基于层次化掩码图自编码框架的APT威胁检测方法

刘尚东[1,2,3,4],杨易润[1],王一诺[1],杜宏煜[1],汪文博[1],季一木[1,2,3,4*]

(1. 南京邮电大学计算机学院,江苏南京 210023;2. 南京邮电大学高性能计算与大数据处理研究所,江苏南京 210023;3. 国家高性能计算中心南京分中心,江苏南京 210023;4. 江苏省高性能计算与智能处理工程研究中心,江苏南京 210023)

摘 要: 高级持续性威胁(Advanced Persistent Threats, APTs)凭借其高度隐蔽性、长周期性及多阶段攻击的特性,已成为当前网络安全防御体系面临的最严峻挑战之一。尽管基于主机日志的溯源图分析技术能够将孤立的系统事件关联为细粒度的行为审计路径,为威胁检测提供了结构化支撑,但现有研究仍面临核心瓶颈:在复杂的系统环境中,攻击者往往通过低频操作来模拟良性行为,导致传统的基于特征码或静态规则的检测方案在应对零日攻击(Zero-day)时极易失效。针对上述挑战,本文提出一种层次化感知的图掩码自动编码器APT威胁检测框架。本框架的核心创新在于引入了层次化拓扑知识来指导掩码过程,而非采用盲目的随机遮蔽。具体而言,模型集成了全局感知遮蔽、局部感知遮蔽与元素感知遮蔽三种策略:全局感知遮蔽旨在保留溯源图的宏观结构稳定性,局部感知遮蔽侧重于刻画实体间的邻域交互逻辑,而元素感知遮蔽则关注实体属性的细粒度特征。这种层次化设计能够在预训练阶段有效过滤非结构性的系统噪声,同时最大程度地保留关键的因果逻辑链条。特别地,节点级一致性约束通过在原子尺度上建模,有效规避了传统图表征学习中全局聚合带来的信号稀释风险。这确保了即便在极端不平衡的样本分布下,微弱的攻击信号仍能通过损失函数获得充分的梯度响应,从而在数学逻辑上保障了训练目标与单点异常检测任务的一致性。在检测阶段,框架采用无监督异常检测算法,基于实体类型的嵌入分布量化节点异常分数,从而精准识别破坏局部因果链的恶意行为。本文在StreamSpot、Unicorn Wget以及DARPA E3等多个公开权威数据集上进行了全面评估。实验结果表明,该框架在平均精确率上达到了98.49%,F1分数达到98.97%。相比于现有基准模型,本文方法在极低攻击基本比率场景下表现出更强的鲁棒性与召回能力,能够有效识别APT攻击全生命周期中的微弱异常信号。

关键词: 高级持续性威胁;异常检测;图掩码自编码器;图神经网络;自监督学习;系统溯源图

基金项目: 国家重点研发计划课题(No.2023YFB2904000,No.2023YFB2904004);江苏省重点研发计划课题(No.BE2023004-2)

中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112(XXXX)XX-0001-17

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20251261

Advanced Persistent Threat Detection Via Hierarchical Masking Graph Autoencoder

LIU Shangdong[1,2,3,4], YANG Yirun[1], WANG Yinuo[1], DU Hongyu[1], WANG Wenbo[1], JI Yimu[1,2,3,4*]

(1. *School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210023, *China*;
2. *Institute of High-Performance Computing and Bigdata, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210023, *China*; 3. *Nanjing Center of HPC China, Nanjing, Jiangsu* 210023, *China*;
4. *Jiangsu HPC and Intelligent Processing Engineer Research Center, Nanjing, Jiangsu* 210023, *China*)

Abstract: Advanced Persistent Threats (APTs) have emerged as one of the most severe challenges to modern cybersecurity defense systems due to their extreme stealthiness, prolonged duration, and multi-stage nature. Although provenance-based analysis of host logs provides structured support for threat detection by correlating isolated system events into granular behavioral auditing paths, existing research still faces a core bottleneck: in complex system environments, attackers often disguise malicious activities as benign behavior through low-frequency operations, rendering traditional detection schemes based on signatures or static rules highly susceptible to failure when encountering zero-day attacks. To address these challenges, this paper proposes a Hierarchical-Aware Graph Masked Autoencoder framework for APT detection. The primary innovation of this framework lies in the introduction of hierarchical topological knowledge to guide the masking process, fundamentally overcoming the limitations of blind random masking. Specifically, the model integrates three targeted strategies: Global-Aware Masking (GAM), Local-Aware Masking (LAM), and Element-Aware Masking (EAM). GAM aims to preserve the macro-structural stability of the provenance graph; LAM focuses on characterizing neighborhood interaction

收稿日期:2026-03-29;录用日期:2026-05-05;责任编委:赵瑞杰

*通信作者:季一木

logic between entities; and EAM addresses fine-grained entity attributes. This hierarchical design effectively filters out non-structural system noise during the pre-training phase while maximizing the retention of critical causal logic chains. Notably, the node-level consistency constraint models at an atomic scale, effectively circumventing the risk of signal dilution caused by global aggregation in traditional graph representation learning. This ensures that even under extremely imbalanced sample distributions, faint attack signals can still obtain sufficient gradient responses through the loss function, thereby mathematically guaranteeing the logical alignment between training objectives and point-wise anomaly detection tasks. During the detection phase, the framework employs an unsupervised anomaly detection algorithm to quantify node anomaly scores based on the embedding distributions of entity types, enabling the precise identification of malicious behaviors that disrupt local causal chains. Comprehensive evaluations were conducted on multiple authoritative public datasets, including StreamSpot, Unicorn Wget, and DARPA E3. Experimental results demonstrate that the proposed framework achieves an average precision of 98.49% and an F1-score of 98.97%. Compared to state-of-the-art baselines, our method exhibits superior robustness and recall in scenarios with extremely low attack base rates, effectively identifying subtle anomalous signals throughout the entire APT lifecycle.

Keywords: advanced persistent threats; anomaly detection; graph masked autoencoder; graph neural network; Self-Supervised Learning; System Provenance Graph

Foundation Item(s): National Key Research and Development Program of China (No. 2023YFB2904000, No.2023YFB2904004); Jiangsu Provincial Key Research and Development Program (No.BE2023004-2)

0 引言

高级持续性威胁(Advanced Persistent Threats, APTs)因其高度隐藏、复杂多变且持续时间长的攻击特性,给全球网络安全构成了严峻挑战[1-2],此类APT攻击通常由组织严密、资源充足的攻击者发起,常常利用零日漏洞或复杂的攻击链,使得传统的入侵检测系统难以有效应对[3]。因此研究和开发能够准确、高效检测此类高级威胁的防御手段已成为网络安全领域亟待解决的关键问题。

在此背景下,从操作系统审计日志中提取的溯源数据因其能够详细刻画系统实体间的因果依赖关系和信息流动路径,为理解和检测复杂的攻击行为提供了强大的数据支撑[4-5]。由这些数据构建成的系统溯源图不仅记录了离散的系统事件,而且揭示了这些事件之间的上下文关联,使得攻击行为即使跨越了较长时间尺度或涉及多个系统组件时,也能在图中有所体现[6]。

尝试利用溯源图进行APT检测的早期研究主要依赖于专家知识构建匹配规则或依赖图模式[7-9]。例如通过将底层审计事件映射到ATT&CK框架中的战术、技术和过程(Tactics, Techniques, Procedures, TTPs)[9],并关联这些TTPs之间的信息流来构建高层攻击场景图。这类方法的可解释性强,但面临两大挑战:一是规则的制定和维护需要深厚的安全知识,成本高昂;二是对于采用新颖TTPs或未知攻击模式的APT攻击难以有效识别,泛化能力有限。

为了降低对先验知识的依赖,一些研究转向基于统计特征的方法[10-11]。这些方法通常分析溯源图中节点的度数、边的频率、路径的罕见程度等可量化的图属性,以此来识别与正常模式偏离的行为。例如,通过识别图中罕见的事件序列[10],或计算节点或边的出现频率来量化其可疑度[11]。虽然这类方法在一定程度上能够发现异常,但它们的主要缺陷在于过度简化了复杂的系统行为,往往忽略了节点交互背后丰富的上下文语义信息。仅凭统计上的罕见不足以区分真正的恶意活动和良性的、偶然发生的系统操作,因此容易产生较高的误报率,增加了安全运营的负担。

随着深度学习技术,特别是图神经网络(Graph Neural Network, GNN)在捕捉图数据结构复杂关系方面展现出的卓越能力,基于GNN的溯源图分析方法已成为一个富有前景的研究方向[12]。GNN能够自动化地学习节点和图结构的嵌入表示,自动捕捉复杂的上下文信息和高阶依赖关系,这在理论上克服了传统统计方法忽略语义的缺点,显著提升了行为建模的精度和表达能力[5,12]。研究者们已探索将多种GNN架构应用于溯源图分析,并在APT检测任务上展现了有效性与潜力[13-17]。

在此背景下,Zian Jia等提出MAGIC[4],该方法在模型学习中采用掩码重构策略,旨在无标签数据上学习通用的、富有意义的系统行为表示被视为一种有前景的方案[18]。然而,MAGIC所采用的随机掩码的策略虽然简单通用,但也在两个重要方面限制了其效果:一是可能无意中破坏了溯源图中对于理解系统行为至关重要的因果依赖链或关键的局部/全局拓扑结构;二是可能产生了过于简单的重构任务,使得模型无需学习深层次的语义信息即可完成任务。这种学习效率和表示质量的不足,限制了模型在下游APT检

