 个报文数，该流在抽样比为 s 的条件下被记录到流记录中的概率 Pr（i）为：

$$\Pr(i)=1-(1-s)^i \tag{5}$$

令流的报文数为随机变量 N，N~P（b, k），其概率密度函数为：

$$f(x)=\begin{cases}\dfrac{kb^k}{x^{k+1}}, x\geqslant b\\ 0, x<b\end{cases} \tag{6}$$

流记录算法对端数的损失率 L 为：

$$L=1-\sum_{i=b}^{u}P(N=i)\Pr(i) \tag{7}$$

再对离散型事件使用连续型概率分布进行估算，可得：

$$L\approx 1-\int_b^u \frac{kb^k}{i^{k+1}}(1-s^i)\mathrm{d}i \tag{8}$$

由 Pareto 分布的参数点估计方法[18]可以计算得到其中的 k：

$$k=n\times\left(\sum_{i=1}^{n}\ln\frac{X_i}{b}\right)^{-1} \tag{9}$$

其中，n 为样本的总数，X_i 表示第 i 个样本的流报文数的取值。

将该定理应用到流记录统计算法中，倘若一个 IP 地址在流记录统计结果中获得的对端数为 m^*，可以得到经过补偿纠正以后的对端数 m^{**}：

$$m^{**}=m^*/(1-L) \tag{10}$$

为了将该阈值模型应用到后面的实验中，对 2019 年 8 月 8 日实验结果中被判为访问超点的相关流的报文数进行统计。统计结果中，b=1，u=3 748 613，n=497 641，k=0.544。将统计结果代入式（8）和式（9）之中，可以得到 L 为 0.911。设超点对端数的真实阈值为 1 024（即 m^{**}=1 024），将其代入式（10），可以得到流记录统计算法纠正后的阈值为 91。

4.2 分析方案

流记录统计算法存在抽样误差，其阈值难以与真实的阈值对等起来，为此，不妨假设流记录统计算法检测出的超点数量与精准统计算法检出的数量相同，在同等数量条件下的检测结果可以直接反映出检测算法的性能。在此假设条件下流记录统计算法获取检测结果的过程如下：设在一个时间窗口内精准统计算法超点检出个数为 P，对检出结果按照对端数进行降序排序，取前 P 个超点与精准统计方法下的检出结果进行比较，以此作为一个实验组（记为流记录 1）。

为了增强实验的说明性，还另设了一个对照组，按照同样的取法获取流记录统计算法的前 1.2P 个超点结果（记为流记录 2）。为了验证第 4.1 节中阈值模型的合理性，另外将使用了纠正后阈值 91 的流记录统计算法结果作为一个对照组（其结果记为流记录 3）。将流记录 1 与流记录 2 和流记录 3 进行比对分析。此外，为了支撑以上比对过程中依据的假设条件，算法运行时需要将阈值设定得相对小一些，保障能够检出足够多的访问超点数量。根据调整测试的结果，最终将流记录统计算法运行时阈值设定为 4。

理论上说，流记录统计算法存在一个阈值区间，当检测阈值落在该区间内时，检测方法可以达到整体性能最优。为了确定性能最优的阈值区间，另设了一个实验。该实验的实施过程如下：在获取线上实验的结果以后，逐步调整阈值对同一组实验结果进行访问超点的二次过滤（从阈值下限 4 到阈值上限 1 024），分别记录在不同阈值条件下得到的访问超点结果并绘制成图像。观察

图像寻找使得算法达到性能最优的阈值区间，并通过判断补偿后阈值91是否落在该区间内来进一步验证第 4.1 节阈值模型的正确性。

4.3　实验结果分析

流记录统计算法的 FPR、FNR 比对如图 1 和图 2 所示。流记录统计算法的检测误差比较大，误报率和漏报率都在 45%以上。比对流记录 1 和流记录 2 可知，增加访问超点的检出个数，可以降低漏报率，但是也会增大误报率。从整体上看，流记录 3 的曲线和流记录 1、流记录 2 的曲线大致保持一致，可知基于第 4.1 节的阈值模型的流记录统计算法的超点检测结果在数量上与真实的超点个数基本相近，在精度上也和基于假设的检测结果大致持平，这说明了基于重尾分布的阈值模型的合理性。

