基于加密流量视角的电子资源异常访问溯源分析

韦雨君，何海涛，赵琼，姚仁龙，黎恩磊
（中山大学网络与信息中心，广东 广州 510275）

摘　要：为了溯源高校图书馆电子资源过量下载、频繁访问等违规使用事件，通过利用中山大学流量大数据分析平台，基于加密流量视角对电子资源异常访问行为进行了分析。讨论了相关网络流量指标，并提出了电子资源异常访问事件的溯源思路。研究证明加密流量特征能够有效识别电子资源异常访问行为，中山大学流量大数据分析平台能够从技术层面为确保电子资源合理使用提供支持。

关键词：高校图书馆电子资源；异常访问行为；过量下载；流量分析
中图分类号：TP393
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024258

Traceback analysis of abnormal access to electronic resources from the perspective of encrypted traffic

WEI Yujun, HE Haitao, ZHAO Qiong, YAO Renlong, LI Enlei
Network and Information Center, Sun Yat-sen University, Guangzhou 510275, China

Abstract: To trace the misuse of university library electronic resources, such as excessive downloading and frequent access, the Sun Yat-sen University traffic big data analysis platform was leveraged to analyze abnormal access behaviors from the perspective of encrypted traffic. Relevant network traffic indicators were discussed and a methodology for tracing abnormal access incidents was proposed. The findings demonstrate that the characteristics of encrypted traffic can effectively identify abnormal access behaviors, and the Sun Yat-sen University traffic big data analysis platform provides technical support to ensure the proper use of electronic resources.

Keywords: university library electronic resources, abnormal access behavior, excessive downloading, traffic analysis

0　引言

随着数字资源的发展，学术资源的在线化和数字化大大便利了高校师生获取相关资源并进行科学研究。各大高校为了满足师生的科研需求，每年都会投入大量经费用于购买各类数据库和电子资源。然而，随着电子资源在高校的使用率越来越高，异常访问、过量下载等事件也愈发频繁发生。部分师生由于版权意识淡薄，可能会采用不当方式使用电子资源，主要包括：1) 使用下载工具批量下载或多线程下载图书馆购买的电子资源； 2) 进行连续、系统、集中、批量的下载、浏览、检索数据库等操作[1]。

电子资源违规使用行为一旦被数据库商成功监测，极有可能会造成学校的整个IP段被对方封禁，严重影响其他用户正常使用电子资源。此外，电子资源异常访问事件的频发会导致高校图书馆无法掌握数据库的实际使用情况，虚高的使用量不但会干扰图书馆决策是否要续订该数据库，还会增加数据库商在下一期合同签订的议价筹码[2]，更会对学校的声誉造成负面影响。因此，如何追溯数据库商反馈的过量下载行为，对电子资源异常访问行为进行监测，已成为各大高校亟待解决的问题。

目前，已有部分高校通过采取技术手段对图书

收稿日期：2024–10–22

馆电子资源异常访问行为进行监控。清华大学通过在校园网出口网关部署电子资源访问控制系统，对用户访问数据资源的情况进行实时监控，根据网络流量和下载频率设置阈值对违规行为进行判定[3]。西安交通大学通过电子资源流量控制系统在校园网出口旁路所有电子资源访问的http流量，针对不同数据库设定阈值，对异常IP实施阻断[4]。大连理工大学同样使用旁路监听技术实时分析校园内电子资源访问流量，监控过量下载行为[5]。北京工业大学借助EZproxy代理软件，对师生在校外远程访问图书馆资源的行为进行分析，发掘读者访问电子资源时的异常行为[6]。

然而，目前的电子资源使用行为的流量监控方法无论是正向代理还是旁路监控均基于http明文设计，直接根据用户访问电子资源的内容判断下载行为并对过量下载进行监测。随着加密技术的不断发展，大部分网络流量均采用了加密技术对传输内容进行了加密[7]，基于明文监测的电子资源异常访问行为监测方法不再适用于当前情况。因此，本文的工作重点是探索在加密网络流量的背景下电子资源异常访问的流量特征。