测任务中的准确性和鲁棒性。

为解决上述问题,我们提出了一种基于层次化感知的APT检测框架,该框架遵循从原始数据处理到最终异常判定的标准流程:首先从主机审计日志中提取系统事件,并构建动态溯源图表示系统行为;针对现有掩码重构策略中随机掩码可能破坏关键拓扑信息、降低学习效率的问题,我们引入了层次化拓扑知识指导的掩码机制[19]。该机制通过全局、局部和元素三个层面的感知,智能选择掩码对象,引导GNN自编码器学习更具判别力的节点表示[20];在特征重建与结构重建的监督下,我们进一步结合节点级上下文一致性约束,通过最大化同一节点在不同掩码视图下的表示一致性,增强模型对局部一致表示的鲁棒性。在学习到高质量的节点嵌入后,进入异常检测阶段,根据每个节点的类型及局部密度动态调整近邻数量的自适应异常检测(k-Nearest Neighbors,k-NN)算法。

本文在公开的StreamSpot、Unicorn Wget、DARPA E3多个基准数据集[21-24]上对所提出的方法进行了广泛的实验评估。实验结果表明,与最先进的SOTA模型相比[4,17,25-26],我们的方法在保持计算效率的同时,降低了误报率,并在多个评价指标上取得了较优的性能。综上所述,本文的主要贡献包括:(1)首次将层次化拓扑知识掩码策略应用于溯源图APT检测的自监督表示学习中,提升了模型的特征提取能力;(2)在图表示学习阶段利用联合重构的监督机制生成信息更丰富、判别力更强的系统实体嵌入表示,并采取自适应k值k-NN算法进行异常检测,增强了异常检测环节对数据密度变化的适应性;(3)通过在多个公开数据集上的实验验证了所提方法的有效性和实用性,为APT检测框架提供了新的思路和技术途径。

1 相关工作

近年来,将主机日志构建溯源图进行高级持续性威胁检测已成为网络安全领域的研究热点,并涌现出多种技术路线[27]。根据其核心检测原理和技术特点,现有基于溯源图的APT检测方法大致可以分为基于规则、基于统计和基于学习三大类。

基于规则的检测方法是最早应用于溯源图APT分析的技术之一,其核心思想是依赖安全专家的先验知识,将已知的攻击行为模式预先定义为一系列规则、签名或图查询模板,然后在溯源图中进行匹配以发现潜在威胁[7-9]。SLEUTH[7]采用基于标签的方法,通过预定义标签策略来追踪信息流并识别可疑行为。Poirot[8]尝试将已知的攻击行为描述与内核审计记录进行对齐,以发现威胁,但其有效性依赖于现有威胁情报报告的准确性和完整性,对于行为变种或未知攻击则难以辨别。HOLMES[9]引入了一个基于ATT&CK TTPs的中间层,将底层审计事件映射为TTPs实例,并通过追踪实例间的信息流依赖关系构建高层场景图以判断攻击链的存在,但其效果受限于已知TTPs的覆盖范围。综上所述,基于规则的方法凭借其明确的匹配逻辑,具有天然的强可解释性,检测到的威胁可以直接关联到具体的规则和攻击模式,并且通常能保持相对较低的误报率。然而,这类方法共同面临的核心局限性在于其对先验知识的强依赖,这不仅使其难以有效识别利用零日漏洞或采用未知TTPs组合的新型APT攻击,导致泛化能力严重不足,而且随着攻击手法的不断演变,规则库的构建和维护成本变得异常高昂,需要持续的安全专家投入才能跟上威胁态势的变化,限制了其在快速变化的威胁环境中的适应性。

为了缓解对显式攻击规则的依赖性,研究者们提出了基于统计特征的检测方法[10-11]。此类方法旨在通过量化分析溯源图的结构属性来识别与正常模式显著偏离的行为,其基本假设是恶意活动会在统计层面上呈现异常。具体而言,这些方法关注的统计特征包括节点的度数、特定类型边的频率、信息流路径的长度或罕见程度,以及图中特定子结构的出现次数等。PrioTracker[10]通过计算事件的罕见度评估其可疑程度,但单纯的统计罕见度并不能直接等同于恶意性。NoDoze[11]则利用事件频率量化可疑度以辅助告警分诊,但也可能因此忽略掉一些新出现的、但暂时看起来频率不低的恶意活动。虽然基于统计特征的方法在一定程度上降低了对具体攻击模式先验知识的需求,能够发现一些偏离常规统计规律的异常行为,但其共同的缺陷在于它们往往过度简化了复杂的系统交互。通过倾向于关注孤立的、低维度的图属性,它们容易忽略节点和事件背后丰富的上下文语义信息。并且,统计意义上的“异常”或“罕见”并不直接等同于恶意行为,良性的系统操作或系统噪声也可能造成统计偏差。这种难以准确区分良性罕见事件和真正恶意活动的能力不足,导致该类方法通常伴随着较高的误报率,增加了安全运营的实际负担,限制了其在复杂环境中的应用效果。

基于学习的检测策略突破了传统统计模型在语义理解上的瓶颈,其中图神经网络在处理溯源图数据方面展现了卓越的能力[5,12-17],其能够自主地从复杂的图结构中发掘深层表示,并精确地建模实体间的上下文交互与高阶依赖关系。ThreaTrace[17]是早期探索者之一,它利用GraphSAGE[28]的归纳学习能力,通过半监督节点分类任务这一代理任务来识别行为异常的节点。FLASH[13]着重解决了GNN在大规模流式

数据上的效率瓶颈,通过嵌入重用机制大幅提升了检测速度,但其效率增益依赖于节点行为模式相对稳定的假设。KAIROS[14]进一步将检测粒度细化到事件级别,利用时序图网络(Temporal Graph Network,TGN)捕捉动态变化并结合社区发现进行攻击溯源,但事件级别的分析和复杂的模型使其计算开销相对较大。APT-KGL[15]尝试通过构建异构图来区分实体和关系类型,并融合外部威胁知识库以缓解训练数据稀疏问题,其效果仍依赖外部知识的质量且对完全未知的TTPs可能泛化不足。ShadeWatcher[16]创新性地借鉴推荐系统思想,预测实体间"异常"的交互偏好,提供了一种新颖的视角。MAGIC[4]通过训练GNN自编码器来重构被随机掩码部分遮盖的溯源图,迫使模型学习图中固有的结构和属性模式。但过于简单的遮挡也可能导致模型仅需浅层信息即可复原,未能充分驱动模型学习深层次的语义关联。这种表示学习环节的潜在信息损失或学习不足,会直接影响模型理解复杂正常行为模式的深度,从而制约其在后续APT检测任务中的表现。针对此问题,本研究借鉴并引入了基于层次化拓扑知识的掩码策略,期望通过一种更具结构感知能力的方式选择掩码目标,引导自监督学习过程更有效地捕捉溯源图的关键特征,最终提升表示质量与检测性能。

2 概述

2.1 动机实例

图1展示了一个源于真实世界APT攻击的简化溯源图,它描绘了一次APT组织通过污染NPM包发起供应链投毒作为入侵起点到窃取关键信息回传远程服务器的完整攻击过程。入侵始于被攻击用户执行npm install命令,该命令在解析package.json文件后,触发node进程执行依赖安装。由于npm本身允许包作者在安装阶段执行任意代码,于是此node进程随即下载并执行了攻击者事先设置好的载荷malicious.js,该javascript代码下载了第二阶段载荷并立即通过bash执行了它。这个被启动的bash进程成为攻击者的主要代理,它首先在本地搜集凭证,读取了.ssh/id_rsa等用户私有的身份凭证,接着打包创建一个恶意的zip文件data.zip隐藏数据。最后bash进程调用curl将zip文件泄露到攻击者控制的远程服务器,后续可能利用窃取到的身份凭证伪装成该用户身份入侵用户拥有的其他计算资产。

2.2 威胁模型

为了清晰地界定本框架的检测范围和能力,我们建立在以下威胁模型和信任假设之上。与该领域中多数基于溯源图的检测工作[13-14,29]相似,我们的框架

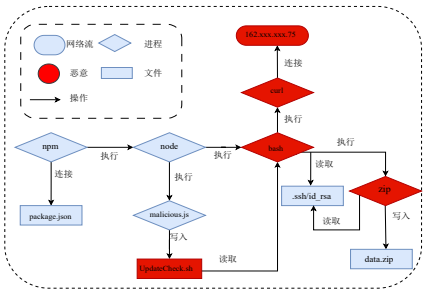

图1 一个通过NPM包发起APT攻击的简化溯源图

Figure 1 Simplified provenance graph of an APT attack vector originating from a malicious NPM package

