文章编号：2095-4980(2022)12-1257-12

内部安全威胁与室内人群移动轨迹基准数据集

赵　颖[1]，赵　鑫[1]，杨　奎[1]，陈思明*[2]，张　卓[3]，黄　鑫[3]

(1.中南大学 计算机学院，湖南 长沙 410083；2.复旦大学 大数据学院，上海 200433；3.奇安信科技集团股份有限公司 雷尔可视化平台部，北京 100015)

摘　要：数据集是众多科学研究得以开展与验证的基础，学术界和工业界已经联合在许多领域打造了丰富的基准数据集，但在一些细分研究领域仍然缺少高质量数据。本文介绍了2个新基准数据集：内部安全威胁基准数据集和室内人群移动轨迹基准数据集。2个数据集经过精心的场景设计、科学的模型构造，嵌入了丰富的数据模式和交错的故事情节，采用程序驱动的合成数据生成方法，数据类型多样，规模适中，有一定的分析难度，曾被用于中国数据可视分析挑战赛。本文旨在进一步宣传和推广这2个数据集，以促进相关领域的科学研究与技术应用的发展。

关键词：基准数据集；内部安全威胁；室内人群移动轨迹；数据可视分析挑战赛

中图分类号：TN520.60　　**文献标志码：**A　　**doi：**10.11805/TKYDA2021143

Benchmark datasets for insider threat detection and indoor crowd behavior analysis

ZHAO Ying[1]，ZHAO Xin[1]，YANG Kui[1]，CHEN Siming*[2]，ZHANG Zhuo[3]，HUANG Xin[3]

(1.School of Computer Science，Central South University，Changsha Hunan 410083，China；
2.School of Data Science，Fudan University，Shanghai 200433，China；
3.Layer Visualization Department，Qi An Xin Technology Group Co.，Ltd.，Beijing 100015，China)

Abstract：Benchmark datasets are crucial for many data-dependent scientific studies and technology applications. Academic and industry communities have closely collaborated to release abundant datasets in many fields. However, there is still a lack of high-quality benchmark datasets in some specific domains. This paper introduces two open-source benchmark datasets, namely, the Insider Threat Dataset (ITD-2018) and the Indoor Crowd Movement Trajectory Dataset(ICMTD-2019). The two datasets are produced by program-driven synthetic data generation methods and are presented with well-defined scenarios, carefully-designed behavior models, rich data patterns, and vivid storylines. The two datasets were used in the ChinaVis Data Challenge. This paper aims to promote the two datasets for the development of the research and technology in relevant domains.

Keywords：benchmark；insider threat；indoor crowd movement trajectory；ChinaVis Data Challenge

数据集是科学研究的基础，是检验科研成果的试金石，对推动各学科领域的发展发挥着重要作用[1-3]。一方面，许多学科领域的科学发现需要依托于已有数据集；另一方面，许多学科领域需要使用已有数据集来重复科学试验，验证其研究成果的真实性与可靠性。因此，科学研究对数据集的需求与日俱增，逐渐表现出数据紧密型和以数据为中心的特征。但在很多细分学科领域，如：内部安全威胁和室内人群管理等，仍然缺少高质量的数据集。

企业安全是企业能正常运行的前提。近年来，企业信息化建设的快速发展显著提升了企业办公和生产效率。但企业安全问题也日趋凸显，尤其是企业内部潜在的安全威胁问题[4-7]。在信息化环境中，企业中可能存在恶意员工组织实施的一系列恶意行为，如信息窃取、漏洞利用、恶意软件嵌入等，对企业内部安全构成了严重威胁。因此，业界、学界等相关研究人员对企业内部安全的研究需求日趋增长[8-10]。企业内部的日志数据记录了员工在

收稿日期：2021-04-08；**修回日期：**2021-06-10

基金项目：国家自然科学基金资助项目(61872388)

*通信作者：陈思明　email:simingchen3@gmail.com

使用信息化服务时的操作，为研究人员进行企业内部安全的相关研究提供数据基础。但由于企业内部数据涉及商业机密和员工隐私问题，迄今为止，公开的企业内部安全威胁数据集很少[11−15]，且在目前已公开的企业内部安全威胁数据集中，普遍存在数据源类型不够丰富，涉及的安全威胁类型不够广泛，包含的故事线比较单一，以及缺乏完整的数据集故事真相和评估方案等限制。与内部威胁检测领域巨大的研究需求形成了鲜明对比，在一定程度上限制了相关研究的发展与进步。

随着城市基础设施的完善和普及，人们大量的时间处于写字楼、车站、地铁、医院、商场等室内环境，随之产生了大量室内人群移动轨迹数据[16−23]。室内人群移动轨迹数据包含了人类在小范围内的细粒度移动模式和行为特征，对智慧场馆服务与管理、群体时空行为模式分析、室内导航、室内热点推荐等方向具有较大的研究与应用价值[24−28]。但由于室内定位技术的应用成熟度和普及度仍不足，导致可收集到的室内移动轨迹数据很少。由于室内场景涉及敏感信息和个人隐私，公开的室内移动轨迹数据集很少，且在目前已公开的室内移动轨迹数据集中，普遍存在故事线不够完整、轨迹多样性不够充分、移动场景不够细致、移动行为的可信度难以评估等限制，远远不能满足研究人员对室内人群行为分析的研究需求。

针对以上现状，本文介绍了2个公开数据集：内部安全威胁数据集(ITD−2018)[29−30]和室内人群移动轨迹数据集(ICMTD−2019)[31−32]。这2个数据集根据现实世界类似场景，通过科学建模和程序驱动的方法生成。2个数据集包含精心设计的多重故事线和完整的故事情节，能够反映出定制场景中逼真化、多样化的人类行为，为相关领域的研究提供了新数据资源，曾经分别应用在2018年和2019年的ChinaVis数据可视分析挑战赛中[33−35]。本文对2018年和2019年挑战赛收到的152份作品，进行了领域专家评估、分析质量评估和事件难度评估，并对参赛选手进行了问卷回访。结果显示，这2个数据集具有较好的完整性、可用性和有效性，数据集场景设计逼真，数据集中丰富且难度各异的人类行为模式和异常事件不仅能充分激发用户的分析热情，还能有效鉴别数据分析方法、技术和系统的可用性与有效性。本文旨在进一步宣传和推广这2个数据集，以促进相关领域的科学研究与技术应用的发展。

1 内部安全威胁基准数据集

1.1 场景设计

内部安全威胁基准数据集的场景设置在一个中等规模的HighTech虚拟互联网公司中。公司正在全力研发一款重量级高科技产品，该数据集的时间线设置在产品发布前夕的一个月内。为了使场景更具真实性、普遍性、多样性和挑战性，HighTech公司的部门与人力设计参考了某真实互联网企业。公司内设置有财务部、人力资源部和研发部3大部门，有行政总裁、部门经理、组长和普通员工4种行政级别的299位员工。员工的行为特征设计丰富多样，所属不同部门的员工表现出各部门代表性的工作行为，如：财务部门员工主要负责财务制度制定、会计核算等工作内容；人力资源部门员工主要负责绩效考核、劳动合同签发、福利保障设置等工作内容；开发部门员工日常工作主要围绕需求分析、软件开发等。同时，每个员工也具有不同的行为偏好。场景中设计有分析难度不同的多重故事情节，包括一系列互联网高科技公司的常见工作事件，如加班、辞职、员工球赛等，还嵌入有一系列与产品发布相关的高威胁事件，如产品数据泄露、关键资产损坏等，其中一些事件相互关联，涉及到多数据源的协同分析与线索融合。

