教育教学与管理 2026年第9期
ISSN: 3079-4439

网络流量大数据的实验室人物画像与精准育人

胡春兵[1]，王旭英[2,*]
[1]四川民族学院智能科学与技术学院，四川康定，中国
[2]四川民族学院图书与信息中心，四川康定，中国
*通讯作者

【摘要】管理粗放、安全教育同质化与数据支撑不足，是高校实验室安全管理长期面对的三类问题。本文据此提出并实现一种基于网络流量大数据的实验室人物画像模型，并将其作为独立实践项目纳入实验室安全育人体系。研究以某高校实验室网络出口防火墙导出的 11，735，807 条会话日志（覆盖 1，176 个内网 IP 实体）与旁路抓包数据为基础，经编码规整、异常清洗与设备流量识别后，构建“身份—行为—风险—教育”四维画像指标体系，并采用 K-means 聚类、孤立森林异常检测与熵权-专家混合加权风险评分方法，实现用户分群、异常识别与风险量化。结果表明：1，125 个真实用户实体可划分为办公型（595 个，52.9%）、普通型（269 个，23.9%）与高风险型（261 个，23.2%）三类；识别出 4 个异常实体，均属高风险型；17 个实体风险评分超过 80 分；10.0%的实体存在风险域名或风险关键词命中。该模型可为实验室安全风险预警、精准安全教育与管理决策提供量化依据，并可成为提升学生数据素养与安全素养的实践育人载体。

【关键词】网络流量大数据；人物画像；实验室安全；用户分群；风险评分；实践育人
【基金项目】基于人工智能视角下的高校实验室安全管理运行机制策略研究——以理工省级实验教学中心（四川民族学院）为例（编号：XJYB202431）

1.引言

实验室是高校人才培养与科学研究的基础场所，其安全管理直接关系师生人身安全与校园稳定。教育部先后发布《高等学校实验室安全管理办法》等文件[1]并部署实验室安全专项检查，对实验室安全管理提出“全覆盖、零容忍”的要求。不过，日常管理仍以制度约束和事后处置为主，短板集中在这三处：其一，风险识别依赖人工巡查与经验判断，缺乏量化数据支撑，难以实现早发现、早预警；其二，安全教育内容同质化，未针对不同风险群体差异化设计，教育效果缺乏可测度性；其三，数据资产闲置——实验室网络出口的防火墙与上网行为管理设备持续沉淀海量会话日志，记录着用户的访问行为、应用偏好与网络质量等信息，但其管理价值尚未被系统开发。

大数据与人工智能技术为精细化安全管理提供了可行路径，然而海量网络数据的汇聚与使用也使数据安全与隐私保护面临新挑战[2-4]。用户画像将海量行为数据转化为可操作的特征标签，已在电子商务、个性化推荐与图书馆服务等领域得到广泛应用[5,6]；将其引入高校实验室安全管理与安全教育的研究仍相对较少。网络流量识别技术能够从海量会话日志中自动提取应用与行为特征[7]，为实验室用户行为建模提供了技术基础。

本研究基于防火墙会话日志与旁路抓包数据，构建实验室网络使用者的“身份—行为—风险—教育”四维人物画像模型，实现安全风险的可视化、量化评估与预警，并将建模全流程设计为不依托特定课程的独立实践项目。主要贡献包括：

（1）本研究通过提出建立从“主机”到“设备”再到“人物”的三级画像实体模型，搭建基于“IP-MAC”映射与身份的融合接口，主要目的是解决会话日志缺乏用户账号关联的问题；

（2）本研究通过构建覆盖时间（Time）、流量（Data traffic）、应用（Application）、空间（Space）、传输质量（Transmission Quality）与风险（Risk）六类共 28 维画像（Dimensional Portrait）特征（Feature），即“T-D-A-S-TQ-R-RP & F”模型，并辅以 K-means 聚类、孤立森林异常检测与混合加权风险评分等算法实现用户分群、异常识别与风险量化；

（3）本研究将模型设计（Model Design）

为数据驱动（Data Drive）的实践育人载体，完成对学生上网行为的数据采集、特征工程到模型构建与安全分析的全流程训练，初步形成安全管理与育人成效的协同提升机制。