中山大学网络与信息中心于2018年提出流量大数据安全分析平台[8]。该平台通过对校园网边界流量进行数据采集和解析提取，形成流量日志数据。通过对日志数据的挖掘，可以进一步对校园网络的威胁情报进行分析。本文通过利用中山大学流量大数据安全分析平台，从加密流量视角对电子资源异常访问行为进行分析，提供了异常事件回溯思路，并以数据库A异常访问事件为例，探索了传输控制协议（TCP）流量特征、网络地址转换（NAT）数据与异常访问行为的相关性。

1　流量数据

中山大学校园网络流量大数据分析平台主要包括流量采集与分发平台、流量预处理模块、Cloudera大数据管理平台，共3个部分[8]，如图1所示。

流量采集与分发平台在校园网核心路由器进行端口镜像实现流量捕获。通过软件定义网络（SDN, software-defined networking）控制器下发OpenFlow流表，由OpenFlow交换机对流量进行采集、过滤和分发。

流量预处理模块使用数据帧处理板卡对网络流量进行识别解析，将数据分发到现场可编程门阵列（FPGA, field-programmable gate array）处理器进行处理。FPGA通过硬件逻辑快速执行预定义的处理任务，如流量过滤、分类、添加时间戳等，并且根据用户定义规则将多组数据帧合并，聚合成单一流数据。对于TCP和用户数据报协议（UDP）流量，流量预处理模块可以根据(源IP, 源端口, 协议, 目标IP, 目的端口)组成的五元组，定位出独立的流数据，然后对应用层协议进行识别，将聚合后的流数据存储到相应的数据结构中。导

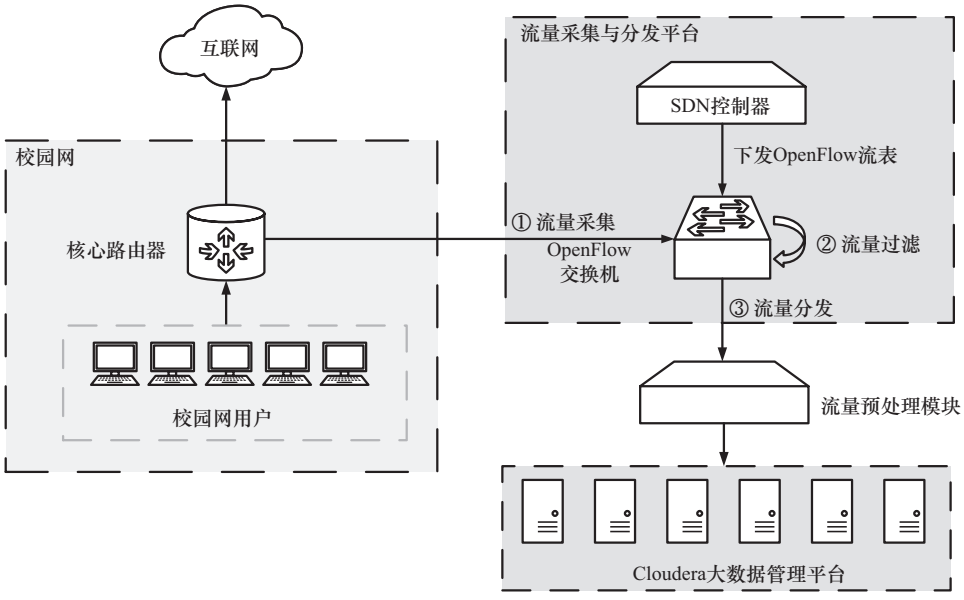


图1　中山大学校园网络流量大数据分析平台

出的流量日志存储在Cloudera大数据分析平台的Hadoop分布式文件系统（HDFS, Hadoop distributed file system）上，通过impala建立元数据，后续可以通过结构查询语言（SQL）语句进行查询分析。

针对高校图书馆电子资源异常访问任务，本文主要关注通过大数据分析平台收集的流量数据，包括：网络地址转换日志（network.nat）、域名解析流量（network.dns）、TCP流量（network.tcpflow）。下面对相关的流量数据进行详细介绍。