1.2 数据建模与数据描述

为了逼真地还原预设场景，设计了对象模型、关系模型和威胁事件模型。对象模型包括员工、部门、资产模型，分别用来刻画定制场景中公司员工的年龄、性别等个人基本属性，部门、级别等工作属性，日常上下班时间、经常浏览的网站等行为属性；定制场景中部门的日程安排，如部门上下班时间、休息时间等；以及定制场景中的公司资产，如IP、服务器等。关系模型主要包括2部分，分别描述公司各部门之间的人事关系和公司资产之间的关联关系。人事关系即部门与部门员工的层级关系，如：CEO作为公司高层全面管理公司；部门经理是各部门的主管，管理其下属的组长和普通员工。公司资产关联关系，即资产与各部门及员工的所属关系，如：电子邮件服务器服务于所有部门员工，开发和备份服务器只分配给研发部员工。威胁事件模型，用于刻画一系列内部威胁行为，这些可能会影响到公司的正常运营与新高科技产品的发布。

基于以上3类模型，设计实现了一个数据生成器，按照单人单日的策略生成员工的行为数据，主要包括以下4大步骤。首先创建并初始化所有模型；然后在时间片的推进下，使用行为调度器为员工在部门日程安排基础上分配个人行为；接下来由行为控制器交付相应的参数化表达给数据转换器，生成相应的背景数据。同时，还设计有威胁事件驱动脚本，来描述公司中存在的恶意行为，并驱动生成威胁事件数据；最后，融合背景数据与威

胁事件数据，消除二者之间的数据冲突，得到最终的内部安全威胁数据集。文献[30]及附加材料中详细介绍了该数据集的场景设计，对象模型、关系模型和威胁事件模型的具体建模方法，生成数据集的程序驱动方法和数据集的评估方法与结果。

内部安全威胁数据集包括员工打卡日志数据、网页访问日志数据、电子邮件日志数据、服务器登录日志数据和TCP流量日志数据5大类数据源，每类数据源之间通过各种各样的方式关联在一起，共计133.4 MB。每类数据源的详细说明如下。

员工打卡日志数据：记录公司每位员工每天的上班、下班打卡时间。当员工没有上班时，也会生成当日的打卡日志数据记录，并设置打卡记录时间为0。而且，该员工还会在次日收到一条旷工提醒邮件。

网页访问日志数据：记录所有员工的网页访问记录，具体包括，记录生成时间，客户端IP地址与端口，服务器IP地址与端口，员工访问的服务器域名。当员工直接通过其主机的IP地址访问目标网站，没有经过域名系统(Domain Name System，DNS)解析时，服务器域名的HTTP报头字段记录的主机名为空字符串。

电子邮件日志数据：记录经过公司邮件服务器的收发邮件信息，具体包括：邮件主题，邮件发送者、发送时间、发送端IP和端口，邮件接收者、接收时间、接收端IP和端口。

服务器登录日志数据：记录员工登录服务器的详细信息。员工可以通过使用公司配备的工作站主机直接登录到目标服务器，也可以通过一些跳转服务器来登录到目标服务器。一条服务器登录日志数据记录中，具体包括：登录时间、用户名、协议、访问端IP和端口、目的端IP和端口、登录结果。

TCP流量日志数据：记录公司内部网络活动产生的所有TCP连接。任何一条邮件收发行为、网页浏览行为或服务器登录行为等，都会产生一条或多条TCP记录。一条TCP流量日志数据记录中，具体包括TCP连接的起止时间、使用协议、发送端IP和端口、目的端IP和端口、字节数。

1.3 数据真相

1.3.1 背景说明

HighTech公司近期有一款重量级高科技产品将要发布，为了保护公司的核心利益，确保新产品的顺利发布，公司高管高度警惕，成立了内部安全威胁情报分析小组分析内部系统采集到的日志数据，其主要任务是分析并处置可能存在的各种安全威胁。具体包括，分析公司内部员工所属部门及各部门的人员组织结构；按部门分析、总结公司员工的日常工作行为模式；分析异常事件以及事件之间可能存在的关联，发现、总结有价值的威胁情报。

1.3.2 主线事件

本文提出的内部安全威胁数据集由2条主线故事和7条支线故事构成，故事线索错综复杂，分散在5类日志合成的多源异构数据中。

主线故事主要包括产品数据泄露和关键资产损坏2大威胁事件。产品数据泄露事件讲述了一名商业间谍员工在公司新产品发布前夕盗取产品相关资料并将其泄露出去的过程。X公司与Hightech公司是2个有竞争关系的互联网公司，X公司为了及时掌握Hightech公司的动向，以便于在竞争中取得优势，派遣了一名商业间谍进入Hightech公司作为一名普通员工，以窃取Hightech公司的重要信息。该商业间谍在Hightech公司的员工编号是1487，在该公司工作已一年有余。近期Hightech公司即将发布一款重量级科技产品，员工1487收到命令需要在产品发布前夕盗取该产品的相关资料并将其提前泄露，以达到打击Hightech公司的目的。员工1487为了完成任务且不被察觉，制定了一个非法计划。首先，他盗取了一位公司领导的账号以获取更高的资料查阅权限；然后，趁机在一次突发数据库故障的维护工作中，使用领导账号登录公司服务器，违规查看新产品的资料；最后，在数天后使用领导账号通过使用一次跳转服务器登录到新产品资料所在的公司服务器，盗取产品资料并上传到外界服务器。由于担心事情败露，员工1487完成任务后在月底申请了辞职。

该主线故事由盗取领导账号、违规查看产品信息、恶意泄露产品信息、商业间谍辞职4个主要的异常事件组成，员工1487的详细作案流程如图1所示。

盗取领导账号(E1)：员工1487在2017年11月03日、2017年11月04日、2017年11月06日分别尝试登录领导1080,1211,1228的账号，频繁登录失败，最终由于1228的密码口令等级较弱而被该员工成功破解。

违规查看产品信息(E2)：2017年11月16，员工1487报名参加了公司的集体打球活动，但由于数据库突发故障需要参与数据库维护，他实际并未参加打球活动。在数据库维护期间，他趁机使用1228的账号登录了目标服务器10.50.50.44，确保服务器上有新产品的相关资料。

恶意泄露产品信息(E3)：2017年11月24日12:43分，员工1487使用1228的账号登录服务器10.50.50.43，并

以此服务器为跳板登录目标服务器 10.50.50.44，向外界服务器 13.250.177.223 上传数据。

商业间谍辞职(E4)：员工 1487 在这期间经常浏览招聘网站，收到很多猎头的邮件，最终于 2017 年 11 月 27 日提出辞职申请。

关键资产损坏事件讲述了一位开发部员工 1376 的工作失误。由于员工 1376 早已计划从 Hightech 公司辞职，因此，他经常对工作心不在焉，应付了事。2017 年 11 月 16 日 19:22 分，员工 1376 由于一次错误操作，导致一台关键服务器的数据库故障，因此，员工 1376 和该部门另外两名员工收到了故障数据库的报警邮件。当晚，这 3 名员工一起对数据库进行了维护。因为员工 1376 的不当行为对公司资产造成了严重的影响，该员工在月底提出了辞职。该主线故事由数据库故障、数据库维护和删除者辞职 3 个主要的异常事件组成，事件具体过程如图 1 所示。