本研究的实验数据来自某高校实验室网络信息中心提供的网络出口设备导出的信息数据（详见第 2 节），该数据经过预处理后，通过对 1125 个真实用户实体建模，构建起完整的数据模型和训练过程。训练结果表明：3 类群体差异显著，4 个异常实体均集中于高风险型群体，17 个实体风险评分超过 80 分，结合训练结果数据，可构成针对这些风险群里的预警与精准教育的重要目标，为下一步实验室安全教育指明方向。

本文后续安排为：第 2 节数据基础与预处理；第 3 节人物画像模型设计；第 4 节实验室安全应用场景；第 5 节实践实施与成效；第 6 节结论与展望。

2.数据基础与预处理

2.1 数据来源

本研究所使用的数据来源于某高校网络信息中心提供的网络出口设备导出的实验室网络出口防火墙与旁路抓包设备，经过分析，得到两类数据。

（1）防火墙会话日志：由上网行为管理设备导出，一共两份。时间覆盖 2026 年 5 月 21 日至 31 日，累计 11735807 条会话记录，涉及 1176 个校内内网 IP 实体，总下行流量约 580.7GB。

（2）旁路抓包数据：导出数据采用经典“.pcap”格式，共 553387 个数据包、数据体量约 128MB，时间范围为 2026 年 6 月 9 日 09：05：22.9 至 09：05：49.9（约 27s），此数据主要为深度解析和呈现协议细节与 TCP 传输异常。

结合两类数据的规模、周期、数据协议包等特征进行研判，本研究认为：防火墙会话日志数据适合用于刻画用户行为规律，而旁路抓包数据该类数据可用于画像模型的验证与补充，即：“会话-行为”模型和“旁路-补充”模型。

2.2 数据字典

防火墙会话日志共 22 个字段，其中关键字段说明如表 1 所示。

表 1.会话日志关键字段说明（节选）

Field	Type	Introduction
Source/Target Address	text	Both of The IP Address for Conversation
Source/Target Port	number	Communication Port Number
Access Time	text	Time About Communication Start
Protocol	text	Trans Layer Protocols And Business Application Identification Result (Like Douyin, Remote Control Application, etc.)
Upload/Download Data Traffic	number	Bytes About Sent and Received
Upload/Download Data Packets	number	Number of UpStream and Downstream Data Packets
Upload/Download Retransmission	number	Number of TCP Retransmission
Upload/Download Reset	number	Number of TCP RST Packets
Domain	text	Access Domain
Target Location	text	Target Location(Like Hangzhou Zhejiang Alibaba Cloud)
User Account	text	Sign in to Verify Account（null）

2.3 数据预处理

第一步须完成编码与格式规整：原始日志为 GB18030 编码，将不符合计算标准的数据统一转换为 UTF-8；时间字段（如“2026-05-22/01：17：49”）标准化为可计算格式。

第二步须将地址与协议规范化：完成对 IPv4/IPv6 地址的表示规范；通过计算分析，过滤 ARP、组播等非 IP 帧（抓包数据中 ARP 帧 65，996 个，约占 12%），避免干扰。

第三步须将异常数据清洗：原始数据抓包文件在采集过程中（抓包工具导出）被截断（文件大小恰为 128MB），必须对尾部零时间戳数据包进行剔除；同时对会话级字段进行缺失值与异常值计算分析，完成初试数据的统计清洗。

第四步须对设备流量识别与剔除。通过对原始数据进行统计发现，51 个 IP 实体的会话数均超过 5 万条、应用种类仅 1 种、目标 IP 不超过 9 个、端口不超过 7 个，重传率中位数约 34%，行为模式高度一致，判定为网络摄像头、终端代理等设备的遥测流量数据。此类流量约占会话总量的 75%，不属于人物行为，属于干扰数据。若直接参与建模将严重影响训练结果影响建模和进行最终数据画

像，须予以剔除，最终保留 1125 个真实用户实体进入建模。

2.4 数据脱敏与合规

原始数据中获取的网络流量日志涉及个人信息，在研究中须遵循最小必要与去标识化原则：（1）本研究中所有 IP 地址均以掩码形式展示；（2）本研究所使用的数据仅用于实验室安全管理与教学实践，在使用上已获学校网络信息管理部门授权；（3）数据处理符合个人信息保护相关法律法规及《信息安全技术个人信息安全规范》（GB/T35273—2020）的要求[8]。