流记录统计算法在不同阈值下的检测精度如图 3 所示，当检测阈值不断增大时，误报率呈现

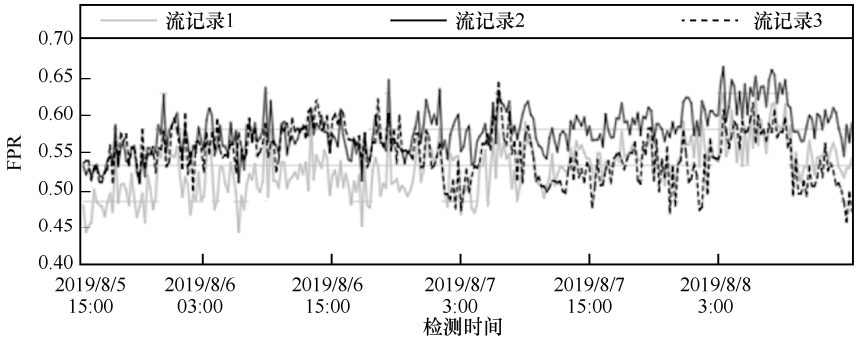

图 1　流记录统计算法的 FPR 比对

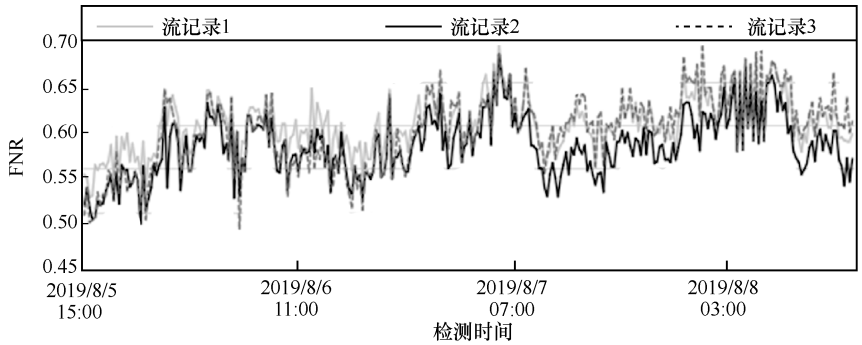

图 2　流记录统计算法的 FNR 比对

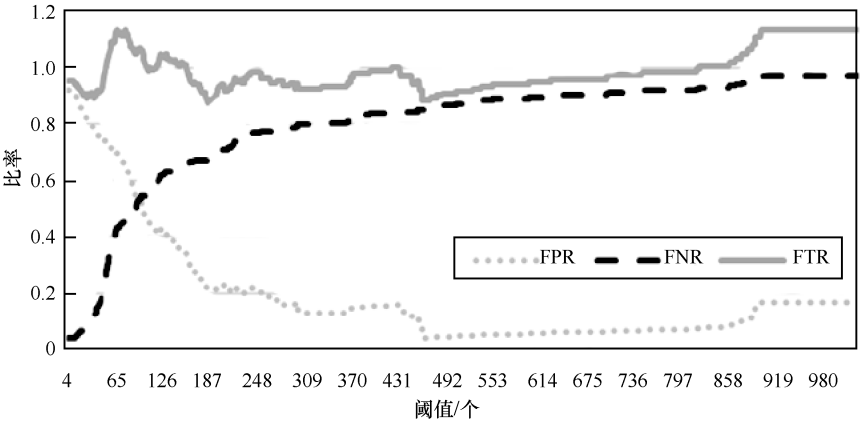

图 3　流记录统计算法在不同阈值下的检测精度

下降的趋势，漏报率呈现持续上升的趋势。显然，流记录统计算法的误报率和漏报率存在负相关的关系，而且存在一个能够使算法整体性能达到最优的阈值区间，在FPR曲线和FNR曲线交点附近，为80~120。第4.1节阈值模型纠正后的阈值91正落在其中，进一步证明阈值模型的正确性，流记录统计算法在阈值被纠正后可以基本达到整体性能最优。在既定抽样比的前提下，当流记录统计算法的检测精度达到最高水平时，生成流的报文抽样比会成为制约算法提升精度的瓶颈。显然，抽样比越小，精度会越高，但是与此同时会削减检测方法的实时性。由此可见，流记录统计算法检测精度的提升空间非常有限。