数据库故障(E5)：2017 年 11 月 16 日 19:22，员工 1376 由于工作失误造成 10.63.120.70 服务器的数据库故障，系统向员工 1284 和员工 1487(X 公司商业间谍)发送了数据库报警邮件。

数据库维护(E6)：员工 1284、1376、1487 当晚一起进行数据库维护工作，于当日 23 点 30 分左右完成工作后离开公司。

删除者辞职(E7)：员工 1376 在本月频繁浏览招聘网站，并多次收到猎头公司的邮件。数据库故障事件发生后，该员工于 2017 年 11 月 27 日提出辞职申请。

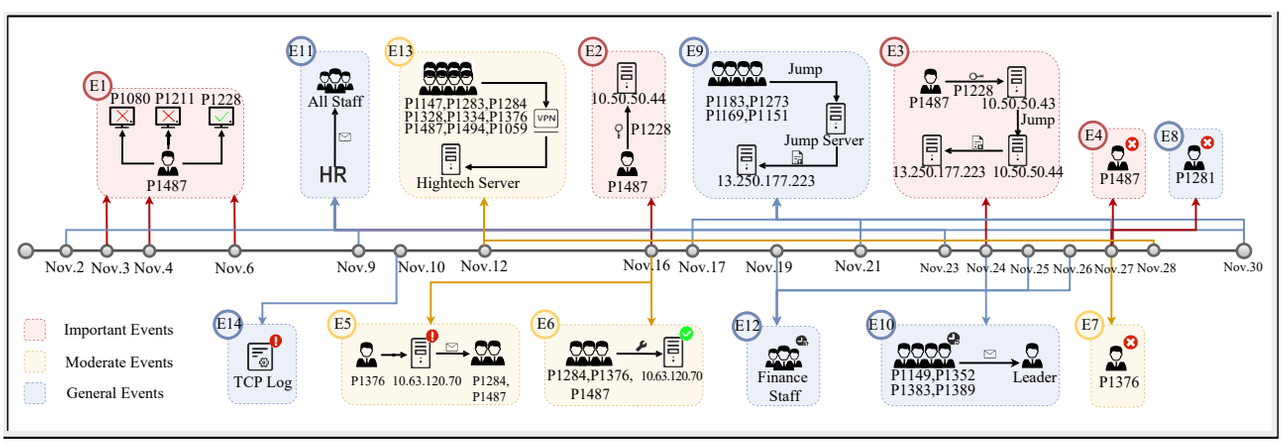


Fig.1 Main plots overview of ITD-2018

图 1 ITD-2018 故事概览

1.3.3 支线事件

除了以上两大错综复杂的主线故事外，提出的内部安全威胁数据集中，还包括以下 7 个经常发生在企业中的独立支线事件，如员工加班、系统故障、员工集体活动等，这些事件不会对公司的新产品发布构成直接威胁，如图 1 所示。

离职事件(E8)：2017 年 11 月 27，员工 1376,1281,1487 提交了辞职申请，并于第 2 天辞职申请审批通过后离职。其中，员工 1376 早已产生换工作的打算，在遇到数据库故障事件后于当月月底辞职。员工 1487 则是已经完成数据泄露任务，担心事情败露而产生换工作的打算，员工 1376 和 1487 的日志数据中显示，他们都频繁浏览招聘网站且邮箱中收到较多猎头和 HR 的邮件。员工 1281 由于家中出现重大变故，故而突然提出辞职申请。

跳板事件(E9)：除了泄露数据的员工 1487 外，在 2017 年 11 月 17 日、2017 年 11 月 21 日、2017 年 11 月 27 日、2017 年 11 月 30 日，员工 1183,1273,1169,1151 也先后通过跳板向外界服务器 13.250.177.223 上传数据，且 4 位员工的网络访问均在正常范围内。

旅游事件(E10)：员工 1149,1352,1383,1389 计划一起出去旅游，他们在 2017 年 11 月 20 至 2017 年 11 月 24 日期间频繁浏览旅游网站，并在 2017 年 11 月 24 日向各自的部门领导发送了请假邮件，请假时间为 2017 年 11 月 27 日至 2017 年 11 月 30 日。

员工集体活动(E11)：每周四早上 9:30 人力资源部门会向所有员工发送打球邀请的邮件，有意向参加的员工回复确认邮件。大部分参加活动的员工均在 19:00~19:20 这段时间区间离开公司。2017 年 11 月 16 日 9:30 分，员工 1487 和 1376 报名了打球活动，但由于数据库故障而参与了数据集维护工作，没有参加打球活动。

财务月底加班(E12)：由于月底财务工作繁忙，在月底的周末，即，2017 年 11 月 19 日、2017 年 11 月 25 日、2017 年 11 月 26 日，财务部绝大部分员工来到了公司加班。

VPN远程访问(E13)：1147,1283,1284,1328,1334,1376,1487,1494八名员工曾在周末通过VPN远程连接到公司内网加班工作；员工1059在2017年11月28日没来公司，通过VPN远程连接到公司内网审批了自己主管部门的两名员工1376和1487的辞职申请。

流量监控系统故障(E14)：由于TCP日志系统存在漏洞，导致2017年11月10日至2017年11月28日，TCP流量日志数据中一些邮件收发记录的网络协议本应该是SMPT，却被标记为了HTTP。

2　室内人群移动轨迹基准数据集

2.1　场景设计

室内人群移动轨迹基准数据集的场景设置在一个大型国际网络安全学术会议中。本次会议的场景设计参考了多个真实的网络安全学术会议，并邀请一位场馆管理人员和两位学术会议组织者参与了整个场景的设计过程，以确保场景的真实性和专业性。该数据集场景中设置有7种类型的会场人员，分别是：VIP嘉宾、普通参会人员、参观人员、媒体记者、黑客大赛参赛者、工作人员和参展单位。每类人员都具备特定参会权限和特有移动模式。根据某大型国际会议中心实景，设计了一个双层式且可容纳5 000人的大型场馆，馆内有主会场、分会场、展区、比赛区等功能区域，服务台、餐厅、茶歇点、休闲区、洗手间等基础便利设施，还提供了VIP休息室、媒体间、工作间等便于相关类型人员休息和工作的房间。参考某实际国际网络安全学术会议，设计了会议活动、普通事件、异常事件这3类事件。会议活动包括会议主旨报告、学术研讨、座谈会等学术性活动。普通事件包括商业展览、黑客大赛、采访、茶歇、晚宴等学术会议中较为常见的交流性活动。异常事件包括卡复制、物品遗失、签售会等12个涉及到细粒度时空特征变化和复杂关联的特殊事件。