3.人物画像模型设计

3.1 画像实体与总体架构

本研究中的画像实体是画像建模的基本对象，考虑到会话日志中的“用户账户”（即表 1 中“User Account”）字段为空、源 MAC 为网关接口，在实现过程中难以直接关联到具体人员。因此，本研究提出建立从“主机”到“设备”再到“人物”的三级画像实体模型：以主机（IP）为最小画像单元，通过 IP 地址与 MAC 地址映射，形成终端设备画像，并预留身份融合接口和终端设备上网认证记录（需更高授权获得），实现主机画像向人物画像的聚合。

本研究建立的总体架构自下而上分为别数据层、特征层、模型层、画像层与应用层（图 1）。数据层通过原始数据解析并融合会话日志、旁路抓包与身份映射接口等三类数据源；特征层完成对多维数据统计分析，建立特征标记，形成完整特征层；模型层使用现有成熟算法，基于特征标记，用于实现用户分群和分类、异常数据检测与风险数据分级；画像层根据数据输出标签体系，进行用户上网行为风险评分；应用层对接安全态势可视化、风险预警，为管理部门提供精准安全教育数据与决策信息支持。

3.2 画像指标体系

本研究所指的用户画像（Persona）概念源于交互设计中对目标用户的抽象与建模[9]。探索搭建一套成熟的用户上网行为画像理论并立足实验室安全管理需求，构建“身份—行为—风险—教育”实验室用户四维画像指标体系（图 2）。

（1）身份画像（Who）：被测用户账号、部门/角色、设备类型、归属实验室、认证方式。由于当前日志缺少身份字段，故此维度需通过身份融合接口授权后接入，确定用户真实信息。

（2）行为画像（What/When/Where）：统计并分析用户上网的时间规律（活跃时段、活跃天数、深夜活跃度）、应用偏好（应用协议分类、娱乐/办公/学习类占比）、流量特征（总量、上下行占比、平均会话字节数、大流量会话数）、空间特征（目标 IP 与地理位置分布、跨境占比），汇总用户上网行为特征。

（3）风险画像（Risk）：根据数据传输质量（重传率、RST 率）、高危访问行为（风险域名、高风险 TLD 命中）、违规外联动作流量（远程控制、游戏、跨境流量）、异常下载行为（大流量会话）、异常探测扫描（多目标多端口），建立风险清单，汇总风险行为，进行风险行为的深度画像。

（4）教育画像（Edu）：通过画像，提高管理部门对用户上网行为安全事件参与度，同时可以提升安全预警情况响应率，用作安全管理培训与考核记录，支撑精准安全教育。

数据字段—画像特征—画像标签的映射关系如表 2 所示。

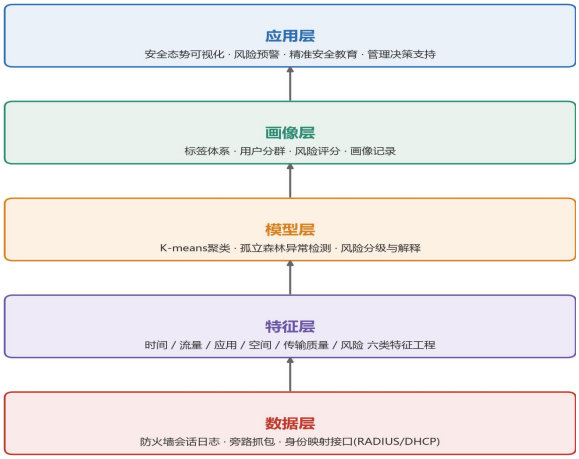


图 1.基于网络流量大数据的实验室人物画像总体架构



图 2.实验室用户画像四维指标体系

表 2.数据字段—画像特征—画像标签映射（节选）

Field	Image Features	Image Tag
Access Time	Percent Proportion of Each Time Period	Activate Time is Night or Daytime
Protocols	Percent Proportion of Each Behaviour	Entertainment, Study, Working
Upload/Download Data Traffic	Total of Data Traffic, Day of Data Traffic, Total of Large Data Traffic	Heave User, Download User
Upload/Download Retransmission	Retransmission Rate	Transmission Error
Upload/Download Reset	RST Rate	Connection Error
Domain	Number of Risky TLD Hits, Number of Risky Keywords Hits	Access Risky Domain
Target Location	Proportion of Cross-Border Conversation Rate	Cross-Boder Outreach