5 基于估值原理的估算算法的检测精度分析

5.1 分析方案

SRLA具有估算时间短、精度高的优势[11]，本文以该算法为研究主体展开对基于估值原理的估算算法精度的分析。将SRLA作为实验组，其检测结果记为SRLA1。第4.3节已证实，流记录统计算法在阈值被纠正后可以基本达到整体性能最优，所以可以取流记录3作为一个对照组。为了增强实验的说明性，还另设了一个对照组，对照组取法如下（同第4.2节）：设SRLA检测超点数量为$X2$，对流记录算法结果按对端数降序排序，取排序之后的前$X2$个超点结果（记为流记录4）。将SRLA1与流记录3和流记录4进行比对分析。

SRLA引入了滑动窗口机制，由此可以通过合并多个邻近时间窗口内的检测结果来获取更完整的访问超点集合。为了证实SRLA的这一优势，另设了一个只针对SRLA的比对实验。为了支持该实验，不仅需要记录下既定时间窗口末端的一组结果，还要保存前后邻近时间窗口的结果各5组，288×（1+5+5）=3 168，如此便有3 168组实验结果。分别将邻近时间窗口内的11组结果合并成一组，最终一共也是288组数据（记为SRLA2）。最终将SRLA2和SRLA1进行比对分析。

5.2 实验结果分析

FPR的结果如图4所示，可以直观地看到SRLA的误报率非常低，基本能够保证在5%以内，要明显低于流记录统计算法。SRLA是基于估值原理的算法，具有不可避免的系统偏差，所以它的这部分非常小的误差是可以被接受的。另外，估值算法还可以通过调整估值载体的参数进一步降低误差。从FNR的结果（如图5所示）可见，SRLA的漏报率虽然没有达到误报率那么低的水平，但仍要比流记录统计算法要低得多。对所有检测结果进行汇总，具体见表1，流记录统计算法检测精度的平均水平：误报率为56.4%、漏报率为59.4%；SRLA检测精度平均水平：误报率为2.3%、漏报率为14.4%。由此可见，SRLA的检测精度要明显高于流记录统计算法。