依赖于一个可信计算基(Trusted Computing Base,TCB)。我们假设操作系统内核的完整性是可信的,攻击者无法在内核层面破坏其核心功能或绕过其监控,且用于生成溯源图的底层审计日志框架本身是完备的,能够安全地记录系统实体间的交互事件。此外,我们假设审计日志的存储和分析过程是安全的,这包括日志被安全地传输到分析服务器,且存储介质具备防篡改机制[30],确保攻击者无法在获取权限后删除或修改其攻击痕迹。本检测框架自身也被假定在TCB内运行,其分析过程不会受到攻击者的干扰[31]。我们的目标就是从这些痕迹构成的溯源图中,识别出由恶意交互所引发的图结构与上下文模式的异常。本框架作为自监督系统,基于两项基本假设:一是训练所用的良性数据集是纯净的,因此数据投毒攻击[32]不在本模型的直接防御范围内;由系统更新或配置变更等非恶意活动引起的正常行为演变[26],我们将其视为一个独立的模型适应性挑战,而非本模型所定义的直接攻击。

2.3 系统架构

如图2所示,我们的模型详细定义了一个从原始日志到最终威胁判定的多阶段流程,主要包括溯源图构建、层次化拓扑掩码生成、图表示学习以及自适应k-NN异常检测四个关键阶段。

(1)溯源图构建:作为系统的统一输入,框架首先从主机端收集原始的操作系统审计日志。这些海量的、非结构化的日志流随后被解析,并根据事件间的因果依赖关系构建成一个大规模的数据溯源图[33]。为了使GNN能够处理这种异构信息,我们首先定义了溯源图的模式,如表1所示。这个模式明确了系统中的核心实体类型、它们之间的交互关系以及用于表征它们的属性。

(2)层次化拓扑掩码生成:本框架采用了一种层次化拓扑知识指导的掩码机制,并将其作为一个离线预处理步骤。该机制并非在训练时动态生成掩码,而是在训练前对良性训练数据集中的每个溯源图进行

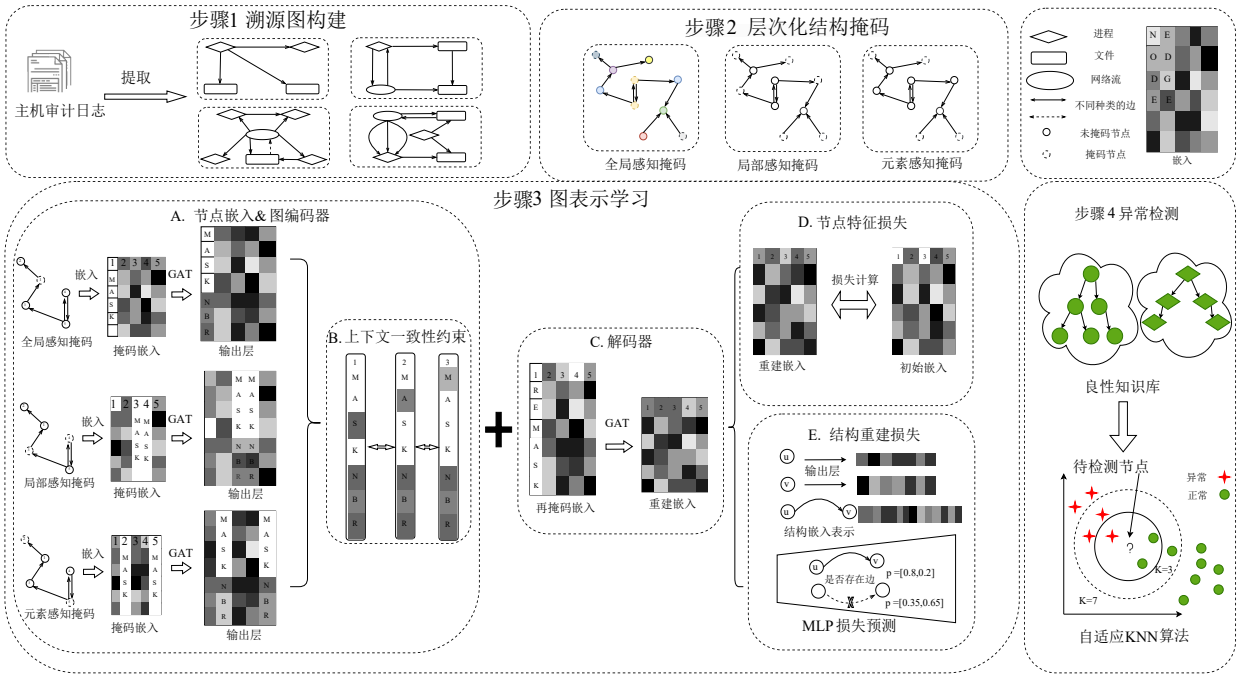

图2 基于层次化感知图自编码器的APT检测框架总体架构

Figure 2 Overall architecture of the APT detection framework based on hierarchical-aware graph autoencoders

表1 系统实体类型、属性、边类型

Table 1 System entity types, attributes, and edge types

类别	类型
节点类型	进程
	文件
	网络
节点属性	进程名
	路径
	IP地址
边类型	读取
	写入
	执行/创建子进程

分析。它通过全局感知、局部感知和元素感知三个层面,智能地选择对模型学习最具信息量的节点进行掩码。这个过程会为每个训练图生成一组对应的掩码视图。这种策略旨在保留关键的图拓扑结构和因果链,并将计算开销转移到离线阶段。

(3)图表示学习:图自编码器模型由一个GNN编码器和一个重构解码器组成,其目标是在给定掩码视图的情况下还原原始图的特征和结构。在编码器部分,本框架采用一个基于图注意力网络[34](Graph Attention Network,GAT)的架构。对于输入的图,GAT编码器通过自注意力机制为邻居节点分配不同的权重,并聚合这些信息来生成节点的低维潜在表示。通过堆叠多层GAT,编码器能够捕捉到高阶邻域信息,最终输出包含丰富上下文和拓扑结构信息的节点嵌入矩阵;解码器负责利用这些学到的节点嵌入来执行两个并行的重构任务:一个GAT的聚合层用于重构被掩码的原始节点特征;同时一个基于两层MLP的样本级结构重构模块用于预测非掩码节点对的边存在概率,以实现图局部结构的间接重构。其次是编码器的局部一致性约束,我们额外引入了节点级上下文一致性损失。该正则化项直接作用于编码器提取的节点嵌入,通过约束同一节点在不同层次化掩码视图下的一致性,迫使模型学习其在局部拓扑波动下的不变特征,确保训练阶段的优化目标与推理阶段的节点级检测任务在粒度上实现逻辑对齐。最终,模型在节点特征重构、结构重构与局部一致性损失的联合监督下进行优化,以确保在还原图信息的同时,学到具备因果稳健性且高度判别性的潜在表示。

(4)自适应k-NN异常检测:在检测阶段,我们利用训练好的GNN编码器,对所有溯源数据进行一次前向传播,高效地提取出图中节点的表示向量。我们采用一种无监督的自适应k-NN算法来判定异常。该算法首先会评估每个节点嵌入向量在其邻域空间中的局部密度,根据该密度动态地为每个节点分配合适的邻居数量,最后通过计算该节点到其动态邻居集的距离来量化其异常分数。如果一个节点的行为模式在表示空间中显著偏离了训练集中学到的良性模式,它将被赋予高异常分并被识别为威胁。

3 基于图表示学习的自适应异常检测方法

本章将阐述我们提出的APT检测框架的各个关键组件。主要由四个主要部分组成:溯源图的构建与节点嵌入表示部分介绍了如何将原始审计日志逐步转换为机器可理解的高维语义特征;层次化拓扑掩码的生成机制,将阐述如何为图学习表示智能地生成保留关键拓扑信息的预训练任务;图表示学习将说明GNN自编码器如何通过联合监督损失来学习良性行为的鲁棒表示;异常检测模块将描述如何通过无监督方式识别与良性模式偏离的异常节点。

3.1 溯源图构建方法与节点嵌入表示

本框架所采用的检测模型以数据溯源图作为核心输入。溯源图是一种能够展示系统执行操作记录的数据结构,它通过将底层的操作系统审计日志抽象为图结构,进而获取系统实体间信息流动和因果关系的信息。我们的框架首先从受监控的主机收集由操作系统审计工具生成的原始事件日志流[4-5,13-14]。这些原始日志由大量半结构化的文本数据构成,以时间顺序记录了内核层面的每一个系统调用和相关活动。随后日志解析对这些原始日志进行处理,用于识别日志事件中的实体,并将它们之间的交互关系转换为图的节点和边。具体而言,我们主要关注三类核心的系统实体作为图中的节点:进程、文件、网络。图中的有向边代表了这些实体之间发生的、具有明确因果关系的系统事件,这些交互通常由系统调用触发,并决定了信息流动的方向。