3.3 特征工程

本研究对“天”和“小时”两个时间窗口聚合以及 1125 个真实用户实体（经处理后的实际数据）会话级数据，最终生成 28 维画像特征，画像建模与安全应用的整体流程如图 3 所示。图中的六个阶段包含数据采集、数据预处理、特征工程、模型构建、画像生成、安全应用，其中底部虚线表示“安全教育效果评估与安全事件复盘→特征与阈值更新”，形成完整的反馈回路，具有一定的可复现性和可重复性。最终，在本阶段的内容中，得到最终结果，累计涵盖六大类：

（1）时间规律：实验室用户的活跃天数、高峰使用的小时数、深夜会话占比、时间活动规律性（时段熵）；

（2）流量特征：被测量用户群体的上下行流量，各标记用户的平均会话字节数、大流量会话数（>100MB）；

（3）应用特征：被测量用户所使用的应用种类数、各应用占比、娱乐/办公/学习类占比；

（4）网络质量：被测量用户在计量期间的网络重传率、RST 率；

（5）空间特征：可能存在被测量用户访问唯一目标 IP 数、唯一域名数、唯一地点数，在流量跨域中跨境会话占比数；

（6）风险特征：被测量用户在风险 TLD 命中数、风险关键词命中数（挖矿、赌博等关键词，实际部署可替换为威胁情报库），通过统计这些风险测量目标，对被测量用户进行高风险/低风险行为标记记录，作为重点关注人群。

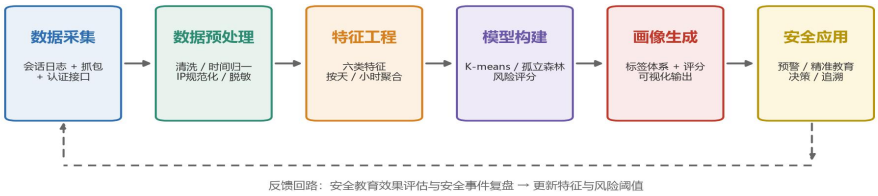


图 3.人物画像模型构建流程

3.4 建模算法

（1）用户分群：在本研究中，采用 K-means 算法对标准化特征聚类[10]，K 值的确定通过在研究中确定的轮廓系数（3 到 8）范围内寻找最优。研究发现最终结果显示 K=3 时轮廓系数最大（0.4121），三类群体特征及结果如表 3 所示，研究结果图像 PCA 二维投影如图 4 所示。

表 3.K-means 用户分群结果

分群类型	实体数	占比	风险评分均值	平均会话数	典型特征
办公型	595	52.9%	52.7	1，729	办公类应用占比 38.6%，风险命中极少
普通型	269	23.9%	34.6	540	活跃度低，各维度指标均衡
高风险型	261	23.2%	59.9	6，597	娱乐类应用占比 20.5%，风险命中显著偏高

（2）异常检测：本研究采用孤立森林算法对多维特征检测离群实体，在研究中采用自实现、100 棵树、每树采样 128 个样本等方法。最终结果显示：共识别出 4 个异常实体，且均属于高风险型，与风险评分结果相互印证。

（3）风险分级（扩展）：在通过“标注→获取”流程后，后续可训练随机森林/XGBoost 分类器实现“低—中—高”三级风险自动分级，并结合机器学习模型“SHAP”库的