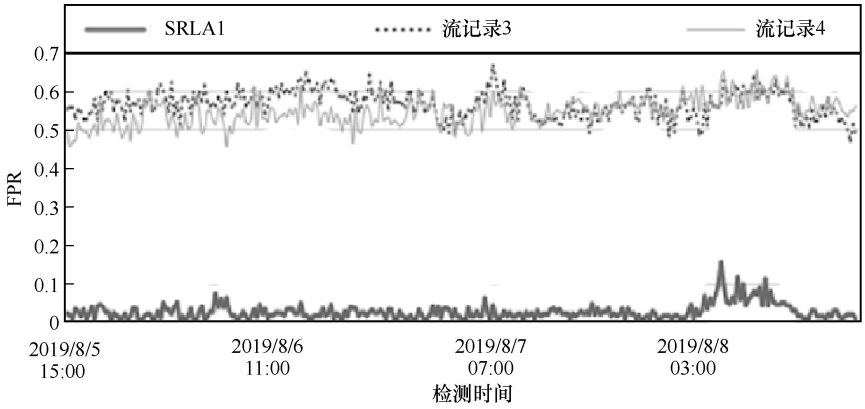

图4 流记录统计算法和SRLA算法的FPR比对

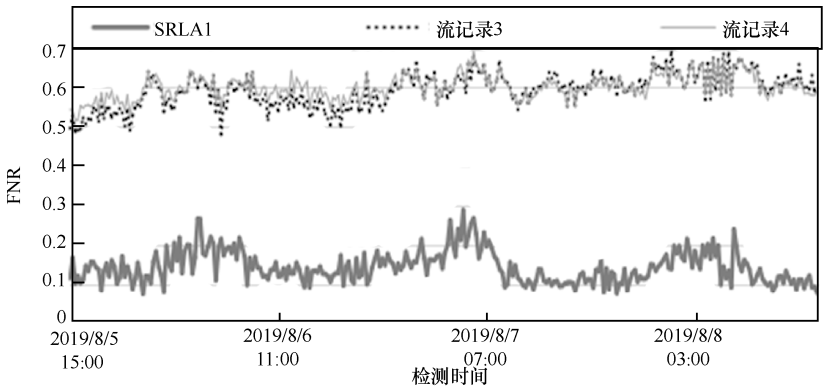

图 5　流记录统计算法和 SRLA 算法的 FNR 比对

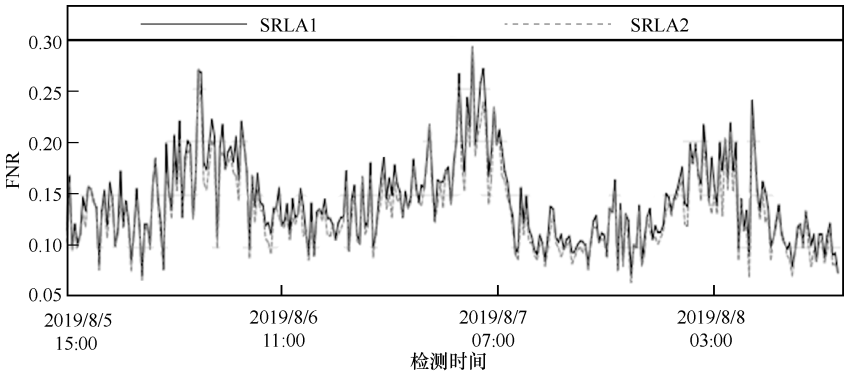

图 6　SRLA1 和 SRLA2 的 FNR 比对

表 1　流记录统计算法和 SRLA 精度统计

对比项	FPR				FNR			
	最大值	最小值	均值	方差	最大值	最小值	均值	方差
流记录统计算法	0.663	0.470	0.564	0.035	0.690	0.481	0.594	0.043
SRLA	0.155	0.000	0.023	0.021	0.289	0.074	0.144	0.040

SRLA1 和 SRLA2 的 FNR 比对如图 6 所示。SRLA2 的漏报率水平要低于 SRLA1，这也证实了 SRLA 算法可以发挥其滑动窗口机制的优势，进一步提升检测精度。由此可以推断，流记录不适合作为访问超点检测的数据源，基于估值原理的估算类算法比流记录统计算法更适合进行高速网络环境下的超点实时检测。

6　结束语

本文从精度分析角度对比了流记录统计算法和基于估值原理的估算算法。先根据网络流报文数的分布特性提出了基于重尾分布的阈值模型，并将其与估算算法进行比对，估算算法采用面向滑动窗口的 SRLA 在 GPU 环境下实现。实验结果表明，基于估值原理的估算算法的检测精度要明显高于流记录统计算法。访问超点检测是网络管理的基础，接下来工作是对检测出的超点进行研究，从流量层面深入分析超点的流量行为，检测异常的流量形态模式等。

参考文献：

[1] CHOI S, CHOI Y, LEE J, et al. Network abnormal behaviour analysis system[C]//Proceeding of 2017 19th International Conference on Advanced Communication Technology (ICACT).

[S.l.:s.n.], 2017: 49-52.

[2] 周爱平. 高速网络流量测量关键问题研究[D]. 南京: 东南大学, 2015.
ZHOU A P. Research on key issues of high-speed network traffic measurement [D]. Nanjing: Southeast University, 2015.

[3] KUCERA J, KEKELY L, PIECEK A, et al. General IDS acceleration for high-speed networks[C]//Proceeding of 2018 IEEE 36th International Conference on Computer Design (ICCD). [S.l.:s.n.], 2018: 366-373.