1.1 network.nat

network.nat中存储了NAT设备日志，其中记录了校园内部设备到外部网络之间的连接信息。相关参数包括时间戳、目的IP、转换前内网IP，转换后公有IP等数据，部分NAT参数如表1所示。由于校内IP地址需要经过网络地址转换后才能访问校外资源。因此，当电子资源异常访问事件发生时，数据库商侧仅能看到学校发起异常访问的出口IP。在事件回溯过程中，NAT日志有助于校内人员对可疑源IP进行排查。

表1 部分NAT参数

字段名	类型	字段说明
time	string	时间戳
dstip	string	目标IP地址
dstport	string	目标端口号
beforeTransAddr	string	转换前地址
beforeTransPort	string	转换前端口
afterTransAddr	string	转换后地址
afterTransPort	string	转换后端口

1.2 network.dns

流量处理平台通过对UDP流量进行识别，然后过滤出应用层协议为DNS类型的流量，生成DNS日志。DNS日志中记录了请求时间、请求的客户端IP地址、DNS服务器地址、解析类型、请求域名、返回的IP地址以及txt记录等内容，部分DNS参数如表2所示。网络流量大数据分析平台收集到的DNS流量数据保存在表network.dns中，每条DNS日志详细记录了校园网流量中用户对域名解析的查询及收到响应的情况。

表2 部分DNS参数

字段名	类型	字段说明
time	string	时间戳
qlient	string	客户端IP地址
qserver	string	DNS服务器地址
type	string	解析类型
qname	string	请求域名
ip_txtcnt	string	域名解析的IP地址、txt记录内容

1.3 network.tcpflow

TCP流数据是通过将同一TCP连接中的TCP包进行重组而生成的。流量处理平台根据TCP包的五元组(源IP, 源端口, 协议, 目标IP, 目的端口)，判断其是否属于同一TCP流。以SYN包作为流的开始，将FIN/Reset包作为流的结束。然后通过计算连接持续时长，发送数据包数、接收数据包数、发送字节数、接收字节数等，对TCP包数据进行聚合，最终生成一条TCP流数据，部分tcpflow参数如表3所示。

表3 部分tcpflow参数

字段名	类型	字段说明
time	string	时间戳
elapsedtime	bigint	持续时长
protocol	string	协议
srcip	string	源IP地址
srcport	string	源端口号
dstip	string	目标IP地址
dstport	string	目标端口号
txpackets	bigint	发送包数
rxpackets	bigint	接收包数
txbytes	bigint	发送字节数
rxbytes	bigint	接收字节数

2 电子资源异常访问行为溯源思路

在电子资源异常访问事件溯源的过程中，数据库商通常会反馈异常访问行为发生的时间、请求流量日志、代表性的问题IP地址等信息。本节对电子资源异常访问事件的网络流量回溯思路进行了详细阐述，说明了如何利用我校流量大数据分析平台收集到的流量数据，验证异常访问行为并对数据库异常访问源头进行确定。利用流量数据对电子资源异常访问事件进行溯源的思路如图2所示。