解释模型算法，获取实验室用户群体的上网行为关键风险特征，进一步支撑管理部门实施精准安全教育。

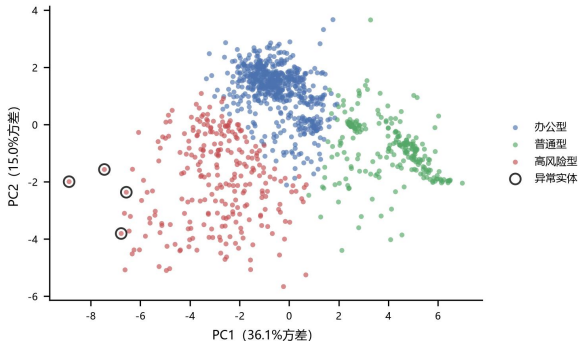


图 4.用户分群 PCA 二维投影

3.5 风险评分模型

在本研究中，为进一步量化实验室用户上网行为安全风险，通过构建 0~100 分的风险评分：先将各风险特征归一化处理，再进行加权求和。结合本研究实际，最终权重采用“熵权法客观赋权+专家先验修正+修整量”的混合策略，即式（1）所示：

$$\omega = 0.5 \times \omega_{熵权} + 0.5 \times \omega_{专家} + \varepsilon_{修整量} \quad (1)$$

其中专家权重由实验室安全管理人员结合实践经验给出，修整量由偏正结果调整。混合权重经归一化后如表 4 所示。

表 4.风险评分维度及混合权重

Risk Dimension	混合权重
Retransmission Rate	0.1809
RST Rate	0.0979
Proportion of Late-night Conversation	0.1427
Number of Heavy Data TrafficConversation （log）	0.1926
Risk TLD Hit（log）	0.0789
Risk Keyword Triggered（log）	0.1229
Proportion of Cross-Boder Conversation	0.0282
Total Data Traffic（log）	0.1559

按上述模型，1125 个实体的风险评分分布为：80 分以上 17 个、60~80 分 282 个、30~60 分 742 个、30 分以下 84 个。风险评分 Top20 用户均以“高流量+大流量会话+风险域名”为主要风险来源（图 5），可为安全员复核与预警处置提供量化依据。值得注意的是，图 5 的标记原则：条形颜色表示风险区间（≥80 红色、60—80 橙色、<60 蓝色）；虚线为 80 分预警阈值；n=20。