[4] VENKATARAMAN S, SONG D, GIBBONS P B, et al. New streaming algorithms for fast detection of superspreaders[R]. 2004.

[5] MODI C, PATEL D, BORISANIYA B, et al. A survey of intrusion detection techniques in cloud[J]. Journal of Network and Computer Applications, 2013, 36(1): 42-57.

[6] KAMIYAMA N, MORI T, KAWAHARE R. Simple and adaptive identification of superspreaders by flow sampling[C]// Proceeding of IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications. Piscataway: IEEE Press, 2007: 2481-2485.

[7] YOON M K, LI T, CHEN S, et al. Fit a compact spread estimator in small high-speed memory[J]. IEEE/ACM Transactions on Networking, 2011, 19(5): 1253-1264.

[8] CHENG G, TANG Y. Line speed accurate superspreader identification using dynamic error compensation[J]. Computer Communications, 2013, 36(13): 1460-1470.

[9] LIU W, QU W, GONG J, et al. Detection of sauperpoints using a vector bloom filter[J]. IEEE Transactions on Information Forensics and Security, 2015, 11(3): 1.

[10] WANG P, GUAN X, TAO Q, et al. A data streaming method for monitoring host connection degrees of high-speed links[J]. IEEE Transactions on Information Forensics & Security, 2011, 6(3):1086-1098.

[11] XU J, DING W, GONG Q, et al. A super point detection algorithm under sliding time windows based on rough and linear estimators[J]. IEEE Access, 2019(7): 43414-43427.

[12] 白磊, 陈超, 田立勤. 基于 TCBF_LRU 的高速网络大流检测算法[J]. 计算机研究与发展, 2014(S2): 122-128.
BAI L, CHEN C, TIAN L Q. High-speed network large flow detection algorithm based on TCBF_LRU[J]. Journal of Computer Research and Development, 2014(S2): 122-128.

[13] 陈楚, 许勇, 张凌. 重尾分布对网络流量性质的影响[J]. 计算机应用, 2009, 29(6): 1520-1522.
CHEN C, XU Y, ZHANG L. Effects of heavy-tailed distribution on the nature of network traffic[J]. Journal of Computer Applications, 2009, 29(6): 1520-1522.

[14] 丁伟, 洪沿, 夏震. 基于流记录的 HTTP 80 端口服务检测和分析[J]. 华中科技大学学报 (自然科学版), 2016, 44 (11): 34-38.
DING W, HONG Y, XIA Z. Detection and analysis of HTTP 80 port service based on stream recording[J]. Journal of Huazhong University of Science and Technology(Natural Science), 2016, 44(11): 34-38.

[15] 程光, 唐永宁. 基于近似方法的抽样报文流数估计算法[J]. 软件学报, 2013(2): 255-265.
CHENG G, TANG Y N. Sampling message flow estimation algorithm based on approximation method[J]. Journal of Software, 2013(2): 255-265.

[16] WHANG K Y, VANDERZANDEN B T, TAYLOR H M. A linear-time probabilistic counting algorithm for database applications[J]. ACM Transactions on Database Systems, 1990, 15(2): 208-229.

[17] XIAO Q, CHEN S, YOU Z, et al. Cardinality estimation for elephant flows: a compact solution based on virtual register sharing[J]. IEEE/ACM Transactions on Networking, 2017(99): 1-15.

[18] 杨永愉. Pareto 分布参数的统计推断及其应用[J]. 北京化工大学学报(自然科学版), 1992(1): 89-97.
YANG Y Y. Statistical inference of pareto distribution parameters and Its application[J]. Journal of Beijing University of Chemical Technology (Natural Science Edition), 1992(1): 89-97.

[作者简介]

于海青（1996– ），男，东南大学网络空间安全学院硕士生，主要研究方向为网络空间安全、网络测量、网络管理等。

丁伟（1962– ），女，博士，东南大学网络空间安全学院教授，主要研究方向为网络测量与行为学、网络管理及网络安全等。

徐杰（1989– ），男，东南大学网络空间安全学院博士生，主要研究方向为网络安全、并行计算和分布式计算等。