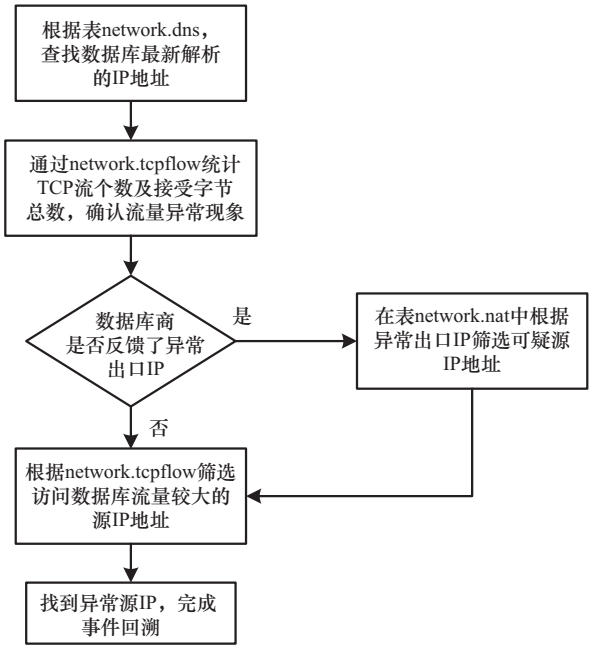


图2　利用流量数据回溯电子异常访问行为

首先，确定目标数据库IP地址。由于数据库的IP地址可能会变动，而数据库的URL基本上是固定的并且供应商会及时反馈。因此，进行数据库异常访问行为回溯时，首先需要确定目标数据库的IP地址。根据network.dns中的历史DNS解析记录，获取数据库URL对应的最新解析IP地址。

然后，确定异常流量特征。通过查看tcpflow数据，分析数据库供应商反馈的日期是否存在流量异常。可以固定目的IP对当天TCP流个数及接受字节总数进行统计，并与近两个月内的数据库访问网络流量情况进行对比，判断是否存在流量异常。

最后，定位发起异常行为的校内源IP地址。若数据库商提供了发起异常行为的IP地址（某个学校出口IP地址），则可以根据network.nat通过定位时间、目标数据库IP、转换后IP，对可疑源IP地址进行筛选。若对方未反馈异常出口IP地址，则根据network.nat筛选访问数据库次数最频繁的源IP，并且利用network.tcpflow筛选与数据库频繁建立TCP连接且有较大通信流量的源IP地址。

3　电子资源异常访问行为溯源分析案例

下面以数据库A异常访问事件为例进行溯源分析。我校于2024年6月收到数据库商反馈数据库A的访问用量存在异常，如图3和图4所示。由图3可知，自2024年5月末开始，我校对数据库A的访