该过程首先将节点的文本属性分解为有意义的词元序列,我们在大规模的良性溯源数据语料库上训练一个词嵌入模型Skip-gram[35],该模型能够学习到词元间的上下文关系,并将每个词元投影到一个低维连续向量空间中。该模型核心目标是通过tokenw_i预测其上下文词元w_c,迫使模型学习语义相似性,以/bin/bash和/bin/ls为例,它们共享了usr和bin等相同的上下文词元。为了在训练中同时成功预测这些共同的上下文,模型通过梯度下降会“被迫”将bash和ls的词元向量v_{bash}和v_{ls}移动到嵌入空间中彼此邻近的位置。相反,/tmp/data由于其上下文环境完全不同,其词元向量v_{tmp}和v_{data}最终会分布在远离bin和bash的区域。训练完成后,我们便可将一个节点n_i的属性分解为其词元序列$T_i=\{t_1,t_2,\cdots,t_K\}$,并检索出每个词元$t_k$对应的嵌入向量$v_{t_k}$。最后,我们通过平均池化将它们聚合成节点初始特征向量X_i:

$$X_i=\frac{1}{K}\sum_{k=1}^{K}v_{t_k} \tag{1}$$

3.2 层次化拓扑掩码生成机制

传统图自编码器(Graph Auto-Encoder,GAE)所采用的随机掩码策略在应用于结构复杂的溯源图时存在固有缺陷,这种策略在选择掩码对象时是与拓扑无关的,它等概率地对待所有节点和边,忽视了图中的结构重要性。这可能导致无意中破坏了图中关键的因果链(例如一个完整的“进程-读-文件-写-网络”流),导致GNN编码器难以从受损的图中学习到有意义的上下文信息;二是可能产生了过于简单的重构任务(仅掩码了某个高度连接节点的几个邻居),模型仅凭局部的简单插值即可恢复,而无需学习深层的语义表示。

为解决此问题,我们不再使用单一的随机掩码,而是引入一种层次化拓扑知识指导的掩码生成机制[19]。该机制的目标是保留关键拓扑信息的掩码视图,从而引导GNN自编码器学习图的深层结构特性,而非表面统计规律。我们的掩码生成器并非单一策略,而是从三个不同的粒度全局感知、局部感知和元素感知来分析图的拓扑结构,并据此生成不同的掩码视图。这一过程作为自监督预训练的任务生成阶段,应用于所有良性训练图,设$G=(V,E,X)$,邻接矩阵A,节点特征$X=\{x_1,x_2,x_3,\cdots x_v\}$,训练时,对被选中节点集合$\mathcal{M}\subseteq V$以可学习向量$\mu$代替输入:

$$x_i=\begin{cases}\mu_i, & i\in\mathcal{M}\\ x_i, & i\notin\mathcal{M}\end{cases} \tag{2}$$

全局感知掩码(Global-Aware Masking,GAM):该策略旨在保留图的全局连通性和结构多样性,确保每个顶点接收到足够的上下文信息。我们首先利用图着色算法[36]对溯源图G中的节点进行着色,确保任意两个相邻的节点v_j,v_j具有不同的颜色$C(v_j)\neq C(v_j)$。在生成掩码视图时,GAM至少选择m种颜色为溯源图上色,并选取所有颜色种类一半的节点进行掩码C_{mask}。由于同一颜色的节点彼此不相邻,这种策略确保了保留下来的子图G_{remain}仍然具有较高的连通性,迫使模型必须依赖更长距离的依赖关系和全局拓扑知识来重构被掩码的节点。

局部感知掩码(Local-Aware Masking,LAM):为抑制节点间的信息冗余问题、迫使模型学习多种节点类型的局部交互。在溯源图中,许多节点的行为高度相似,其信息可以轻易地从彼此中推断出来。该策略通过计算相邻顶点间的共同邻居数量来评估这种局部冗余[37]:

$$cn(v_j,v_j)=\left|\mathcal{N}(v_j)\cap\mathcal{N}(v_j)\right| \tag{3}$$

其中,$\mathcal{N}(v_i)$代表顶点v_i的邻居集合,$|\cdot|$代表集合中满足条件的元素个数,利用邻接矩阵以C_N^2组合方式

计算所有顶点对共同邻居的相似性,并根据节点之间的相似性降序排序,选取 $Top-k$ 的顶点对并据此遮蔽约一半数量的顶点。

元素感知掩码(Element-Aware Masking, EAM):基于节点在图拓扑中的重要性来进行掩码,以平衡核心结构和边缘模式的学习,因此我们通过计算每个节点的度,对边缘节点适度提高被遮蔽概率;我们首先根据节点度 $d(v)$ 和一个超参数 a 来为图中每个节点 v 分配一个掩码权重 $w_v=\left(d(v)\right)^{\alpha}$。然后,通过对全图 V 中所有节点的权重进行归一化,我们得到了每个节点 v 被选中进行掩码的最终概率:

$$P(v)=\frac{\left(d(v)\right)^{\alpha}}{\sum_{u\in V}\left(d(u)\right)^{\alpha}} \tag{4}$$

其中 $d(v)$ 是节点 v 相连的边的数量,通过调控 a 的值,我们可以控制不同度的节点的采样权重,当设置 $a<0$ 时,会显著增加对 d_v 较小的边缘节点的采样权重,从而提升了对良性稀有行为的辨识力。

3.3 采样预训练机制

基于层次化拓扑知识的掩码策略,特别是局部感知掩码,其计算复杂度远高于随机掩码,如果在每个训练批次中即时地重新计算这些掩码,将引入巨大的计算开销,严重拖慢模型训练速度,使其在大型溯源图上变得不可行。

为了解决这一效率瓶颈,我们设计了一种层次化掩码采样与预生成机制,该机制的核心思想是以空间换时间,将掩码生成过程与掩码应用过程进行解耦。具体是指由三种策略生成的多种掩码组缓存到磁盘中,随后利用掩码采样算法进行批次训练模型。

算法1阐述了层次化掩码采样与预生成机制的完整流程。行3-7执行初始化操作,为全局、局部和元素三种策略分别创建空的掩码缓存池及对应的循环索引。随后行8-15执行计算开销集中的离线预生成阶段,通过调用三个掩码生成函数来计算掩码,并将结果填充到各自的缓存池中。行16-36为在线动态采样阶段,根据该随机数的值从三个缓存池中采样一个预先计算好的掩码视图添加到当前批次中。最后第37行返回包含所有用于训练批次的队列。

我们对上述算法的渐近时间复杂度分为两部分进行分析,设图 G 由 V 个节点,E 条边构成,共计 L 轮训练,N 种掩码策略。其中离线预生成阶段由三部分构成:其中全局感知掩码的时间复杂度为 $O(V^2)$,局部感知掩码同样需要 $O(V^2)$ 的时间,而元素感知掩码的时间复杂度约为 $O(V\log V)$,因此总的预生成时间复杂度为 $O(LV^2)$。而在在线训练阶段仅涉及常数级 $O(L)$ 的池采样次数及 $O(LV)$ 次掩码应用,因此每步训练时间开销仅 $O(V)$,使其与即时随机掩码的渐近时间复杂度持平,确保了训练效率。在空间复杂度方面,需要额外的 $O(N\cdot L\cdot V)$ 空间存储 $N\cdot L$ 个预生成的掩码图,这保证了离线的预生成策略与在线训练互不影响。

算法1 层次化掩码采样预生成

输入:图 $G=(V,A,X)$,掩码比率 p,训练步数 E,视图个数 L

输出:掩码后的特征对 $Q=\left\{\left(\tilde{X}^{(v)},M^{(v)}\right)\right\}$

1. 步骤0:初始 $i_s\leftarrow 0,i_l\leftarrow 0,i_g\leftarrow 0$;
2. $P_g\leftarrow\varnothing,P_i\leftarrow\varnothing,P_e\leftarrow\varnothing,Q\leftarrow\varnothing$
3. 步骤1:掩码预生成
4. FOR t=1 TO E DO
5. $M_g\leftarrow$ *全局掩码*(G,p)
6. $M_l\leftarrow$ *局部掩码*(G,p)
7. $M_e\leftarrow$ *元素感知掩码*(G,p)
8. $P_g\leftarrow P_g\cup\{M_g\}$
9. $P_l\leftarrow P_l\cup\{M_l\}$
10. $P_e\leftarrow P_e\cup\{M_e\}$
11. END FOR
12. 步骤2:固定比例采样与应用
13. FOR t=1 TO L DO
14. batch $\leftarrow\varnothing$
15. FOR v=1 TO E DO
16. r~Uniform(0, 1)
17. IF r < 0.2 THEN
18. $M\leftarrow P_g\left[i_g\right]$
19. $i_g\leftarrow\left(i_g+1\right)\%\left|P_g\right|$
20. ELSE IF r < 0.7 THEN
21. $M\leftarrow P_l\left[i_l\right]$
22. $i_l\leftarrow\left(i_l+1\right)\%\left|P_l\right|$
23. ELSE
24. $M\leftarrow P_e\left[i_e\right]$
25. $i_e\leftarrow\left(i_e+1\right)\%\left|P_e\right|$
26. END IF
27. $\tilde{X}^{(v)}\leftarrow$ *应用掩码*(X,M)
28. batch $\leftarrow$ batch $\cup\left\{\left(\tilde{X}^{(v)},M\right)\right\}$
29. END FOR
30. $Q\leftarrow Q\cup\{batch\}$
31. END FOR
32. RETURN Q

3.4 图表示学习

我们的图表示学习阶段由图编码器、节点级上下文一致性模块、图解码器构成,其核心任务是接收层次化拓扑知识掩码过的图 $\hat{G}=(V,A,\hat{X})$,通过编码器学

习节点潜在表示H,节点级一致性模块通过约束同一节点在不同掩码视图下嵌入表示的一致性,迫使模型学习不随局部邻域扰动而改变的因果不变特征。最终由解码器通过特征重构与结构重构提供监督信号,利用H重建被掩码的节点特征,从而训练模型具备从受损上下文中恢复原始信息的能力。