2.2　数据建模与数据描述

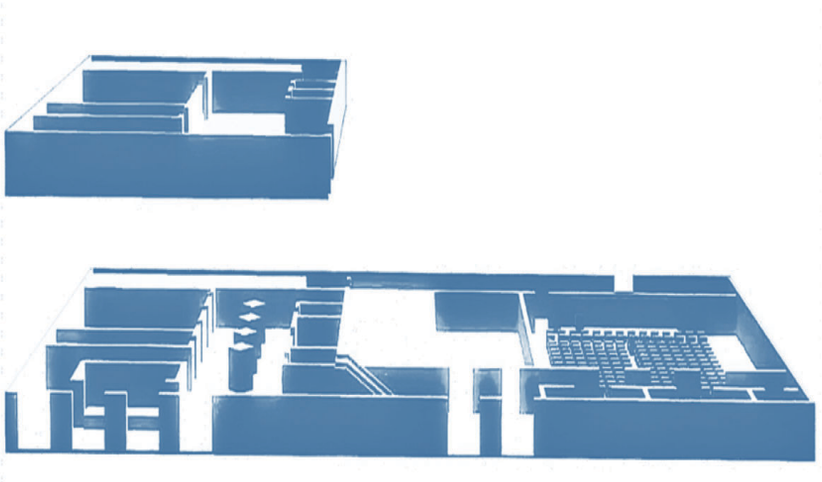
为了生成符合上述预定义场景的室内人群移动轨迹数据集，首先，构造了会议场馆中客观存在的人、事件、空间3大对象模型，并通过基本属性、运动与行为状态属性刻画移动人员的背景身份，和人员在会议场馆中的实时状态；通过基本属性、权限属性、优先级属性和状态属性刻画事件的背景信息与实时进度；通过基本属性、容量属性和状态属性刻画空间的基本结构、功能分区以及实时容纳情况。在3大对象模型的基础上，构造了行为模型，包括一系列非线性衰减与恢复函数来约束人的行为，以及一系列兴趣驱动函数来驱动移动人员对会场事件产生兴趣从而自发移动，以及一系列行为决策策略，来控制移动人员在会议场馆中按照一定的条件切换感兴趣的事件，控制人的入场与离场，从而使人表现出接近于真实世界的移动模式和分布特征。

最后，基于以上模型，设计了一种由程序驱动的数据生成系统驱动数据生成。首先，对3大对象模型和行为模型进行创建与初始化。然后，在时间片的推进下，使用状态更新器更新所有人员的状态属性，使用行为控制器对人进行行为约束、兴趣驱动和兴趣控制，使用路线分配器对需要移动的人分配路线并记录人的位置信息，得到背景数据。同时，设计有异常事件驱动脚本，配置了预设的异常事件并驱动异常事件发生，生成异常事件数据。最终合并背景数据和异常事件数据，消除数据冲突，得到室内人群移动轨迹数据集。文献[32]详细介绍了该数据集的场景设计，3个实体模型与3个行为模型的建模方法，生成数据集的程序驱动方法和数据集的评估过程与结果。

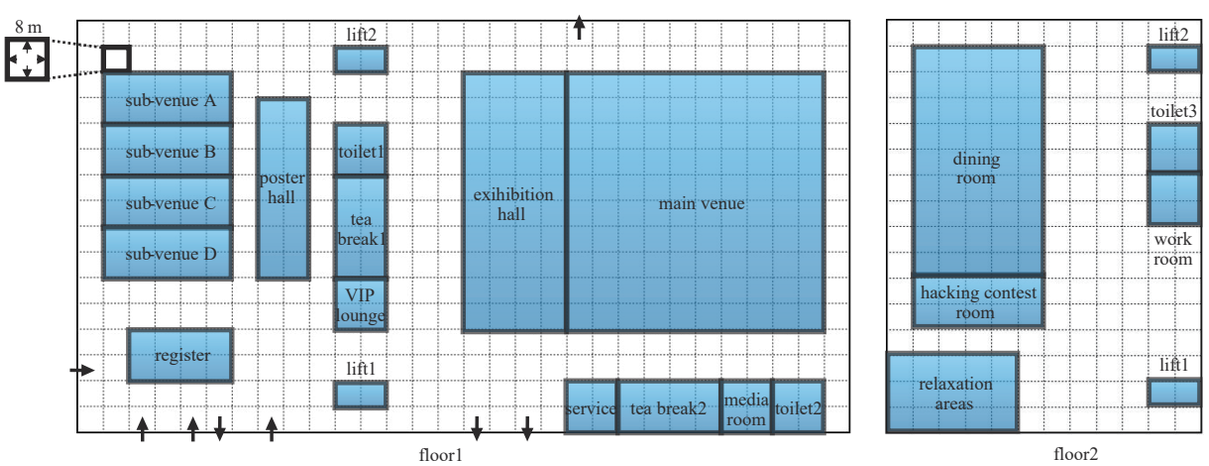
室内人群移动轨迹数据集共包括2大数据源：会议场馆传感器分布数据，描述传感器在场馆内布置的具体位置；传感器日志数据，记录会议期间每个传感器收集的参会人员移动轨迹信息。数据集中共包括1,879,485个移动轨迹点，合计34.1 MB。

传感器分布数据描述了所有传感器在会场空间中的位置分布情况，包括传感器分布地图和传感器分布表。传感器分布地图如图2所示，图2(a)为会场地图的三维空间模型，图2(b)为会场地图的二维俯视图，描述会场空间结构与功能分区，图中的每一个正方形格子表示一个传感器的覆盖范围，对应真实世界中8 m×8 m的区域。传感器分布表记录了每个传感器的编号、所在楼层，以及在会场二维地图中对应的横、纵坐标位置。

传感器日志数据以CSV格式分3天给出，一条数据记录中具体包括人员编号、传感器编号和时间戳记录。考虑到当一个人专注于会议活动而没有移动行为时，他们会在连续时间内在相同位置产生许多轨迹点。因此，删除了这种重复的轨迹点，仅保留了人员产生位置变化时的传感器日志数据，以减小数据集规模，突出人员在会场中的移动行为。



(a) 3D venue model



(b) 2D top view of venue

Fig.2 Sensor distribution map

图 2 传感器分布地图

2.3 数据真相

2.3.1 背景说明

提出的室内人群移动轨迹数据集中，记录了会议期间 5 256 人在会场中活动的移动轨迹。本次会议是一个虚构的智能网络安全领域的学术会议，会议为期 3 天，邀请了多位资深专家做主旨报告，设置了数据安全、物联网安全、移动安全、隐私保护、智能场馆和智能安全技术创新 6 大研讨主题，还举办了商业展览和黑客大赛。本次会议的组委会为了便于会场管理人员及时合理地调动资源，处理各种突发状况，为会议各项议题的顺利进行保驾护航，临时成立了一个数据分析小组，其主要任务是根据采集到的传感器日志数据，总结各类型人员的移动规律，发现存在的异常事件，以便于协助组委会管理会场，响应和处理各类应急事件。

2.3.2 人员说明

本次会议中，会场中移动的人员包括 VIP 嘉宾、普通参会人员、参观人员等 7 种类型，每类人员具有不同的会场权限，包括午餐权限、晚餐权限、VIP 休息室权限，只有具有相应权限的人员才可以进入对应会场区域。而且，每一类会场人员也呈现出不同的基本移动模式，如表 1 所示。

2.3.3 异常事件

将会场人员遵循其对应会场权限和移动模式下的常规移动视为正常行为，除此以外，会议期间发生了以下 12 个值得关注的异常事件，使得事件相关人员的移动轨迹呈现出特定模式或移动轨迹，可能对会场的管理与安全带来一定影响，如图 3 所示。

