基于大数据的 APT 检测技术应用研究

黄亭

西安翻译学院，陕西 西安 710105

DOI: 10.61369/TACS.2026090051

摘　　要： 网络环境越来越复杂，叠加 AI 工具在网络攻击的应用，传统的 APT 检测手段已经无法满足安全需求。本文从 APT 检测技术、安全运营、数据治理三个核心维度，系统梳理了当前 APT 溯源检测领域的共性痛点，如告警日志爆炸、安全运营效率低、数据标准不统一等。然后本文基于大数据技术，提出了构建 APT 立体防护体系，通过数据建模、构建威胁情报知识库和数据治理体系，建立全域、高效、实时的安全检测体系，全方位提高安全分析的实时性、可靠性。

关 键 词： APT；大数据；数据治理

Application Research of APT Detection Technology Based on Big Data

Huang Ting

Xi'an Fanyi University, Xi'an, Shaanxi 710105

Abstract: The increasingly complex network environment, coupled with the application of AI tools in cyber attacks, has rendered traditional APT detection methods unable to meet current security requirements. This paper systematically sorts out the common pain points in the field of APT traceability and detection from three core dimensions: APT detection technology, security operations, and data governance, such as the explosion of alert logs, low security operation efficiency, and inconsistent data standards. Then, based on big data technology, this paper proposes the construction of a three-dimensional APT protection system. Through data modeling, the establishment of a threat intelligence knowledge base and a data governance system, it builds a comprehensive, efficient, and real-time security detection system to fully improve the real-time performance and reliability of security analysis.

Keywords: APT; Big Data; Data Governance

引言

随着生成式 AI 工具的日益普及，网络攻击软件工具的开发成本越来越低。与此同时，网络环境越来越复杂，攻击暴露点越来越多，而传统的单点安全检测防控系统已经无法满足当前安全需求，特别是 APT（高级持续性威胁）攻击具有隐蔽性强、持续时间长和组织性高等特点，使得网络安全问题愈发严峻。在此背景下，网络攻击追踪溯源成为保障网络安全的关键议题 [1]。一方面，可以系统性分析网络攻击追踪溯源面向核心问题，从技术、基础设施、运营、数据管理等角度全面分析 APT 溯源检测面临的挑战。另一方面，本文从大数据技术体系出发，可以结合前沿技术发展趋势，探索建立立体的全方面的检测防护体系。

一、AI 工具普及背景下 APT 溯源检测的主要挑战与问题

（一）技术层面的面临的主要挑战

智能制造发展迅猛，工业控制系统与传统网络融合深入，移动网及物联网规模越来越大，云原生技术虚拟主机引入更多安全新特征，网络环境日益复杂，暴露点越来越多，技术层面面临越来越多的全新挑战。

第一，安全日志数据规模给现有技术带来巨大冲击。裴嘉睿等提出了基于多尺度图神经网络和双向门控循环单元 - 时间注意力机制的 APT 攻击检测模型，此模型结合溯源图结构、时序演化和攻击归因的联合建模框架，能够刻画攻击链关键阶段及长期演化模式，此方法显著提升了 APT 攻击的检测效果 [2]。王成瑞提出了一种基于告警日志关联分析技术，基于关键威胁与告警关联重构 APT 攻击的实现路径，能够有效压缩日志规模，帮助安全操作人员提升分析效率 [3]。尽管 APT 检测技术融入深度学习后，检测效果得到巨大提升，但是主机日志、网络流量的数据规模越来越大，导致处理性能难以满足实时性要求。尤其是 ICS 网络的攻击日益增多，大量的设备日志不间断产生，数据量庞大，攻击相关的关键信息高度稀疏隐藏在海量正常日志中，APT 攻击通常会持续 1-4 周的低频攻击，导致检测算法很难在较大尺度上保持检测效果。

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

第二，攻击相关数据的种类、粒度的变化导致原检测机制失效。叶华欣基于系统日志、网络流量日志、审计日志、其他数据源（情报、安全事件），提出了一种基于图神经网络的 APT 攻击链溯源方法，将网络实体集合定义为节点构建攻击图，并构建邻接矩阵，能够深入挖掘攻击行为的传播路径，以图的方式描述 APT 阶段性的攻击行为，可以有效降低误报，但该方法的适用边界仅覆盖传统企业网络场景，未针对工控、移动网络等做适配优化 [4]。卢钱长针对 Docker 容器逃逸，重新定义异构溯源图，增加了容器标签，扩充溯源图定义，能够将主机层次的溯源方法应用在容器场景 [5]。然而在移动网络、ICS 网络同样会导致原有 APT 检测方法失效。

第三，防御效率低于攻击效率。在 AI 工具加持下，攻击工具开发速率大幅提高、攻击的组织时间越来越短，而安全分析师日常淹没在大量的告警日志处理上，要求快速、精准发现关键攻击，过滤误报，对安全分析师而言难度持续提升。

（二）安全运营服务自动化率低

对于攻防双方，效率决定着主动权，实战中，由于攻击方属于隐藏的一方，通常掌握着主动权，防守方通常被动分析攻击，如没有高效的工具辅助，难以在第一时间锁定高危的 APT 攻击。