其中在被掩码的输入图$\tilde{G}$中,节点特征矩阵$\tilde{X}$包含两部分:对于未被掩码的节点$v_i \notin \mathcal{M}$,其特征$\tilde{X}_i$保持原始语义嵌入X_i;对于被掩码的节点$v_m \in \mathcal{M}$,其特征$\tilde{X}_m$被替换为一个可学习的向量X_{mask}。

3.4.1 图自编码器

图编码器的目标是从输入的掩码图$\tilde{G}$中提取每个节点的上下文感知嵌入,最终生成一个潜在嵌入矩阵$H \in \mathbb{R}^{N\times d'}$。

我们选择图注意力网络(GAT)作为编码器的骨干架构。GAT的核心优势在于它能够利用自注意力机制,在聚合邻居信息时为不同的邻居节点分配不同的重要性权重。这在溯源图分析中至关重要,因为一个系统实体的邻居对其行为模式的贡献显然是不同的。例如,一个进程与cmd.exe(一个强大的命令行外壳)的交互,相比其与svchost.exe(一个通用的Windows服务宿主)的交互,前者往往携带更强、更明确的行为意图信号[38]。

编码器由L层GAT堆叠而成。在第l层,GAT接收上一层的节点嵌入$H^{(l-1)}$(其中$H^{(0)}=\tilde{X}$)作为输入,并计算新的嵌入$H^{(l)}$。

对于图中的任意一个节点i,首先GAT会对节点i的$l-1$层特征$h_j^{(l-1)}$应用一个共享的可学习线性变换,该变换由权重矩阵$W^{(l)}$参数化,将其投影到目标嵌入空间:

$$z_j^{(l)} = W^{(l)} h_j^{(l-1)} \tag{5}$$

下面模型会为节点i的每一个邻居$j \in \mathcal{N}_i$计算一个原始的、未归一化的注意力系数e_{ij}。该系数e_{ij}表示节点j的特征对节点i的重要性。我们采用一个单层前馈神经网络权重$a^{(l)}$实现该注意力机制,并通过LeakyReLU激活函数增加非线性,其中$\|$表示向量拼接。

$$e_{ij}^{(l)} = \text{LeakyReLU}\left((a^{(l)})^T\left[z_i^{(l)} \| z_j^{(l)}\right]\right) \tag{6}$$

为了使注意力系数易于比较且数值稳定,我们使用softmax函数对节点i的所有邻居$k \in \mathcal{N}_i$的e_{ik}进行归一化,得到最终的注意力权重a_{ij}:

$$\alpha_{ij}^{(l)} = \text{softmax}_j(e_{ij}^{(l)}) = \frac{\exp(e_{ij}^{(l)})}{\sum_{k\in\mathcal{N}_i}\exp(e_{ik}^{(l)})} \tag{7}$$

$\alpha_{ij}^{(l)}$满足$\sum_{j\in\mathcal{N}_i}\alpha_{ij}^{(l)} = 1$,代表了邻居$j$在第$l$层对节点$i$的贡献权重。

节点i第l层嵌入表示$h_i^{(l)}$通过对其邻居变换后的特征$z_j^{(l)}$进行加权求和得到,我们使用一个非线性激活函数σ来完成该层的计算:

$$h_i^{(l)} = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^{(l)} z_j^{(l)}\right) \tag{8}$$

为了稳定学习过程并使模型能够捕获邻域中不同方面的语义信息,GAT采用了多头注意力机制。该机制并行地执行K个独立的注意力机制,每个头使用其独立的参数$(\mathbf{W}^{(l)})_k$和$(a^{(l)})_k$来计算一组注意力权重$(\alpha_{ij}^{(l)})_k$和该头输出的嵌入$(h_i^{(l)})_k$。在我们的编码器中,我们将K个的输出进行拼接作为第l层的最终输入嵌入:

$$h_i^{(l)} = \Big\|_{k=1}^{K}\left(\sigma\left(\sum_{j\in\mathcal{N}_i}(\alpha_{ij}^{(l)})_k\,(\mathbf{W}^{(l)})_k h_j^{(l-1)}\right)\right) \tag{9}$$

通过堆叠L层这样的多头GAT层,图编码器能够捕获节点的多跳邻域信息。编码器的最终输出$H^{(L)}$是一个包含图中所有节点的深层上下文感知嵌入的矩阵。

3.4.2 局部上下文一致性约束

我们发现,GAT编码器学习到的节点嵌入表示的稳定性易受输入掩码随机性的干扰。例如训练集中的一个良性进程,在某一训练轮次中,其邻居可能因局部掩码策略被大量遮蔽,导致其上下文信息稀疏。嵌入$H_i^{(1)}$偏向异常;而在另一轮次中,其嵌入在全局掩码策略的作用下$H_i^{(2)}$表现正常,这种嵌入波动会混淆后续的异常检测器,导致误报。

我们引入了一种节点级上下文一致性约束机制[39],旨在通过挖掘节点在不同局部扰动视角下的因果不变特征。该机制与主GAT编码器并行工作,对每个图$\tilde{G}^{(v)}$的节点特征矩阵X应用了L种不同策略的掩码,得到$\tilde{X}^{(v)}$。

我们首先利用图同构网络(GIN)[40]更新不同视图上的节点i表示,第l层GIN对节点i的表示更新为:

$$h_i^{(l)} = \text{MLP}^{(l)}\left(h_i^{(l-1)} + \sum_{j\in\mathcal{N}_i} h_j^{(l-1)}\right) \tag{10}$$

在K层GIN聚合后,我们计算各视图间对应的节点级一致性损失:

$$\mathcal{L}_{\text{sub}} = \frac{1}{|V|\cdot C_k^2}\sum_{i=1}^{|V|}\sum_{1\le a<b\le k}\| z_i^{(v_a)} - z_i^{(v_b)}\|^2 \tag{11}$$

其中,$z_i^{(v_a)}$和$z_i^{(v_b)}$分别表示节点i在视图a和视图b下的潜在嵌入,k为视图个数。通过这种约束,模型学习不随局部结构波动而改变的稳健特征。

3.4.3 图解码器

图解码器核心任务是接收编码器生成的潜在嵌入矩阵$H \in \mathbb{R}^{N\times d'}$,输出三个独立损失函数构成的联合损失$\mathcal{L}_{\text{Total}}$。通过特征重构、结构重构、局部上下文一致性协同监督来共同优化模型,旨在确保模型学习到的嵌入表示既能捕获系统实体的内在属性,又能理解其拓扑交互模式,同时抵抗掩码带来的随机扰动。

节点特征重构是图自编码器的核心自监督任务,该任务的目标是使解码器重构出的掩码节点特征$\hat{X}_i$与其原始特征X_i尽可能相似。首先我们接收到图编码器包含所有节点的潜在嵌入矩阵$\hat{H}$。对在编码器输入阶段的掩码节点集合$\mathcal{M}$执行重掩码操作,并提取出被掩码节点所对应的潜在嵌入$H_{\mathcal{M}}=\{h_i|i\in\mathcal{M}\}$。我们重新将$H_{\mathcal{M}}$通过GAT层聚合邻居上下文,并应用解码器权重$\mathbf{W}_{dec}$将高维嵌入降维投影回原始特征维度$d$,最后计算这些被掩码节点的重构特征$\hat{X}_i(i\in\mathcal{M})$与其对应的原始节点特征$X_i$之间的缩放余弦误差(Scaled Cosine Error,SCE)[41]来定义特征误差:

$$\mathcal{L}_{fr}=\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}\left(1-\frac{X_i\cdot\hat{X}_i}{\|X_i\|\cdot\|\hat{X}_i\|}\right)^{\gamma} \tag{12}$$

其中,X_i和$\hat{X}_i$分别是节点i的原始与重构特征,γ是放大样本误差的缩放因子。

在特征重构的同时我们引入了样本级的结构重构损失$\mathcal{L}_{sr}$,我们从图G中采样两组节点对:

正样本集$\mathcal{P}_{\text{pos}}$:从图中随机采样的真实存在的节点对$(i,j)\in A$。

负样本集$\mathcal{P}_{\text{neg}}$:从图中随机采样的不存在连接的节点对$(i,k)\notin A$。

我们通过一个轻量的多层感知机(Multilayer Perceptron, MLP)[42]来预测任意两个节点的潜在嵌入(H_i,H_j)之间存在连接的概率p_{ij}:

$$p_{ij}=\text{Sigmoid}(\text{MLP}(H_i /\!/ H_j)) \tag{13}$$

$\mathcal{L}_{sr}$采用标准的二元交叉熵(Binary Cross-Entropy, BCE)[43]损失函数来优化MLP,目标是使正样本的预测概率接近1,负样本的预测概率接近0:$\mathcal{L}_{sr}=$

$$-\frac{1}{|\mathcal{P}_{\text{pos}}|}\sum_{(i,j)\in\mathcal{P}_{\text{pos}}}\log(p_{ij})-\frac{1}{|\mathcal{P}_{\text{neg}}|}\sum_{(i,k)\in\mathcal{P}_{\text{neg}}}\log(1-p_{ik}) \tag{14}$$