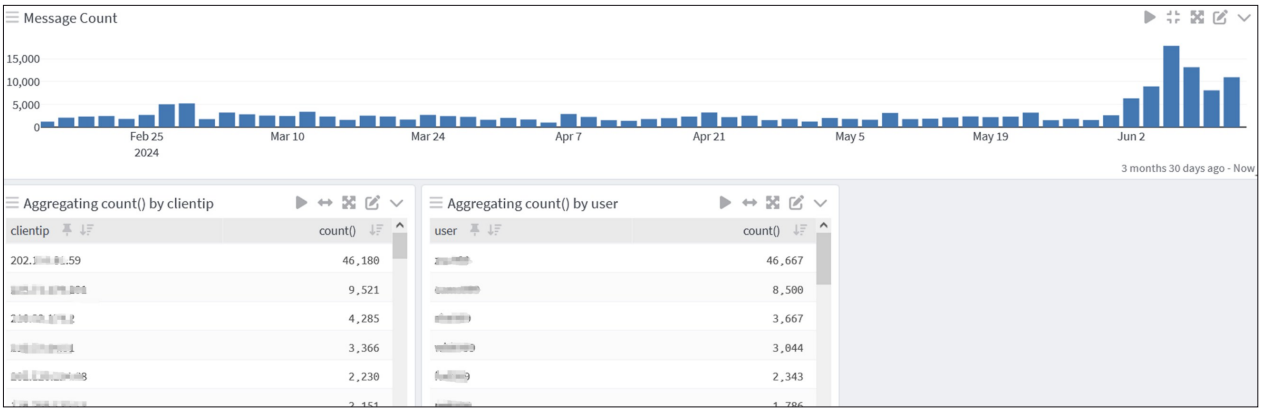


图3　我校IP于2月到6月访问数据库A的访问量变化

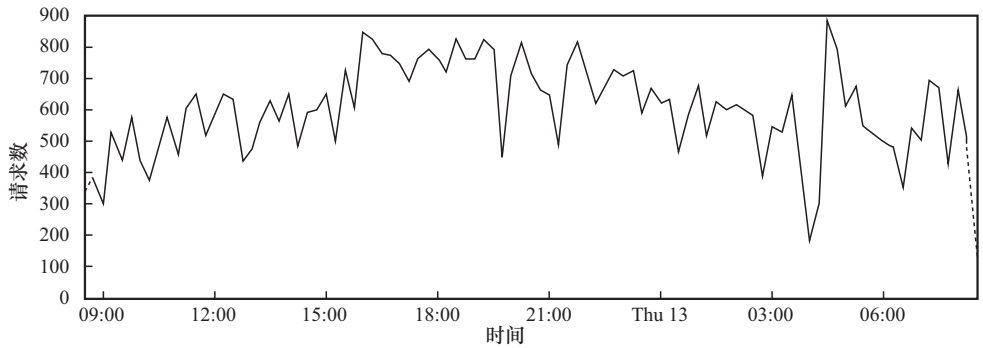


图4　数据库A自6.12 8:35到6.13 8:35收到的网络请求数

问量急剧升高，6月份的访问量与之前的访问量相比有大幅提高，且代表性IP为202.x.x.59。此外，由图4可知，自6月12日8点35分到6月13日8点35分（UTC时间），我校IP地址对数据库A存在不间断访问，频繁发送消息。

基于以上背景，我校网络与信息中心利用流量大数据分析平台对数据库A的异常访问事件进行了回溯。首先，基于数据库A的URL设定qname字段，于network.dns表中查询数据库A最新的解析IP地址为172.x.x.2和104.x.x.254，如图5所示。

然后，对网络流量异常情况进行验证。利用network.tcpflow和network.nat，本文对4月到6月校内IP地址访问172.x.x.2和104.x.x.254的TCP流接受字节总数和地址转换次数按天进行了统计。如图6和图7所示，与4月和5月相比，6月TCP流接受字节总数和NAT次数有明显的增长，仅有部分天数流量较小，印证了数据库A的访问量在6月份急剧增加，确实存在访问流量异常情况。

接着利用network.nat对发起异常行为的校内源IP地址进行定位。将时间限定为6月12日8点35分到6月13日8点35分，转换后IP地址设为202.x.x.59，目的IP地址设为172.x.x.2和104.x.x.254，对NAT流量进行筛选，并根据NAT转换次数降序排序，排查可疑源IP地址。如图8所示，x.x.12.248的NAT次数为14533，明显高于其他IP地址，基本锁定x.x.12.248为可疑源IP。

最后，利用network.tcpflow筛选与数据库频繁建立TCP连接且有较大通信流量的源IP地址。通过统计每个源IP地址访问目标数据库IP所建立的TCP连接数量和接收字节总数，查找流量较为异常的源IP。如图9所示，与x.x.12.248相关的TCP流数量（tcpcounts）和接收字节总数（sumrx）明显高于其他源IP，进一步锁定了x.x.12.248为异常访问数据库A的源IP。

为了进一步验证我校流量大数据分析平台采集到的流量数据和数据库商侧的监控流量匹配程度。本文通过统计校内IP地址访问目标数据库的地址转换次数，对图4数据库商反馈的网络请求趋势进行了模拟。由于数据库商仅反馈了6月12日8点35分到6月13日8点35分我校IP请求数据库的网络请求数量趋势图，未反馈原始数据和计算方式。因此，根据图3的采样频率，每15 min计算一次NAT次数的累计值，图10为相应的NAT次数变化趋势图。将图10与图4进行对比，如图11所示，实线

	qname	ip_txtcnt
1	[redacted].com 数据库A的URL	,,,
2	[redacted].com 数据库A的URL	172.[redacted].2,104.[redacted].254,,
3	[redacted].com 数据库A的URL	104.[redacted].254,172.[redacted].2,,