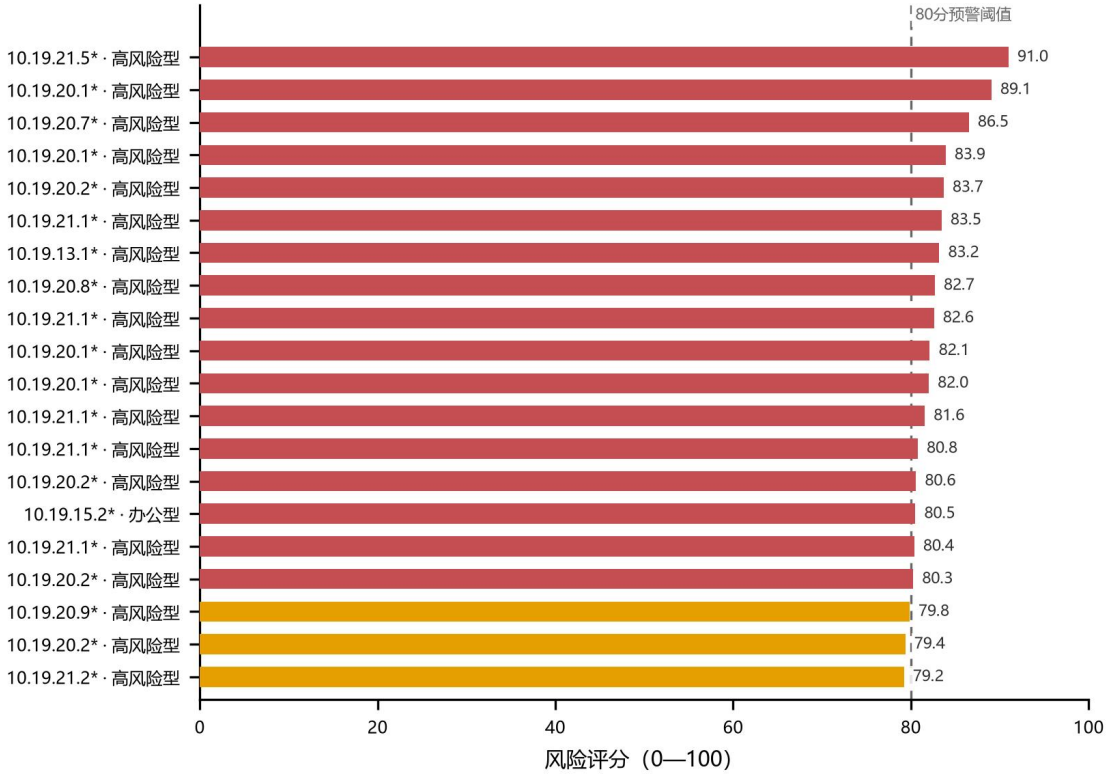


图 5.风险评分 Top20 实体

3.6 画像输出与可视化

本研究的最终画像结果以“一个实体一行记录”的画像方式记录表输出，结果包含了实体标识、分群类型、风险评分、主要风险与教育建议。在结果可视化方面，采用 K-means 分群以及 PCA 二维投影（图 4）与得出的风险评分 Top20 榜单（图 5）已在前文给出，此处进一步生成三类代表用户的雷达

图（图 6）与 Top20 活跃用户逐小时会话热力图（图 7），直观呈现群体差异与风险分布，为实验室安全管理提供决策支持。

图 6 表示的用户画像雷达图是通过对六个维度经 99 分位归一化（风险评分除以 100）后绘制；在结果中选取办公型、普通型代表，通过各群风险评分获得中位数附近用户，高风险型代表取风险评分最高用户；实体用户数 1125（其中办公型 595、普通型 269、高风险型 261）。