第一，传统的基于关键攻击特征的检测方法失效。传统高危攻击行为自动检测方法，通常依赖 MD5 哈希值等静态攻击特征来匹配溯源背后的攻击组织，以此判定当前攻击行为属于 APT 攻击，但是在攻击的早期识别中，很难发现这些确定信息。

第二，威胁情报与攻击算法描述模式分离，难以自动化。在实际安全运营中，如果安全检测系统能够将攻击行为溯源，关联出有效的威胁情报，较好解释攻击行为，则能够大大提高安全运维效率。杨秀璋构建了一个“检测 - 溯源 - 推理 - 解释”的高隐蔽攻击整体防御框架，提出了 APT 攻击溯源及推理方法、APT 知识图谱构建与攻击行为画像生成方法，将 ATT&CK 情报信息融入算法，提高了算法可解释性 [6]。然而，威胁情报是安全专家的文本报告，和攻击溯源图无法直接进行关联，这极大降低了高危 APT 攻击的分析识别效率。

第三，安全检测系统可解释性低。何恬提出一种基于攻击链构建的物联网攻击溯源 ATT&CK-OAHP-BNS 方法，能够构建基于攻击行为的攻击链，并且包括完整的攻击行为及其概率，但是缺乏证据辅助判断检测结果正确性 [7]。攻击分析环节，若系统自动展示攻击步骤、清晰推导攻击目的，将大幅提高安全运营效率。

（三）安全数据治理体系落后

异构图级别的 APT 攻击阶段识别已经成为 APT 检测主流技术，何厚翰在构建异构溯源图时，节点代表主机中的系统实体，涵盖进程、文件、Socket 连接等不同主体类型，边代表主机中的系统事件，包括文件读写、进程调度、网络通信等多种操作行为 [8]。同时有效利用威胁情报、安全事件、审计日志、资产信息，更有助于判断 APT 攻击的可信度。

第一，不同类别数据的描述语言模型不一致。例如，威胁情报通常由安全专家发布，包含攻击的关键特征和步骤，而溯源图刻画的是底层的攻击行为，二者无法有效关联。

第二，进程、网络、文件操作、安全事件等信息构建存在断点。例如，异常进程会引起告警，攻击行为会产生这两类信息，存在重复，需要剔重；网络日志大多为数据包粒度，而进程行为日志对应的是网络流粒度，二者天然存在粒度差，进一步关联容易导致数据不一致，错失关键信息。

第三，不同产品的数据难以对齐。通常安全设备的来源不止一个厂家，厂家之间的数据标准不一致，难以对齐。

第四，安全数据治理意识薄弱。数据是安全的基础，目前很多安全从业者普遍存在“重技术，轻数据体系建设”的认知偏差，而底层数据治理的缺失，恰恰是攻击分析效率难以提升的核心根源之一。

（四）难以对 0-Day 攻击做出快速响应

钱锦等设计并实现了一个融合大语言模型语义解析能力与攻击描述语言（ADL）的自动化攻击链复现系统，可在动态部署的靶场中完成攻击链的执行闭环 [9]。在 AI 工具加持下，攻击者可以快速生成攻击软件或脚本，低成本组织 0-Day 攻击。高频出现的 0-Day 攻击将对现有网络安全防御体系造成极大冲击。

二、基于大数据 APT 检测溯源的防护体系建设策略

（一）构建基于大数据的全域、高效、实时的安全检测体系

第一，接入安全全域数据，构建安全数据湖。通过 ETL 技术将企业网、移动网、ICS 的系统日志、网络流量日志、审计日志、以及威胁情报、安全事件等多源异构数据接入数据湖，进一步构建基于安全的数据仓库，为安全分析打下坚实基础。

第二，建立安全数据的元数据，构建面向 APT 检测的数据仓库。针对数据粒度不统一、类型异构、关联链路断裂等问题，建议企业统一的元数据管理体系，设计元模型，持续维护，保证数据之间通过清晰的血缘关系能够关联一块；根据 APT 检测全部攻击阶段识别的数据需求，构建数据仓库，满足后续的准实时分析。

第三，对数据进行“行为 -> 事件 -> 攻击模式”三个层次汇聚。行为层对应攻击溯源图，数据来源包括系统日志、网络流量日志，以图的方式刻画进程级别的序列攻击行为；事件层面向全量日志进行大规模压缩，输出实例化的攻击步骤，进一步通过溯源图把审计日志、事件告警日志关联去重，形成一个攻击图；攻击模式层是攻击知识库，来源于威胁情报或安全分析专家，是对攻击模式的高度抽象，攻击模式与攻击图的映射依赖安全专家或者 AI 技术，详情见本节第二部分。基于三层的大数据检测体系，将检测、求证融为一体，大幅提高安全分析师分析效率。

第四，建立以预防为主的态势感知系统。将安全检测的重点前移，通过大数据平台，建立网内资产安全基线，实时检测异常的行为，根据检测结果对异常资产优先进行攻击溯源分析，从而提高检测的实时性。