图5　获取数据库A的IP地址

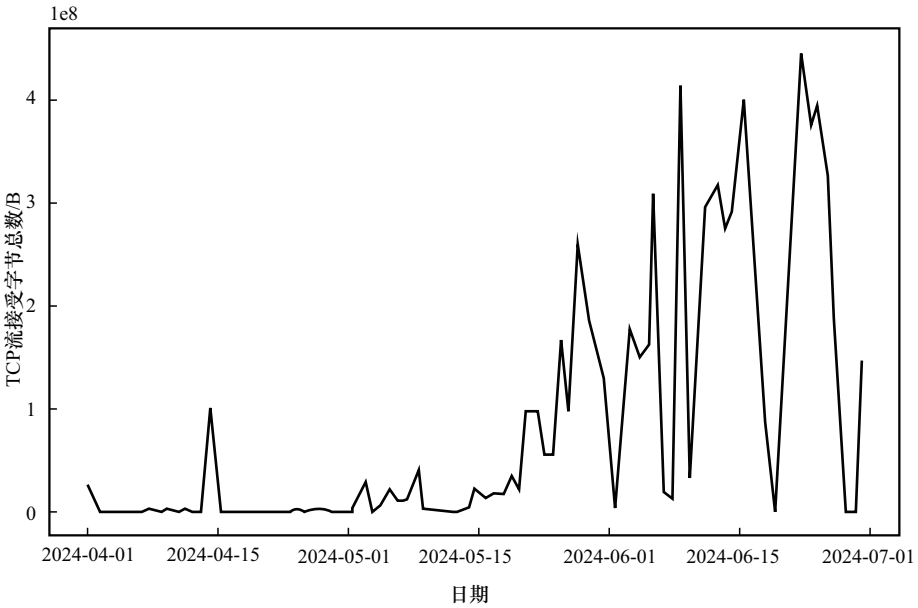


图6　访问数据库A的TCP流接受字节总数（4月到6月）

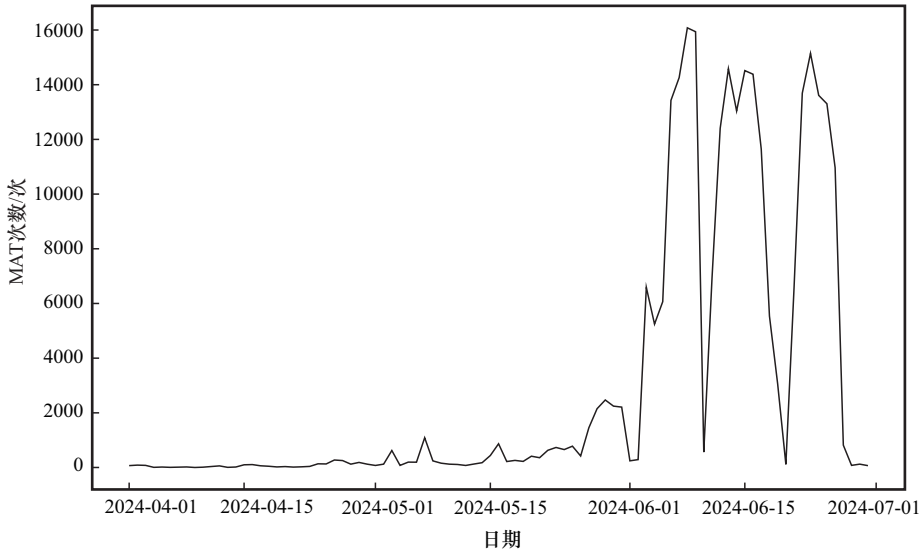


图7　访问数据库A的NAT次数(4月到6月)

	srcip	natcounts
1	12.248	14533
2		31
3		1

图8　通过NAT流量数据筛选可疑源IP

	srcip	tcpcounts	sumrx
1	12.248	13294	305952572
2		11	69341
3		10	552359
4		8	13641
5		7	30061
6		7	184910
7		7	19678
8		7	4437150
9		4	13308
10		4	7778