通过最小化$\mathcal{L}_{sr}$,编码器将拓扑上相连的节点在潜在空间中映射得更近,从而使H蕴含丰富的结构信息。

最后我们通过一个联合损失函数实现端到端优化:

$$\mathcal{L}_{\text{Total}}=\mathcal{L}_{fr}+\alpha\mathcal{L}_{sr}+\beta\mathcal{L}_{mv} \tag{15}$$

其中α和β两个超参数用于平衡三个任务在总损失中的权重。

在APT检测场景中,攻击节点仅占极小比例,传统的全局判别容易导致微弱信号被稀释。本框架在训练阶段通过设计节点级多尺度协同机制:节点级重构损失确保了模型对实体微观交互模式的细粒度;而通过约束同一节点在不同掩码视图下的表示一致性,迫使编码器学习良性行为的局部结构稳定性。这种约束机制实现了训练目标与推理任务在粒度上的逻辑对齐,通过强化模型对良性局部因果链的理解,显著提升了其识别结构性偏差的敏感度。

3.5 自适应异常检测

图编码器通过训练阶段的节点级上下文一致性约束,能够深刻捕捉良性行为在局部邻域内的结构不变性,促使具有相似行为特征的系统实体在嵌入空间内紧密聚类[44-47]。相比之下,恶意或异常活动由于破坏了良性逻辑中的局部因果链,其嵌入表示将偏离这些良性簇,呈现为高维空间中的孤立点。因此本文将威胁检测任务转化为一个基于节点距离的异常检测问题。

然而固定k值的k-NN算法挑战在于溯源图嵌入空间的密度不均性:不同类型的系统实体(如高频交互的文件与低频交互的网络套接字)其良性嵌入本就具有截然不同的局部密度,固定k值易导致稀疏良性簇的误报或密集恶意簇的漏报。

因此我们基于一种基于局部密度算法(Local Correlation Integral,LOCI)[48]来执行异常检测,该算法的详细流程如算法2所示,算法3-10行在训练集上构建良性知识库$\mathcal{B}$,并为每种节点类型t计算独有的平均距离$d_{\text{avg}}[t]$;11-23行计算出每个待测节点嵌入z_v到同类型良性嵌入之间的最小距离$d_{\min}(z_v)$,并统计落在该类型平均距离的节点数量记为k_{adaptive}。

最终异常判定通过将节点的异常分数S_v综合嵌入的相对偏离度和局部密度,与异常阈值θ进行比较来完成,计算公式如下:

$$\begin{aligned}
&d_{\min}(z_v)=\min_{z_b\in\mathcal{B}[t_v]}\|z_v-z_b\|_2\\
&k_{\text{adaptive}}(z_v)=\left|\{\|z_v-z_b\|_2\le d_{\text{avg}}[t_v]\quad|z_b\in\mathcal{B}[t_v]\}\right|\\
&S_v=\left(\frac{d_{\min}(z_v)}{k_{\text{adaptive}}(z_v)+\delta}\right)\times\left(\frac{d_{\min}(z_v)}{d_{\text{avg}}[t_v]}\right)\\
&\text{Detection}(z_v)=\begin{cases}\text{Anomalous} & \text{if } S_v>\theta\\ \text{Benign} & \text{if } S_v\le\theta\end{cases}
\end{aligned} \tag{16}$$

4 实验与评估

4.1 实验环境配置

我们提出的框架基于Python 3.8和PyTorch 1.12.1实现。在数据预处理与图构建阶段,我们首先使用

算法2 自适应异常检测

输入：良性训练样本嵌入E_{train}，测试样本嵌入E_{test}，实体类型集合T，平滑因子δ

输出：异常分数列表S

1. 步骤1:预构建良性知识库
2. FOR ALL t in T DO
3. $E_t \leftarrow \{z \in E_{\text{train}} \mid \text{type}(z)=t\}$
4. $d_{\text{avg}}[t] \leftarrow$ *计算平均距离* (E_t)
5. $B[t] \leftarrow \left(E_t, d_{\text{avg}}[t]\right)$
6. END FOR
7. 步骤2:在线检测与异常分数计算
8. $S \leftarrow \varnothing$
9. FOR ALL z_v in E_{test} DO
10. $t_v \leftarrow \text{type}(z_v)$
11. $(E_t, d_{\text{avg,t}}) \leftarrow B[t_v]$
12. $d_{\min} \leftarrow \min_{z_b \in E_t} \| z_v - z_b \|_2$
13. $k_{\text{adaptive}} \leftarrow 0$
14. FOR ALL z_b in E_t DO
15. IF $\| z_v - z_b \|_2 \le d_{\text{avg,t}}$ THEN
16. $k_{\text{adaptive}} = k_{\text{adaptive}} + 1$
17. END IF
18. END FOR
19. $S_v \leftarrow \left(d_{\min} / \left(k_{\text{adaptive}} + \delta\right)\right) \times \left(d_{\min} / d_{\text{avg,t}}\right)$
20. 将S_v添加到S
21. END FOR
22. RETURN S

Python标准库解析原始主机审计日志，提取系统实体及其交互关系。随后我们利用NetworkX库构建初始的溯源图结构。在模型训练阶段，图神经网络模型利用DGL 0.9.1库实现。在最终的检测阶段，本文提出的自适应异常检测算法利用了scikit-learn 1.2.2库中的K-D树索引来实现高效的近邻搜索。所有实验均在同一台配置有NVIDIA GeForce RTX 4090 GPU的高性能服务器上进行。

超参数的设置上，隐藏层维度设为128，掩码率设为0.5，初始学习率设为1×10^{-3}。对于本文提出的层次化拓扑掩码策略，全局感知、局部感知和元素感知的掩码率分别设为0.3、0.2和0.5。在联合损失函数中，结构重构损失和节点级上下文损失的权重系数分别设置为1和0.1。

为全面评估模型的检测性能，我们采用了四个广泛使用的标准分类指标：精确率、召回率、F1分数和误报率（False Positive Rate，FPR）。在入侵检测领域，FPR是衡量模型在海量良性数据中产生错误告警的关键指标，FPR越低，模型的实用性越高。

4.2 数据集介绍

为了全面评估所提出框架的有效性，我们在多个广泛使用的公开溯源图数据集上进行了实验。我们将我们的模型与多种最先进的基线模型[4,17,25-26]在不同检测粒度上进行了比较。我们将数据集分为两大类，以支持不同粒度的检测任务：批量级（Batch-level）和实体级（Node-level）。

批量级数据集分布情况如表2所示，此类数据集旨在评估模型对整个图进行良性或恶意判定的能力，其Ground Truth仅在图级提供，用于指示整段系统执行或行为批次是否包含攻击行为，因此我们通过对图中节点级异常分数进行统计聚合来刻画系统整体行为的异常程度，我们采用了两个标准数据集：

Streamspot[22]：它总共包含600个溯源图，其中包括五种良性活动场景和一种恶意攻击场景。

Unicorn Wget[23]：模拟APT攻击中的恶意wget下载行为。它包含150个图，其中125批是良性的。

表2 批量级检测数据集分布情况 单位：个

Table 2 Data distribution for batch-level detection unit: count

数据集	图	节点	边
Streamspot	600	50 463	897 709
Unicorn Wget	150	522 580	1 925 113

对于实体级数据集，该类数据集要求模型在海量节点中精确定位异常实体。我们采用了DARPA E3[24]数据集，该数据集通过DARPA TC Engagement 3演习收集，在一个企业网络环境中混合了海量的良性系统活动与由专业红队执行的多阶段APT攻击。如表3所示，我们采用了三个主要的子数据集Theia、Trace和Cadets，其主要攻击向量包括Firefox、Nginx后门、浏览器扩展攻击等。针对DARPA E3数据集，我们参考官方发布的红队攻击时间线和ThreaTrace[17]提供的Ground Truth报告，对图中的异常节点进行了标注。

表3 实体级检测数据集分布情况 单位：个

Table 3 Data distribution for entity-level detection unit:count

数据集	良性节点	恶意节点	边
E3Theia	1 598 647	25 319	2 874 821
E3Trace	3 220 594	68 082	4 080 457
E3Cadets	1 614 189	12 846	3 303 264

4.3 对比实验

为全面评估本文所提出框架的有效性，我们首先在五种数据集上对我们的模型进行了独立测试。随后将其与现有的APT检测方法进行比较。

本文所提出模型的详细检测性能如表4所示，结果表明我们的模型在所有数据集上均表现出了卓越的性能。

表4 本模型在不同数据集具体检测结果

Table 4 Detailed detection results of the proposed model across different datasets

数据集	真实标签良性	恶意	真负例	假负例	真正例	假正例	精确率	召回率	F1分数	AUC
Streamspot	500	100	499.21	0.43	99.57	0.79	99.21%	99.57%	99.39%	99.67%
Unicorn Wget	125	25	124.53	0.47	24.51	0.46	98.13%	98.02%	98.22%	98.83%
DARPA E3Theia	319 475	25 319	319 126	19	25 300	349	98.64%	99.92%	99.28%	99.01%
DARPA E3Trace	616 021	68 086	615 775	19	68 067	246	99.64%	99.97%	99.81%	99.63%
DARPA E3Cadets	344 327	12 846	343 910	59	12 787	417	96.84%	99.54%	98.17%	99.42%