图 7 表示的会话热力图是按行进行总会话数降序排序，不同颜色表示日志记录归一化强度，白色表示无会话，累计选取 20 个实体。

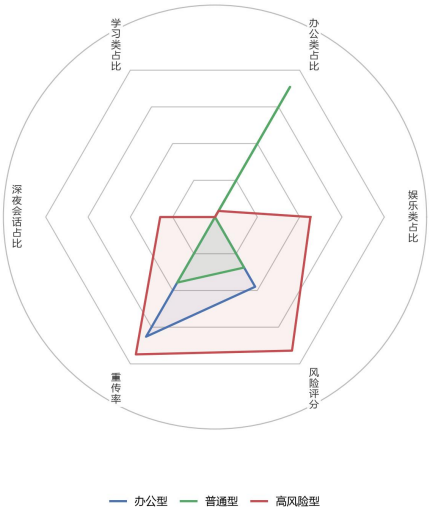


图 6.三类用户画像雷达图

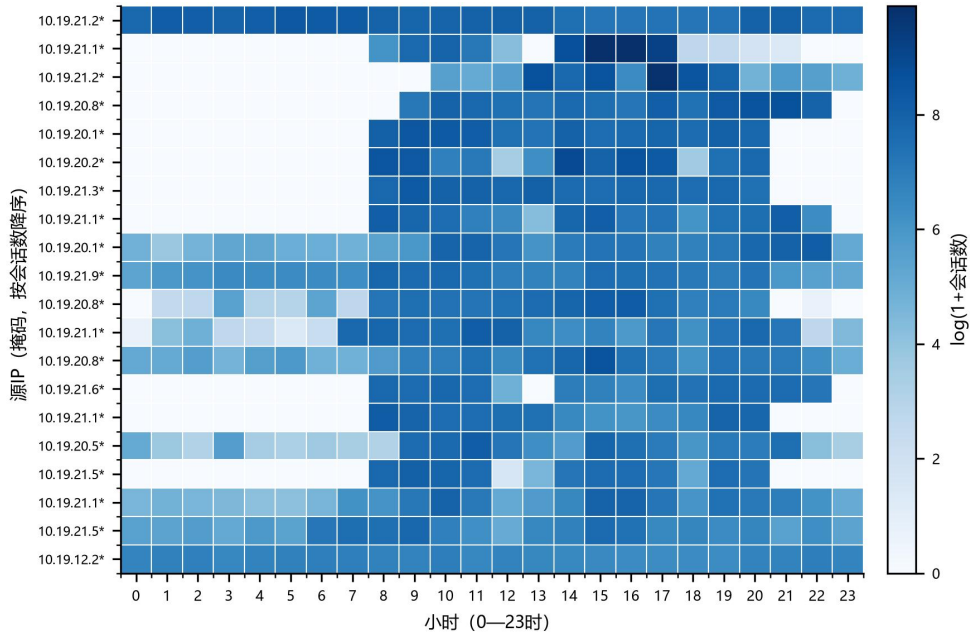


图 7.Top20 活跃用户逐小时会话热力图

4.实验室安全应用场景

4.1 安全态势可视化

本研究基于网络流量进行实验室用户画像，构建实验室网络安全态势仪表盘，搭建网络安全管理监测体系，从群体的分群占比和风险分布并结合个体用户画像两个层面呈现安全态势，实验室管理人员以及网络安全管理部门可直观掌握实验室网络使用的整体状况，通过授权后台认证信息接入，可以快速识别重点关注对象，精准实现育人目标。

4.2 高风险行为预警

本研究将风险评分与异常检测结果接入预警流程：针对风险评分超过阈值以及孤立森林识别出的异常实体，还有深夜大流量、风险域名访问、远程控制工具等高危行为，自动生成预警信息并推送至安全管理人员复核处置。

4.3 精准安全教育

本研究通过依托“教育画像”维度实现安全教育精准化：对高风险型群体进行实验室安全规范、网络安全行为规范、数据安全与隐私保护等专题内容实行定点、定向推送；对深夜活跃群体推送非工作时间上网行为安全注意事项；同时将预警响应情况纳入教育画像，形成“风险识别—教育干预—效果评估”闭环。

4.4 管理决策支持

实验室用户画像统计结果可以给实验室安全资源调配管理工作提供数据支撑。如提高高风险时段网络安全巡查频率，对可能存在高风险用户、高风险应用接入网络定制安全专项教育引导和管控策略；同时，画像数

据作为实验室管理的准入审核以及实验室年度安全评审依据。

4.5 安全事件追溯

当发生网络安全事件时，可使用成熟的画像记录技术，根据特殊或异常的 TCP、HTTP 流量明细，形成技术联动，快速响应安全事件的事故回溯并定位实施实体的行为特征，精准实施教学过程干预手段。

5.实践实施与成效

5.1 组织形式

本研究以“实验室安全大数据分析与画像建模”为主题，依托实验室安全管理部门和网络信息技术部门的授权以及技术支持，采用数据流量抓包辅以上网行为专用监测设备，不依赖特定课程，仅采集专用实验室在常规教学活动中的的网络数据流量情况，不会影响教学活动。

5.2 实施流程

本研究按“数据探索”测试、“数据预处理”模拟、“特征工程”实施、“模型构建”预设、获取“画像输出”、形成“安全分析报告”和“成果汇报”七个阶段实施，周期约 6 周。在每个研究阶段设置了明确的任务和预期成果以及成果形式，同时邀请实验室管理、网络安全管理等多部门技术成员对研究内容和成果进行技术评审和安全评估，确保分析结果符合安全实验室安全管理实际需求。

5.3 考核评价

采用过程性评价与成果性评价相结合：过程性评价主要是考核对预设阶段任务完成度考核，还有算法实现代码与过程文档质量、团队协作情况；成果性评价包括画像分析报告、最终形成的画像结果、在实验室安全管理、精准育人效果等的方面可行性和可复现性、以及将来在实验室安全管理和网络安全监测等方面提出的建议成效预估进行评估，最终做出合理推荐。

5.4 实施成效

表 5.基于模型的实施成效预测

成效指标	预测值	测算口径与依据
首批预警对象规模	18 个实体	≥80 分 17 个与异常 4 个取并集，其中 1 个低于 80 分
预警名单风险域名/关键词命中率	72.2%（13/18）	命中风险 TLD 或关键词，为全体基线（10.0%）的约 7.3 倍
预警名单显性风险复核命中率	100%（18/18）	均含风险域名/关键词、大流量或跨境等可复核证据
定向教育对象规模	261 个实体（占 23.2%）	K-means 高风险型分群
教育对象风险证据覆盖率	66.3%（173/261）	含风险域名/关键词、大流量或跨境任一项
风险关键词命中率（高风险型对比全体）	37.9%对 9.2%（约 4.1 倍）	高风险型分群画像与全库基线对比
风险 TLD 命中率（高风险型对比全体）	6.5%对 1.5%（约 4.3 倍）	高风险型分群画像与全库基线对比
人工复核工作量（预警对比全量普查）	减少约 98.4%	复核由 1，125 个实体减至 18 个（1−18/1，125）