图9　通过TCP流量数据筛选可疑源IP

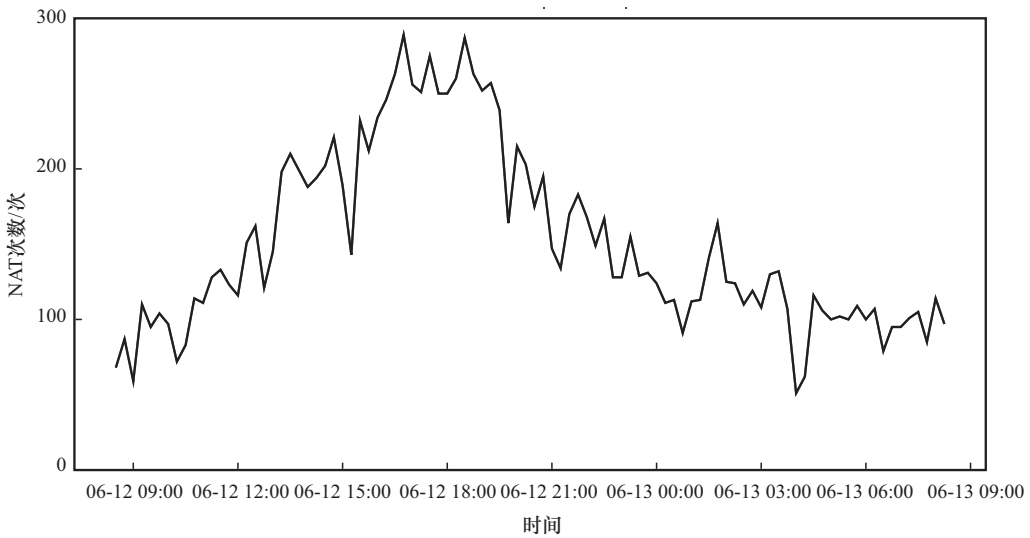


图10　访问数据库A的NAT次数(6.12 8:35－6.13 8:35)

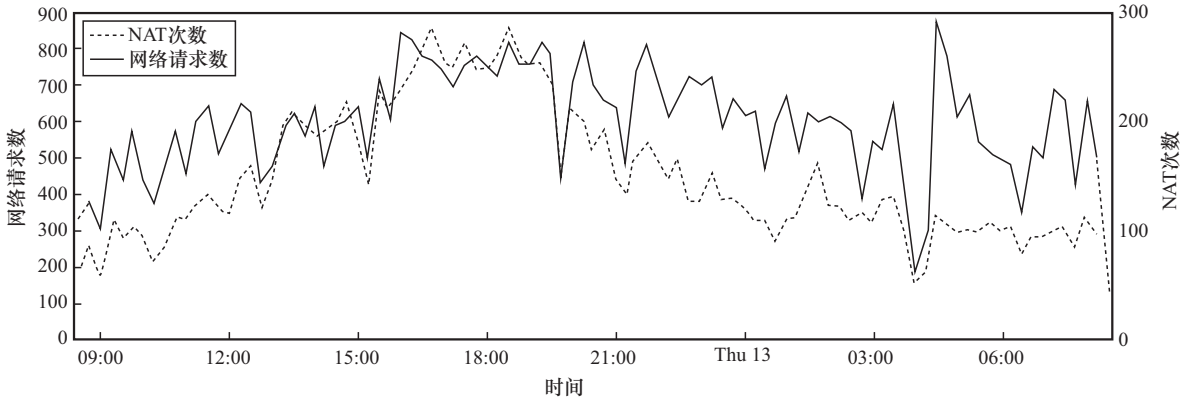


图 11　采用 NAT 流量数据模拟数据库商侧监控流量

为数据库商反馈的网络请求数量，虚线为通过流量平台采集的 NAT 次数。由于 NAT 次数和请求次数并不完全对应，并且发出请求和服务器接收到请求可能存在一定时延。因此，两者不能完全匹配。然而，仍然可以看出 NAT 次数能够较好地模拟整体请求数量的变化趋势，部分关键节点的上升下降趋势能够有效匹配。这说明本文的流量数据能够有效模拟数据库商侧的监控流量。