在批量级检测任务中,我们的模型在Streamspot数据集上取得了卓越的检测性能。这一近乎完美的检测表现主要得益于Streamspot数据集的特性,其攻击与良性活动均由单一用户在隔离场景下执行,使得良性图与恶意图在拓扑结构上呈现出高度可区分性。如图3所示的t-SNE[49]可视化结果进一步证实了这一点,良性嵌入和恶意嵌入在特征空间中形成了距离较远且清晰可分的簇,从而显著降低了检测难度。面对更具挑战性的Unicorn Wget数据集,该数据集模拟了更为隐蔽的APT攻击,我们的模型依然保持了优异的检测性能,F1分数为98.22%,精确率为98.13%,召回率为98.02%,这充分证明了模型在处理隐蔽攻击场景时的鲁棒性。

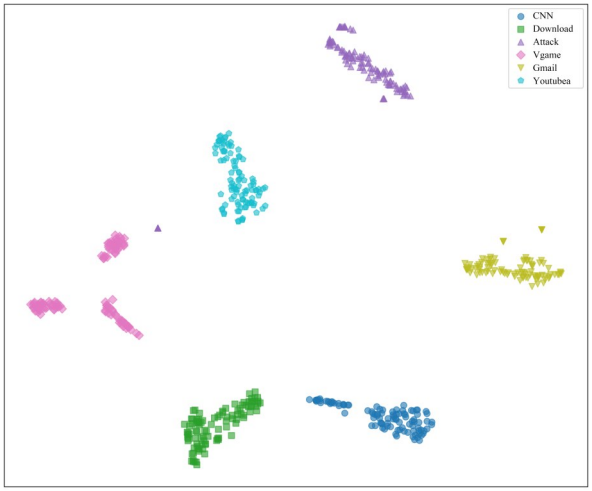

图3 批量级数据集Streamspot分布聚类情况

Figure 3 Clustering distribution of the StreamSpot dataset at the batch level

在实体级检测任务中,DARPA E3数据集以其海量的良性数据和复杂的多阶段APT攻击,对模型的表示能力和检测算法的鲁棒性提出了极高要求。我们的模型在所有三个子数据集上均取得了优异的性能表现。在数据量最大、攻击最为复杂的Trace数据集上,模型实现了99.80%的F1分数,误报率仅为0.04%。在Theia数据集上,模型取得了99.26%的F1分数,误报率为0.09%。在Cadets数据集上,F1分数达到98.18%,误报率为0.12%。我们进一步对Trace和Theia数据集的检测难点进行了深入分析,发现即使面对ztmp端口扫描和profile负载执行这类与良性系统服务高度相似的恶意行为,我们的模型仍能有效识别潜在的攻击链。

为了全面评估我们模型的性能,我们将其与现有的最先进检测方法进行了对比实验,如表5所示。在Streamspot数据集[22]上,我们将我们的模型与Streamspot[25]、Unicorn[26]、ThreaTrace[17]和MAGIC[4]等基线方法进行了比较。实验结果表明,我们的模型在召回率和F1分数方面均达到了最优水平,虽然精确率略低于MAGIC,但误报率显著优于其他模型。值得注意的是,Streamspot数据集由于场景分离特性,不需要额外的复杂信息即可学习节点行为模式,这使得多种方法在此数据集上都能取得较好的性能,但我们的模型依然在综合指标上保持了竞争优势。在Unicorn Wget数据集[23]上,我们的模型同样在大多数指标上表现得更优,这一优势在实际部署中能够显著降低误报的人工审查成本。在DARPA E3数据集[24]的三个子数据集上,我们与ThreaTrace和MAGIC这两个最先进的半监督检测方法进行了全面对比并取得了较强的检测性能指标,这一全面的性能优势主要归因于层次化拓扑掩码策略通过保留全局和局部结构信息,使编码器能够构建更鲁棒的良性行为模式,节点级上下文一致性约束通过维持不同掩码视角下节点表示的稳定性,强化模型对良性行为局部逻辑的捕捉,从而降低对正常系统活动的误报;自适应异常检测根据数据集实体异构性,为不同密度的实体类型提供了专属的检测范围。

为进一步验证层次化掩码策略相较于传统随机掩码的优越性,本研究针对DARPA E3 Trace数据集中的敏感数据泄露场景进行了分析:在该场景中,红队通过受损进程执行了scp操作进行跨机器数据传输,该行为在庞大的良性背景日志中极度稀疏。传统的随机掩码策略在处理此类长程攻击链时,由于缺乏对拓扑结构的感知,极易将该路径上的关键中间节点掩蔽,导致图编码器接收到的上下文信息出现语义断裂。相比之下,本文提出的层次化掩码策略体现了显

著优势。局部感知策略识别出大量的常规文件读写具有高度邻域冗余性的特性，并对其进行高比例掩码，从而迫使模型通过非冗余信息还原节点属性；全局感知策略则通过着色方案保留了跨层级的路径骨干。这解释了为何本模型在E3 Trace数据集上的误报率显著低于MAGIC。

表5 本文方法与现有检测模型对比实验结果 单位：%

Table 5 Experimental comparison between the proposed method and existing detection models unit: %

数据集	方法	精确率	召回率	F1分数	误报率
Streamspot	Streamspot	73.23	89.14	80.40	9.82
	Unicorn	92.02	93.21	98	2.61
	ThreaTrace	98.54	99.23	99.01	**0.59**
	Magic	**99.41**	99.57	99.23	0.61
	(Ours)	99.21	**99.57**	**99.57**	0.75
Unicorn Wget	Unicorn	84.52	89.91	87.13	14.35
	ThreaTrace	93.25	**98.31**	95.71	7.41
	Magic	98.02	96.00	96.98	2.00
	(Ours)	**98.13**	98.02	**98.22**	**1.42**
DARPAE3 Theia	ThreaTrace	88.77	99.01	93.61	0.62
	Magic	96.01	99.49	99.11	0.14
	(Ours)	**98.64**	**99.92**	**99.28**	**0.10**
DARPAE3 Trace	ThreaTrace	72.23	99.01	83.52	2.78
	Magic	99.17	99.73	98.92	0.12
	(Ours)	**99.64**	**99.97**	**99.81**	**0.04**
DARPAE3 Cadets	ThreaTrace	89.45	99.26	94.10	0.28
	Magic	93.24	98.27	95.68	0.21
	(Ours)	**96.84**	**99.54**	**98.17**	**0.12**

4.4 超参数分析

我们框架的最终性能受到多个关键超参数的影响。为了深入理解模型对这些参数的敏感性，并为我们的实验确定最佳配置，我们进行了一系列参数调优实验。所有调优实验均在DARPA E3 Theia数据集上进行。

我们首先评估了GNN编码器输出的嵌入维度d对模型性能的影响。如图4所示，我们将d从32增加到256。实验结果表明，当$d=32$或$d=64$时，模型的F1-Score相对较低，这表明较低的维度不足以捕获溯源图节点丰富的语义和复杂的拓扑信息。当维度增加到$d=128$时，模型性能达到峰值。继续增加到$d=256$时，性能未见显著提升，反而带来了更高的计算和内存开销，且存在过拟合风险。因此，我们在所有实验中均采用$d=128$作为最佳嵌入维度。

我们探究了GNN编码器消息传递层的数量L对性能的影响，测试范围为$L\in\{1,2,3,4\}$。如图5所示，$L=1$时模型性能最差，因为编码器只能聚合1跳邻域

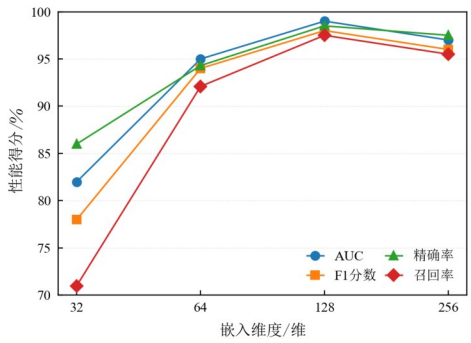

图4 潜在嵌入维度对比实验

Figure 4 Comparative experiment on latent embedding dimensions

信息，无法捕获APT攻击中常见的长距离依赖关系。当层数增加到$L=2$和$L=3$时，性能有显著提升，并在$L=3$时达到最优。这表明3跳邻域信息对于构建鲁棒的节点表示至关重要。然而当$L=4$时，每个节点的感受野会指数级扩大，导致它聚合了图中几乎所有节点的信息。这种过度的、无差别的信息混合会淹没节点自身的局部邻域特征，使得原本具有独特拓扑结构的节点的特征表示变得越来越相似，最终趋于一致，性能开始下降，导致不同节点的表示趋同。因此，我们选择$L=3$作为最佳层数。

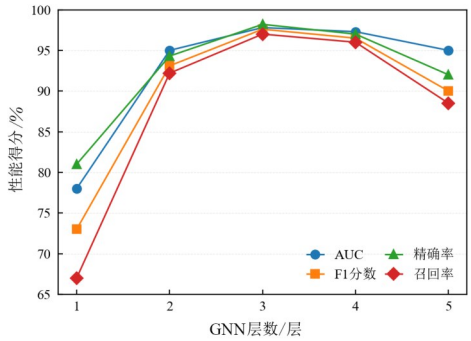

图5 GNN消息层数对比实验

Figure 5 Comparative experiment on the number of GNN message-passing layers

学习率对模型的收敛性和稳定性至关重要。我们在{0.0005, 0.001, 0, 002, 0.005, 0.01}范围内进行了测试。如图6所示，较小的学习率导致模型收敛速度慢，在固定周期内未能达到最优解。而较大的学习率则导致训练过程不稳定，损失函数波动较大，最终性能反而下降。实验证明，学习率为0.001时，模型能够实现最快且最稳定的收敛。