胸牌复制(E1)。人员 A 使用复制了某 VIP 人员 B(16632)的胸牌，并佩戴复制胸牌进入会场 VIP 休息室，停留 1 h 左右，怀疑 A 有行窃行为。

物品遗失(E2)。VIP 人员 C(11260)午餐后在 VIP 休息室休息，发现物品遗失后频繁去服务台处问询。考虑到

A 在 VIP 休息室停留时间很长，且在 C 返回 VIP 休息室 4 min 后就匆匆离开会场，因此怀疑 A 盗窃了 C 的物品。

表 1 各类人员的会场权限和基本移动模式

Table1 Venue permissions and basic movement patterns for each type of venue personnel

type	venue permissions			movement patterns
	lunch	dinner	VIP lounge	
VIP guest	yes	yes	yes	enter the venue without signing in; mainly attend conferences in the main venue and sub-venues and take a seat at the front of the venue;
ordinary guest	yes	no	no	enter the venue with sign-in; mainly in the venues, exhibition hall, poster hall and other areas; self-organized conference attendance and exhibition schedule;
visitor	no	no	no	similar to regular guests, but not available to attend the conferences in the main venue;
media reporter	yes	yes	no	enter the venue with sign-in; mainly in the media room and various areas of the venue for interviews, minutes of meetings, etc;
hacking contestant	yes	no	no	enter the venue with sign-in; mainly in the hacking contest area;
staff	yes	no	no	enter the venue earliest; mainly service in various areas of the venue and often take breaks or lunch in the work room;
exhibitor	yes	yes	no	enter the venue with sign-in; mainly in the exhibition hall

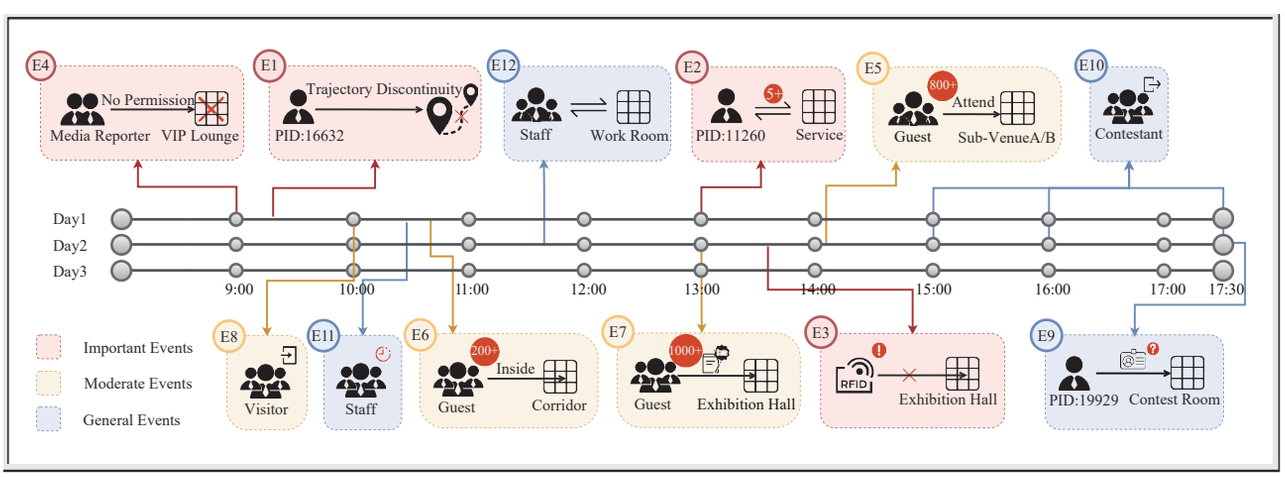


Fig.3 Abnormal events overview of ICMTD-2019

图 3 ICMTD-2019 异常事件概览

设备故障(E3)。会议第二天 13:33:21~13:56:15，传感器(10715,10716,10717,10718,10815,10816,10817,10818)故障，导致日志数据缺失。

人员越权(E4)。VIP 休息室、媒体间分别是专属 VIP 嘉宾和媒体记者的房间，其他类型人员无权进入。但是会议第一天，2 位媒体记者(11201 和 16473)分别在 9:00~9:20,10:30~10:50 越权进入 VIP 休息室；会议第二天，VIP 人员(13344)在 12:29:04~12:39:50 进入媒体间。

部分分会场爆满(E5)。有 3 场分会议受到很多参会人员的关注，人数爆满。分别是第二天 14:00~16:15 在分会场 B 举行的物联网安全论坛，第三天 9:30~11:30 在分会场 B 举行的移动安全论坛，第三天 10:30~11:30 在分会场 A 举行的智能安全技术创新论坛。

会场拥堵(E6)。会议中途有多个茶歇时段，但茶歇点、卫生间等区域空间有限，工作人员也没有很好地引导大家有序活动，导致茶歇期间，茶歇点和各会场走向茶歇区的部分通道以及各厕所出现人群拥挤。

签字售书(E7)。第二天 13:00~14:30 期间，展厅区域举办了新书签售会。大量会场人员向签售会区域聚集，传感器编号为 10215,10315,10415 的区域在该时间段检测到人员密度较大。

团体参观事件(E8)。会议期间共有 4 个参观团体到会场的展厅、海报区域参观游览。分别是第一天 10:00~11:00, 15:00~16:00，第二天 10:00~11:00, 15:00~16:00。参观团体一般集体行动，规模 100 人左右。

胸牌遗忘(E9)。VIP 人员 G(19929)是黑客大赛第二天主持人。第二天 9:00，G 来到黑客大赛会场，把胸牌遗忘在讲台，直到 17:30 左右黑客大赛结束，G 才回到讲台处取回胸牌离开会场。

黑客大赛参赛者提前离场(E10)。黑客大赛分为基础考核和附加考核两个环节。会议第一天上午和第二天上午进行基础考核环节，黑客大赛参赛者须在规定时间内答题。第一天下午、第二天下午和第三天上午进行附加考核环节，该环节采取淘汰赛制，部分被淘汰的参赛者呈阶段式时间离开会场。

工作人员迟到(E11)。会议期间，工作人员会提前进入会场，到达各自工作岗位，但部分工作人员存在迟到

现象，包括人员 18347,10345,14859,18059,12856,11396,14678,10762,17576。

工作人员轮流制午餐(E12)。会场内的工作人员分为两组交替去工作间吃午餐，就餐时间分别为每天 11:40~12:10,12:10~12:40。

3 数据可视分析挑战赛

3.1 挑战赛介绍

ChinaVis 中国可视化与可视分析大会[35]每年汇集国内外数百名可视化研究人员与领域专家。ChinaVis 于 2015 年首次设立数据可视分析挑战赛[33−34]，至今已成功举办 7 届。作为大会的一个重要环节，数据可视分析挑战赛提供一系列精彩的剧本、详实的数据和具体的问题，邀请研究人员、开发人员和爱好者使用一系列可视分析技术和工具完成数据分析任务。数据可视分析挑战赛旨在帮助参赛者评估他们使用的技术和工具在解决复杂问题中的有效性和新颖性，提供锻炼和竞技交流的平台，从而推动可视化与可视分析的专业人才培养，促进相关研究与应用的发展与进步。数据可视分析挑战赛在过去 7 年里迅速发展壮大，吸引了众多师生和企业参加，如今已成为国内可视分析领域的重要竞赛。