4　结束语

本文利用中山大学网络流量大数据分析平台，基于加密流量视角对高校图书馆电子资源异常访问行为进行了分析。首先对流量大数据分析平台能够采集到的相关流量数据进行了详细介绍，然后提供了对电子资源异常访问事件进行溯源的问题排查思路。以数据库 A 异常访问事件为例，通过对 TCP 流量特征、NAT 数据进行统计分析，验证了异常流量情况，并且定位到了校内的可疑源 IP 地址。事件的溯源过程同时验证了我校流量大数据分析平台监控到的网络流量能够和数据库商侧的监控流量有效匹配。本文的工作主要是基于离线的流量数据对电子资源异常访问进行检测，证明了我校流量大数据分析平台在该场景下的技术支撑作用。未来的工作将进一步探索如何基于我校流量大数据分析平台，利用加密流量数据对电子资源异常访问行为进行实时监测，以更有效地保障电子资源的合理利用。

参考文献：

[1] 中山大学图书馆电子资源版权公告[EB/OL]. (2018-07-05)[2024-08-04].

[2] 肖敖夏. 高校图书馆电子资源过量下载的成因及对策[J]. 图书情报导刊, 2018, 3(2): 29-32.
XIAO A X. The causes and countermeasures for excessive downloading of electronic resources in university library[J]. Journal of Library and Information Science, 2018, 3(2): 29-32.

[3] 邹荣, 张成昱, 姜爱蓉, 等. 电子资源访问管理与控制系统的设计及应用[J]. 图书情报工作, 2010, 54(1): 121-124.
ZOU R, ZHANG C Y, JIANG A R, et al. Design and application of electronic resources access management and control system[J]. Library and Information Service, 2010, 54(1): 121-124.

[4] 邵晶, 阎晓弟, 周琴, 等. 电子资源流量控制需求分析及其解决方案[J]. 大学图书馆学报, 2012, 30(4): 11-13.
SHAO J, YAN X D, ZHOU Q, et al. The demands analysis and the solution for the electronic resources usage control[J]. Journal of Academic Libraries, 2012, 30(4): 11-13.

[5] 王政军, 俞小怡, 金玉玲. 利用旁路监听技术约束数字资源过量下载[J]. 现代图书情报技术, 2015(12): 95-100.
WANG Z J, YU X Y, JIN Y L. Using sniffer technology to constraint electronic resource excessive downloading[J]. New Technology of Library and Information Service, 2015(12): 95-100.

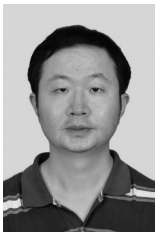
[6] 雷东升, 郭振英. 基于 EZproxy 日志的电子资源异常访问行为研究[J]. 现代情报, 2016, 36(7): 101-106.
LEI D S, GUO Z Y. Research on abnormal access to electronic resources based on EZproxy logs[J]. Journal of Modern Information, 2016, 36(7): 101-106.

[7] C. Fu, Q. Li, and K. Xu. Detecting Unknown Encrypted Malicious Traffic in Real Time via Flow Interaction Graph Analysis[C]. In Proc. NDSS, 2023.
FU C P, LI Q, XU K. Detecting unknown encrypted malicious traffic in real time *via* flow interaction graph analysis[C]//Proceedings 2023 Network and Distributed System Security Symposium. Reston, VA: Internet Society, 2023: 1-18.

[8] 杨敏, 何海涛, 赵琼. 流量大数据安全分析平台的设计与实现[J]. 通信学报, 2018, 39(S1): 104-109.
YANG M, HE H T, ZHAO Q. Design and implementation of traffic big data security analysis platform[J]. Journal on Communications, 2018, 39(S1): 104-109.

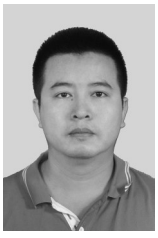
[作者简介]

韦雨君（1997–），女，贵州贵阳人，中山大学助理工程师，主要研究方向为网络安全、流量分析。



何海涛（1975–），男，安徽淮北人，博士，中山大学高级工程师，主要研究方向为因特网流量行为、大数据等。

姚仁龙（1992–），男，安徽安庆人，中山大学助理工程师，主要研究方向为人工智能、流量分析。



赵琼（1983–），男，湖北黄冈人，中山大学工程师，主要研究方向为计算机网络技术。

黎恩磊（1995–），男，河南漯河人，中山大学助理工程师，主要研究方向为数字取证、流量分析。