（二）建设威胁情报攻击图知识库，提高自动识别的可靠度

第一，建设威胁情报攻击图知识库。基于威胁情报，安全专

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

家构建攻击图的知识库，知识库内定义的攻击类型、攻击实体类型，需与前文“事件层”的节点类型、边类型完全对齐。APT 攻击溯源分析时，系统通过图嵌入技术对事件攻击图做特征编码，生成对应的图向量表征，将其与知识库中攻击图做相似度匹配，即可输出对应的攻击模式。安全分析师不仅能看到溯源图，并且能够看到相近的攻击模式，大大提高分析的可靠性和效率。

第二，以“事件层”攻击图的形式实时展示攻击进度。系统在攻击溯源检测过程中，通过事件驱动的方式，实时关联攻击图知识库，一旦匹配度超过阈值，系统触发告警。攻击图形式一方面提高了检测结果的可解释性，另一方面，可以聚集多个攻击事件的信息，有利于更早发现 APT 攻击。

（三）构建网络安全的领域数据治理体系

全面梳理安全检测业务需求，搭建可持续迭代的网络安全领域数据治理体系。

第一，全域异构数据治理。搭建覆盖企业网、移动网、工控场景的安全数据湖，通过 ETL 技术统一接入系统日志、流量日志、威胁情报、安全事件、审计日志、资产信息等多源数据，制定跨厂商统一的元数据标准，打通不同安全设备的数据壁垒。

第二，引入 NLP 技术辅助完成数据汇聚建模。按“行为层 - 事件层 - 攻击模式层”完成数据分层治理，引入 NLP 技术，以实体智能识别分析为主，人工校正为辅，自动对齐三层数据之间的节点、边属性定义，打通全链路数据的关联通道。

第三，建立长效的数据治理机制。搭建带血缘关系的安全数据仓库，强化全员数据治理意识，将数据治理能力建设融入日常工作流程，逐步提升数据质量，促进从数据接入、关联分析到攻击溯源的自动化闭环。

（四）构建基于 AI 的安全分析智能体

尽管攻击的具体实现手段迭代速度极快，但底层攻击模式的演化节奏十分平缓，核心原因是 APT 攻击组织的战术体系更新周期较长。因此，基于安全分析专家分析经验，搭建安全分析智能体 skill，复用专家经验，实现 0-Day 攻击的自动化分析。

三、结束语

综上所述，网络环境的日益复杂，传统特征匹配类检测手段已难以适配当前攻防态势；叠加 AI 工具的使用，网络攻击门槛持续降低，攻击手段更新越来越频繁，因此对高效、可靠的安全分析系统需求越来越紧迫。本文从技术层面、安全运营、数据治理等维度，梳理了 APT 溯源检测存在的问题，如告警日志爆炸、安全运营自动化率低、数据标准不一致等。为了应对相关挑战，本文引入大数据技术与思想，构建 APT 立体防护体系：对网络安全相关数据进行数据建模，建立全域、高效、实时的安全检测体系，同时构建威胁情报攻击图知识库和网络安全领域数据治理体系，融入了最新 AI 智能体技术。全方位提高安全分析的实时性、可靠性。

参考文献

[1] 杨刚，董航，辛冉 . 网络攻击追踪溯源技术研究综述 [J/OL]. 智能计算机与应用 .https://doi.org/10.20169/j.issn.2095-2163.26012103

[2] 裴嘉睿，高见. 基于 MS-GNN 和 BiGRU-TAM 的溯源图 APT 攻击检测模型 [J/OL]. 计算机科学 .https://link.cnki.net/urlid/50.1075.tp.20260630.1833.008

[3] 王成瑞 . 基于告警日志关联分析的攻击场景重构技术研究 [D/OL]. 北京：北京邮电大学 ,2025.https://doi.org/10.26969/d.cnki.gbydu.2025.002807

[4] 叶华欣 . 基于图神经网络的 APT 攻击链智能溯源方法 [J/OL]. 电脑编程技巧与维护 .2026(02)https://doi.org/10.16184/j.cnki.comprg.2026.02.046

[5] 卢钱长 . 基于异构溯源图的 Docker 容器逃逸检测与攻击还原方法 [D/OL] 南京：东南大学 ,2024.https://doi.org/10.27014/d.cnki.gdnau.2023.001635

[6] 杨秀璋 . 面向高隐蔽攻击的行为检测关键技术研究 [D/OL]. 武汉：武汉大学 ,2024.https://doi.org/10.27379/d.cnki.gwhdu.2023.000605

[7] 何恬 . 基于攻击行为的网络攻击溯源关键技术研究 [D/OL]. 无锡：江南大学 ,2024.https://doi.org/10.27169/d.cnki.gwqgu.2023.000803

[8] 何厚翰 . 异构溯源图在 APT 攻击检测中的应用研究 [D/OL]. 北京：中国人民公安大学 ,2025.https://doi.org/10.27634/d.cnki.gzrgu.2025.000313

[9] 钱锦，谭诗雨，徐李冰，罗俊，李振源. 大模型驱动的攻击过程分析、描述与复现方法 [J/OL]. 微电子学与计算机 .https://link.cnki.net/urlid/61.1123.TN.20260319.1132.008

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License