内部安全威胁数据集应用在 2018 年的 ChinaVis 数据可视分析挑战赛中，要求参赛者完成以下 3 个任务：a) 描述公司的组织结构；b) 分析公司员工的日常行为模式；c) 通过异常事件总结出有价值的威胁情报，分析事件之间可能存在的关联。本届数据挑战赛共收到了 342 名参赛者提交的 77 份参赛作品。

室内人群移动轨迹数据集应用在 2019 年的 ChinaVis 数据可视分析挑战赛中，要求参赛者完成以下 4 个任务：a) 推测各会场的日程安排；b) 分析会场中的人员类型并总结各类人员的移动规律；c) 找出至少 5 个值得关注的异常事件；d) 总结本次会议在组织和管理方面的不足。本届数据挑战赛收到了 359 名参赛者提交的 75 份作品。

基于参赛者提交的参赛作品，挑战赛组委会邀请了公共安全专家和可视分析专家共同评审参赛作品，以评估参赛作品对分析任务的完成度和准确率。每个参赛作品被随机分配给 4 到 6 名专家，专家依据数据真相和个人经验，从分析质量、视觉设计、交互设计、新颖性和可扩展性 5 个方面评分。作品得分采用 5 分制，5 分为最好，1 分为最差。根据专家对参赛作品的评审意见、参赛者答题情况和专家建议，分析、总结每份参赛作品对异常事件的分析准确率，以达到事件难度评估的目的，从而验证数据集难度设置的有效性、科学性与合理性。

3.2 事件难度

3.2.1 评估方法

本文提出的数据集采用定量化方法评估事件难度。事件难度量化的核心要点主要包括：事件的发生时间、发生地点、涉及的人物或实体对象以及事件描述。对于每个事件的上述 4 个要点，给出 0 到 1 区间的分数，最后计算各要点的平均得分，将其转化为百分数作为该事件的平均准确率，并以此为参考评估事件的难度。本章节分别统计了内部安全威胁数据集和室内人群移动轨迹数据集中各异常事件的平均准确率。

3.2.2 事件难度评估

在内部安全威胁数据集中，根据网络安全专家对数据集中 14 个异常事件在企业中造成安全威胁程度的评估，先将异常事件分为重要异常事件、中等异常事件和普遍异常事件 3 类，然后依次评估了每个异常事件的得分准确率，并将异常事件的识别难度分为困难、中等、简单 3 个程度，分别用 D(Difficult),M(Moderate),S(Simple)表示，如表 2 所示。

14 个异常事件被分为 4 个简单事件，2 个中等难度事件和 8 个困难事件。简单事件由于其经常在企业中发生，且需要分析的数据源较为单一，数据表象比较明显，参赛者可以很容易在对应日志数据中发现这些异常事件。比如，商业间谍盗取账号的过程中(E1)，生成了许多账号登录失败的日志数据记录；辞职事件中(E4,E7,E8)，参赛者只需要分析员工打卡日志数据，就可以发现 3 名提交辞职申请的员工在 2017 年 11 月最后几天没有来上班。中等难度事件的识别需要参赛者非常细心且善于总结规律，比如，数据库故障(E5)本身并不难识别，但参赛者需要在大量电子邮件日志数据中发现数据库系统发出的报警主题的邮件。员工集体活动(E11)中，人力资源部门会在每周四用专用邮件(allstaff@hightech.com)邀请所有员工参加集体体育锻炼，该事件需要参赛者在电子邮件日志数据中识别出这一规律。财务月底加班(E12)需要参赛者在员工打卡日志数据中识别出各部门人员出勤的规律，才能发现财务部门的绝大部分员工经常在月底加班。困难事件的识别主要体现在 3 个方面：a) 相关数据少，较为隐蔽，需要参赛者在正常的数据表象下挖掘出一些潜在的异常威胁行为与模式。如，违规查看产品信息(E2)和恶意泄露产品信息(E3)事件的相关日志数据较少，且数据表征不明显，需要参赛者具有敏锐的洞察力。而且，违规查看产品信息的行为在数据库维护(E6)的掩护下进行，这增大了事件的识别难度，需要参赛者发现各事件背后潜

在的关联。跳板事件(E9)中，部分员工在访问目标服务器时经过了服务器跳转，跳转服务器的隐藏程度非常高，需要参赛者分析员工访问目标服务器时的每一个环节；b) 需要参赛者协同分析多数据源之间的关联。如，旅游事件(E10)中，虽然4名员工连续多日不在公司很容易识别，但要想推测他们要一起旅行，需要对这几位员工的邮件日志数据和网站浏览数据进行深度综合推理，才能发现他们之间的关联。VPN远程访问事件(E13)中，需要参赛者协同分析员工打卡日志数据和TCP流量日志数据之间的关联；c) 需要参赛者具有一定的网络安全专业知识。如，参赛者需要识别出维护人员使用SSH协议连接到故障服务器这一现象，才能推测出数据库维护事件(E6)。流量监控系统故障(E14)需要参赛者了解各网络协议的通用端口号，因此，只有极少数参赛队伍发现了这一异常事件。

表2 内部安全威胁数据集中异常事件的重要程度、识别难度和识别准确率

Table2 The importance, recognition difficulty level and accuracy of abnormal events in ITD-2018

importance	abnormal event	accuracy/%
important events	E1: account stealing (S)	23.0
	E2: product data peeping (D)	7.0
	E3: product data leakage (D)	12.0
	E4: the spy resignation (S)	27.0
moderate events	E5: database failure (M)	14.6
	E6: database maintenance (D)	9.4
	E7: DB deleter resignation (S)	22.0
general events	E8: resignation (S)	21.0
	E9: jump server (D)	5.0
	E10: tourist (D)	3.0
	E11: group activity (D)	5.0
	E12: overtime work (M)	15.9
	E13: VPN remote access event (D)	10.0
	E14: traffic monitor system failure (D)	3.0

在室内人群移动轨迹数据集中，根据领域专家对数据集中各异常事件在室内场馆管理中的影响程度，先将异常事件分为重要异常事件、中等异常事件和普遍异常事件3类，分别用D(Difficult),M(Moderate),S(Simple)表示；然后依次评估了12个异常事件得分的准确率，并总结了各异常事件的识别难度，如表3所示。

表3 室内人群移动轨迹数据集中异常事件的重要程度、识别难度和识别准确率

Table3 Importance, recognition difficulty level and accuracy of abnormal events in ICMTD-2019

importance	abnormal event	accuracy/%
important events	E1: copy of name badge (S)	21.9
	E2: item missing (D)	1.0
	E3: equipment failure (M)	8.7
	E4: personnel ultra vires (D)	1.8
moderate events	E5: packed sub-venues (M)	6.2
	E6: venue congestion (S)	21.0
	E7: book signing (M)	19.2
	E8: group visit (D)	3.3
general events	E9: forgotten badge (D)	1.7
	E10: early exit of hacking contest (D)	4.7
	E11: staff lateness (M)	15.6
	E12: staff lunch turns (D)	2.7