受完整教学全过程监督和教学活动周期等因素影响，本研究暂处于模拟假设阶段，暂无完整的实测数据，无法精确完整地预估最终的育人效果。此处暂按支撑方案预设与目标设定，以本文中已完成的对上千个实体终端设备（未获的实名授权认证信息），对现有数据进行了成效预测（表 5）。

由表 5 可见：其一，预警名单在风险域名/关键词上的直接命中率（72.2%）为全库基线（10.0%）的约 7.3 倍，且名单内全部实体均具备可复核的显性风险证据，说明模型预警具有较高的对象指向性；其二，若按高风险型群体定向实施安全教育，其风险关键词与风险 TLD 命中率分别为全库基线的 4.1 倍与 4.3 倍，风险信号在拟教育对象中显著聚集，为差异化教育提供了可测度的人群边界；其三，模型预警可将人工复核范围由 1，125 个实体收窄至 18 个预警实体，工作量减少约 98.4%，使有限的安全管理资源集中于高风险对象。

上述预测仅为基于历史流量数据的模型回代结果，不等同于干预后的实测成效。实施完成后，需按参与人数与学生成果清单、安全员复核命中率、安全教育覆盖率、干预前后高风险群体风险关键词/TLD 命中率及风：表 5 并评估“风险识别—教育干预—效果评估”闭环（4.3 节）的实际效果；实测结果亦将反馈至特征与风险阈值的迭代更新（图 3）。

6.结论与展望

本文面向高校实验室安全管理需求，基于真实防火墙流量数据构建了“身份—行为—风险—教育”四维人物画像模型，通过用户分群、异常检测与风险量化，实现对实验室网络使用者安全风险的精准刻画，为安全预警、精准教育与决策支持提供了数据支撑，并探索了以真实数据项目为载体的实践育人模式。将设备流量与人物流量分离、采用混合加权风险评分，是提升画像可用性的两个关键做法。

本研究也存在明显局限：数据观测周期为 11 天，时间规律建模的稳定性有待延长采集后验证；会话日志缺乏用户账号字段，人物级画像依赖身份融合接口的后续接入；风险识别依赖静态关键词与 TLD 清单，对加密流量与新型风险覆盖不足。

下一步将围绕以下几点展开：接入上网认证系统，实现主机画像向人物画像的聚合；延长旁路采集周期，提升时间规律建模的稳定性；引入威胁情报库与加密流量识别技术，提高风险识别能力；探索差分隐私等隐私保护技术，在保障数据安全的前提下实现跨单位推广应用。

参考文献

[1] 中华人民共和国教育部.高等学校实验室安全管理办法[Z].2015.
[2] 王元卓，靳小龙，程学旗.网络大数据：现状与展望[J].计算机学报，2013，36（6）：1125-1138.
[3] 孟小峰，慈祥.大数据管理：概念、技术与挑战[J].计算机研究与发展，2013，50（1）：146-169.
[4] 冯登国，张敏，李昊.大数据安全与隐私保护[J].计算机学报，2014，37（1）：246-258.
[5] 刘海鸥，孙晶晶，苏妍嫄，等.用户画像模型构建研究综述[J].数据分析与知识发现，2019，3（3）.
[6] 高广尚.用户画像模型构建方法研究综述[J].情报科学，2019，37（9）.
[7] 张玉清，董颖，柳彩云，等.深度学习应用于网络流量识别的综述[J].计算机研究与发展，2018，55（9）.
[8] 国家市场监督管理总局，国家标准化管理委员会.信息安全技术个人信息安全规范：GB/T35273—2020[S].北京：中国标准出版社，2020.
[9] Cooper A. The Inmates Are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity[M]. Indianapolis: Sams Publishing, 1999.
[10] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques[M]. 3rd ed. San Francisco: Morgan Kaufmann, 2011.