12个异常事件被分为2个简单事件、4个中等难度事件和6个困难事件。简单事件的发现率和准确率较高，其数据特征较为明显，参赛者可以通过观察会场人员的整体移动和分布情况识别异常事件，如：轨迹不连续或重复(E1)，以及大规模人群的聚集(E6)。中等难度事件需参赛者从较小的时空粒度分析会场人员移动和分布情况，如：数据缺失(E3)和小范围人群的聚集(E5,E7)；还需要参赛者正确识别具有独特移动轨迹特征的人员类型，如，部分工作人员的迟到和早退(E11)。困难事件分析难点来自两个方面：a) 需要参赛者正确识别出所有会场人员类型、权限和移动规律，并识别出图2(b)地图中未公开实际用途的VIP休息室、媒体间、黑客大赛区、茶歇点1、茶歇点2和工作间，如：E4需要参赛者正确识别媒体间和VIP休息室，以及对应允许进入的媒体记者和VIP嘉宾，并正确分析这两种类型人员的移动规律与权限。b) 需要参赛者识别出会场人员的突发性变化，并结合会场中的其他异常事件联合探索数据真相，具体包括E2,E9,E10，如：E2需要参赛者识别出会场人员突然频繁在VIP休息室和服务台之间移动的情况，并结合E1事件中人员A的移动轨迹，发现二者曾经在同一位置停留，猜测该会场人员物品被盗，因此频繁去服务台询问。

3.3 主观评估

以调查问卷的形式收集参赛者对数据集场景设计、数据集总体质量、作品评审结果和任务分析难度的满意度，内部安全威胁数据集共收到47人次有效反馈，室内人群移动轨迹数据集共收到55人次有效反馈，主观评估结果如图4所示。

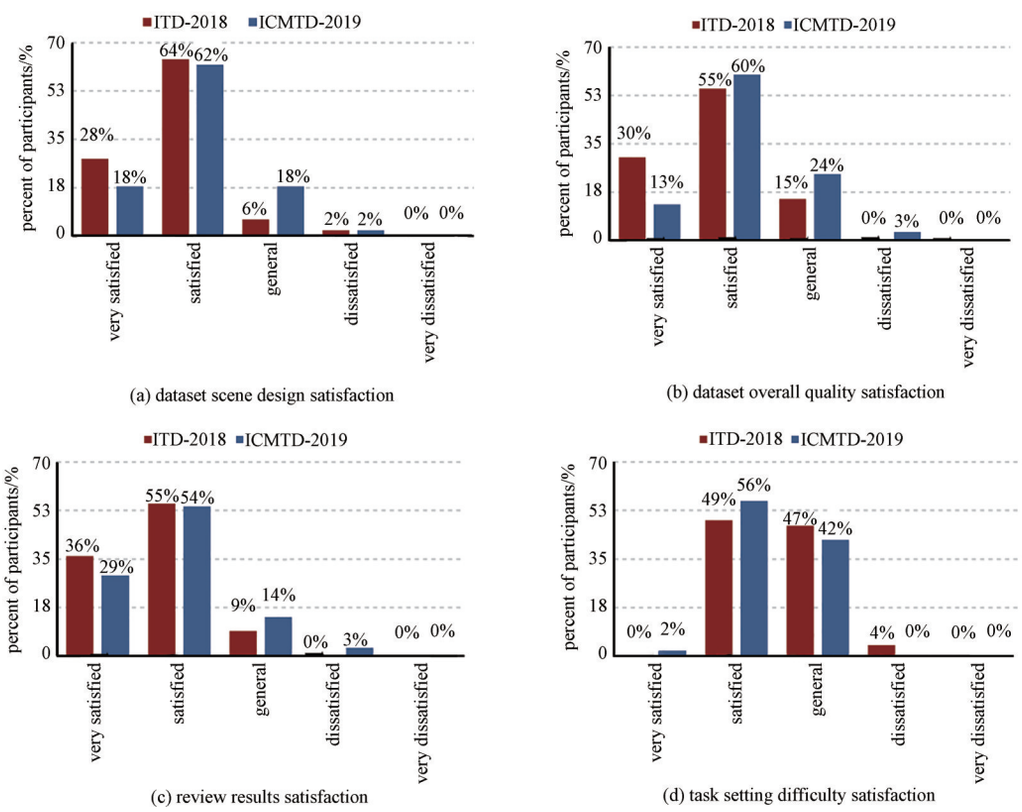


Fig.4 ChinaVis Data Challenge 2018 and 2019 survey results

图4 ChinaVis Data Challenge 2018和2019调查问卷结果

约80%~92%的参赛者对2个数据集的场景设计感到满意(图4(a))，参与者普遍反映数据集的场景设计逼真、生动，背景故事完整，各个实体对象的设计也比较合理，具有较好的现实意义。约73%~85%的参赛者对2个数据集的总体设计感到满意(图4(b))，他们认为数据集规模和时间跨度适中。特别的是，内部安全威胁数据集的参赛者认为，对多源异质数据的综合分析，尤其是从不同类型的数据源中寻找线索和关联的过程是相当有趣和具有挑战性的。室内人群移动轨迹数据集的参赛者认为，数据集中包含了许多有趣的线索和丰富的人群移动时空模式，能激发他们的分析热情。大约83%~91%的参赛者对评审专家给出的评审意见和分数表示满意(图4(c))，这表明我们对采用的整体评估方案和参赛作品评审公平、合理。大约49%~58%的参赛者认为分析任务具有一定挑战性(图4(d))，也能够有效地鉴别他们采用的方法、技术和系统的有效性，可以促进相关研究与技术应用的发展。

4 结论

本文介绍了2个基准数据集，分别是内部安全威胁基准数据集和室内人群移动轨迹基准数据集，以推动相关领域的科学研究与技术应用。本文的工作也存在一些局限性，对于内部安全威胁基准数据集的工作，简化了数据集的场景设计，以便于在技术上提高其可实现性。如，公司部门和层级的划分还可以更丰富，公司的行政总裁数量可以更多，并分管不同的部门。这些简化的场景设计使得内部安全威胁基准数据集的复杂程度低于真实世界的数据。对于室内人群移动轨迹基准数据集的工作，在数据建模时很难完整地构建和模拟人在参加学术会议过程中思想和行为的复杂性和不确定性。另外，数据集中只有移动轨迹一种数据源，无法满足异构数据协同分析任务。

致谢：感谢中国数据可视分析挑战赛(ChinaVis Data Challenge)的所有组织者、评审专家、参赛者、参会者和志愿者。感谢奇安信科技集团雷尔可视化平台部提供技术支持。

参考文献：

[1] GUO H,WANG L,LIANG D,et al. Big earth data from space:a new engine for earth science[J]. Science Bulletin, 2016,61(7):505-513.

[2] 夏勇,丁岐鹃,尤路. 基于大数据的雷达健康管理系统[J]. 太赫兹科学与电子信息学报, 2019,17(4):686-690. (XIA Yong, DING Qijuan,YOU Lu. Radar prognostic and health management technology based on big data[J]. Journal of Terahertz Science and Electronic Information Technology, 2019,17(4):686-690.)

[3] 赵颖,王权,黄叶子,等. 多视图合作的网络流量时序数据可视分析[J]. 软件学报, 2016,27(5):1188-1198. (ZHAO Ying,WANG Quan,HUANG Yezi,et al. Collaborative visual analytics for network traffic time-series data with multiple views[J]. Journal of Software, 2016,27(5):1188-1198.)

[4] CAPPELLI D M,MOORE A P,TRZECIAK R F. The CERT guide to insider threats: how to prevent, detect, and respond to information technology crimes(theft,sabotage,fraud)[M]. Boston,MA,USA:Addison-Wesley Professional, 2012.

[5] 屈正庚,吕鹏. 中小型企业网络安全方案的设计[J]. 太赫兹科学与电子信息学报, 2019,17(1):158-161. (QU Zhenggeng,LYU Peng. Network security scheme design for small and medium-sized enterprises[J]. Journal of Terahertz Science and Electronic Information Technology, 2019,17(1):158-161.)

[6] ZEADALLY S,YU B,JEONG D H,et al. Detecting insider threats: solutions and trends[J]. Information Security Journal:A Global Perspective, 2012,21(4):183-192.

[7] 赵颖,樊晓平,周芳芳,等. 网络安全数据可视化综述[J]. 计算机辅助设计与图形学学报, 2014,26(5):687-697. (ZHAO Ying, FAN Xiaoping,ZHOU Fangfang,et al. A survey on network security data visualization[J]. Journal of Computer-Aided Design and Computer Graphics, 2014,26(5):687-697.)

[8] PARVEEN P,EVANS J,THURAISINGHAM B,et al. Insider threat detection using stream mining and graph mining[C]// 2011 3rd IEEE International Conference on Social Computing. Boston,MA,USA:IEEE, 2011:1102-1110.

[9] LEGG P A. Visualizing the insider threat:challenges and tools for identifying malicious user activity[C]// 2015 IEEE Symposium on Visualization for Cyber Security. Chicago,IL,USA:IEEE, 2015:1-7.

[10] SPITZNER L. Honeypots: catching the insider threat[C]// Proceedings of the 19th Annual Computer Security Applications Conference. Las Vegas,NV,USA:IEEE, 2003:170-179.

[11] Visual analytics benchmark repository[EB/OL]. [2021-06-10]. https://www.cs.umd.edu/hcil/varepository/benchmarks.php.

[12] GRINSTEIN G,SCHOLTZ J,WHITING M,et al. VAST 2009 Challenge:An Insider Threat[C]// Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology. Atlantic City,NJ,USA:IEEE, 2009:243-244.

[13] SCHONLAU M. Masquerading user data[EB/OL]. [2021-06-10]. http://www.schonlau.net.

[14] SCHONLAU M,DUMOUCHEL W,JU W H,et al. Computer intrusion:detecting masquerades[J]. Statistical Science, 2001,16(1):58-74.

[15] CAMIÑA J B,HERNÁNDEZ-GRACIDAS C,MONROY R,et al. The windows-users and-intruder simulations logs dataset(wuil): an experimental framework for masquerade detection mechanisms[J]. Expert Systems with Applications, 2014,41(3):919-930.

[16] YUAN J,ZHENG Y,XIE X,et al. Driving with knowledge from the physical world[C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego,CA,USA:ACM, 2011:316-324.

[17] YUAN J,ZHENG Y,ZHANG C,et al. T-drive: driving directions based on taxi trajectories[C]// Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. San Jose,CA,USA:ACM, 2010:99-108.

[18] PIORKOWSKI M,SARAFIJANOVIC-DJUKIC N,GROSSGLAUSER M. CRAWDAD dataset epfl/mobility[EB/OL]. [2021-06-10]. https://crawdad.org/epfl/mobility/20090224.

[19] PIORKOWSKI M,SARAFIJANOVIC-DJUKIC N,GROSSGLAUSER M. A parsimonious model of mobile partitioned networks with clustering[C]// 2009 1st International Communication Systems and Networks and Workshops. Bangalore,India:IEEE, 2009:1-10.

[20] HERRERA J C,WORK D B,HERRING R,et al. Evaluation of traffic data obtained via gps-enabled mobile phones:the mobile century field experiment[J]. Transportation Research Part C, 2010,18(4):568-583.

[21] CHAN A B,LIANG Z,VASCONCELOS N. Privacy preserving crowd monitoring: counting people without people models or tracking[C]// 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage,AK,USA:IEEE, 2008:1-7.

[22] IEEE Vast challenge 2016 homepage[EB/OL]. [2021-06-10]. http://www.vacommunity.org/VAST+Challenge+2016.

[23] IEEE Vast challenge 2016 benchmark[EB/OL]. [2021-06-10]. http://www.cs.umd.edu/hcil/varepository.

[24] LIN Y, ZHAO H, MA X, et al. Adversarial attacks in modulation recognition with convolutional neural networks[J]. IEEE Transactions on Reliability, 2021,70(1):389-401.

[25] WANG P,GAO F,ZHAO Y,et al. Detection of indoor high-density crowds via WiFi tracking data[J]. Sensors, 2020,20(18):5078-1-15.

[26] 陈思翰. 基于Fang算法的TDOA室内定位技术[J]. 太赫兹科学与电子信息学报, 2017,15(5):752-755. (CHEN Sihan. TDOA indoor location technology based on Fang algorithm[J]. Journal of Terahertz Science and Electronic Information Technology, 2017,15(5):752-755.)

[27] HAN D,JUNG S,LEE M,et al. Building a practical WiFi based indoor navigation system[J]. IEEE Pervasive Computing, 2014,13(2):72-79.

[28] MEHRAN R,OYAMA A,SHAH M. Abnormal crowd behavior detection using social force model[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami,Florida,USA:IEEE, 2009:935-942.

[29] ITD-2018 dataset homepage[EB/OL]. [2021-06-10]. https://github.com/csuvis/InsiderThreatData.

[30] ZHAO Y, YANG K, CHEN S, et al. A benchmark for visual analysis of insider threat detection[J]. Science China-Information Sciences, 2022,65(9):199102-1-4.

[31] ICMTD-2019 dataset homepage[EB/OL]. [2021-06-10]. http://github.com/csuvis/IndoorTrajectoryData.

[32] ZHAO Y,ZHAO X,CHEN S,et al. An indoor crowd movement trajectory benchmark dataset[J]. IEEE Transactions on Reliability, 2021,70(4):1368-1380.

[33] ChinaVis data challenge homepage[EB/OL]. [2021-06-10]. http://www.chinavis.org/2019/english/challenge_en.html.

[34] 赵颖,张卓,袁晓如. 数据可视分析挑战赛三年回顾[J]. 网络与信息安全学报, 2018,4(2):55-61. (ZHAO Ying,ZHANG Zhuo, YUAN Xiaoru. ChinaVis data challenge from 2015 to 2017[J]. Chinese Journal of Network and Information Security, 2018,4(2):55-61.)

[35] ChinaVis homepage[EB/OL]. [2021-06-10]. http://www.chinavis.org.

作者简介：

赵　颖(1980-)，男，博士，教授，主要研究方向为可视化与可视分析.email:zhaoying@csu.edu.cn.

杨　奎(1996-)，男，硕士，主要研究方向为可视化与可视分析.

张　卓(1988-)，男，学士，高级工程师，主要研究方向为高级威胁检测、大数据与网络安全.

赵　鑫(1997-)，女，硕士，主要研究方向为可视化与可视分析.

陈思明(1989-)，男，博士，副教授，主要研究方向为可视化与可视分析，以人为中心的人工智能.

黄　鑫(1992-)，男，学士，高级工程师，主要研究方向为安全可视化与可视分析.