我们分析了预训练阶段的掩码率。如图7所示，我们测试了0.1~0.9之间的不同掩码率。当掩码率过低时，重构任务过于简单，模型无法学习到有意义的深度特征，导致下游检测性能不佳。随着掩码率的提

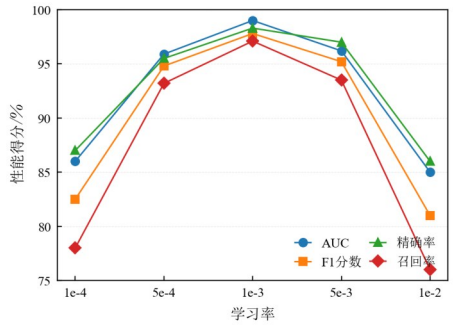

图6　学习率对比实验

Figure 6　Comparative experiment on learning rates

高，重构任务的难度增加，迫使模型学习更鲁棒的拓扑和语义信息，性能随之提升。若掩码率过高，则会破坏过多的图结构，导致上下文信息不足，模型难以收敛，性能急剧下降。因此我们选取掩码率为0.5，确保模型表现最优。

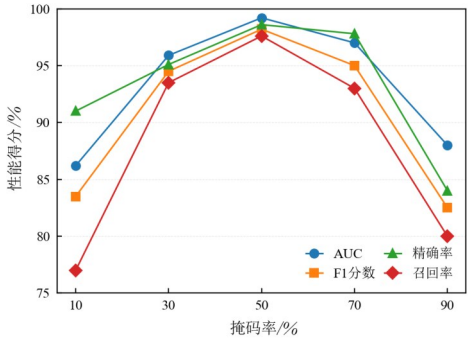

图7　掩码率对比实验

Figure 7　Comparative experiment on masking rates

最后，我们分析了本文引入的层次化掩码策略内部的分配比例。如图8所示，我们测试了三种不同的配比：(0.2，0.3，0.5)、(0.5，0.2，0.3)和(0.2，0.5，0.3)。实验结果表明，(0.2，0.5，0.3)的配比获得了最佳的实验结果。我们分析认为，这一配比是最高效的：它在局部感知上分配了最高权重，在溯源图中，APT攻击的核心特征往往体现在局部的、异常的实体交互模式上(例如，一个良性进程与其不应访问的文件或套接字交互)。通过重点掩码和重构局部邻域结构，该策略迫使模型深入学习这种关键的、细粒度的上下文信息，这是捕获良性行为画像的基础；虽然全局拓扑对理解整个攻击链很重要，但在预训练阶段过度强调它，可能会分散模型对关键局部信号的注意力[50]。保留一个较小的比例足以让模型感知到宏观的图结构；最后，元素感知掩码作为一个重要的补充。它不关注拓扑，而是专注于节点自身的特征。这一比例确保了模型防止在学习邻域结构时过度平滑，忽略了节点本身的语义信息。

4.5　消融实验

为了验证我们框架中各个核心组件的有效性，我们设计了一系列消融实验。我们主要分析了三个方面：(1)层次化掩码策略相较于随机掩码的优越性；(2)节点级局部上下文损失对模型表示能力的影响；(3)自适应异常检测方法相较于其他检测算法的优势。所有实验均在E3-Trace数据集上进行。

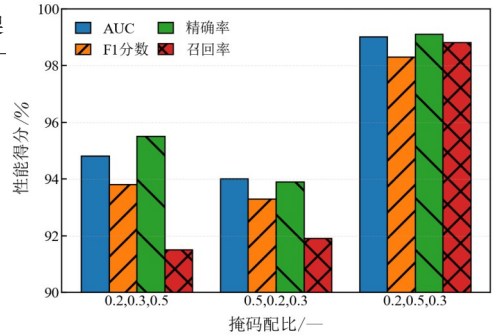

图8　层次化掩码配比消融实验

Figure 8　Ablation study on hierarchical masking ratios

(1)我们对比了随机掩码、单一策略掩码、两两融合掩码以及我们最终采用的三策略融合掩码。如图9所示结果表明，我们最终采用的三策略融合方案取得了最佳性能，这并非简单地叠加，而是一种多尺度协同机制：它强制模型同时处理全局、局部和元素属性三个粒度的重构任务。这种多视图学习机制不仅互补了各自的短板，还通过联合监督构建出信息量最丰富、抗噪性最强的节点嵌入。而传统的随机掩码策略无视了溯源图中节点角色的差异性，极易破坏关键的攻击因果链条，导致模型难以学习到深层次的语义信息。相比之下，我们提出的层次化拓扑掩码策略通过引入领域知识指导掩码过程，无论是单独使用还是组合使用，其性能均显著优于随机基线。这证实了有针对性地保留图结构对于溯源图表示学习至关重要。

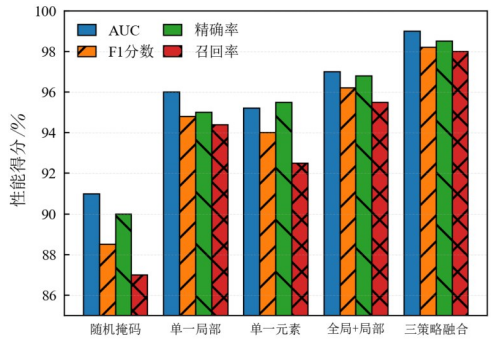

图9　不同掩码策略消融实验

Figure 9　Ablation study on different masking strategies

(2)图10展示了节点级上下文一致性损失对最终检测结果的影响。实验结果表明，移除该一致性约束后，模型检测性能出现显著下降。经过分析，在缺乏一致性约束的情况下，编码器对于同一节点在不同局部扰动所产生的表示存在显著差异。这种不稳定

性使得模型难以捕捉良性系统行为的稳健模式，导致其对正常的、非结构性的行为变体产生过高的敏感度。引入节点级上下文一致性约束后，该机制促使编码器挖掘出在不同视图间保持不变的核心因果结构，不仅确保了模型能基于受损上下文精准重构缺失信息，还为识别结构性异常提供了一个高分辨率的基准，使得任何破坏了局部因果链的恶意行为在检测阶段都能作为显著的离群点被识别，从而更精准地界定异常边界。

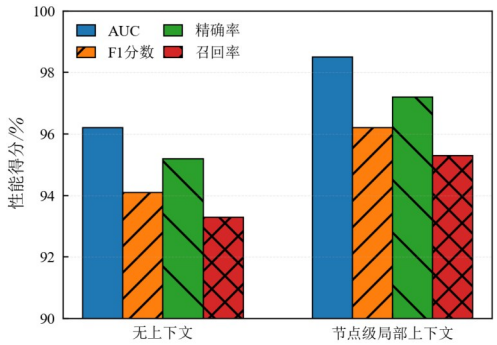

图10 节点级局部上下文消融实验

Figure 10 Ablation study on node-level local context

（3）在获得高质量节点嵌入后，检测算法的选择直接决定了最终效果。如图11结果所示，传统的检测方法如孤立森林和K-均值在溯源图场景下表现不佳，因为它们通常假设异常点远离所有良性簇的中心，而APT攻击往往作为局部异常伪装成正常的系统交互，嵌入在良性节点簇的边缘；标准的k-NN算法利用全局训练样本平均距离作为阈值与待检测节点与其附近的k个邻居的距离进行比较，但其固定的k值无法适应溯源图中极端的密度差异。实验表明，自适应异常检测算法通过动态调整k值解决了这一难题，它能根据目标节点的类型和所处环境的局部拓扑复杂性，智能地确定最佳参考邻域大小。这种机制实际上为每种类型的系统实体定制了独有的异常判定边界，从而在保持对未知威胁高敏感度的同时，最大限度地减少了因环境异质性导致的误报。

4.6 性能开销

除了检测精度，计算效率和资源开销是衡量一个APT检测框架是否具有实际部署价值的关键指标。本节将从理论时间复杂度和运行开销两个角度，评估我们所提出框架的性能。

我们的框架流程共涉及四个步骤：图构建、掩码预训练、图自编码器训练和异常检测。

假设输入的溯源图为$G=(V,E,X)$，其中$N=|V|$为节点总数，$M=|E|$为边总数，d为GNN的嵌入维度，K为训练周期，L为GNN层数。

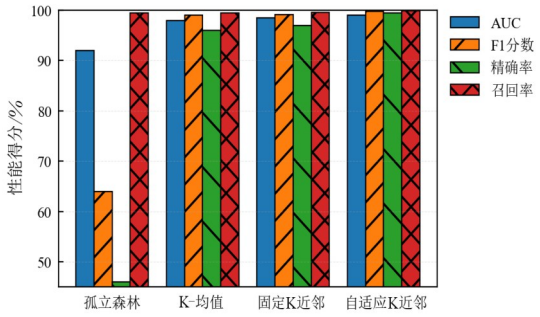

图11 不同检测方法消融实验

Figure 11 Ablation study on different detection methods

图构建阶段涉及解析原始日志并构建图的邻接关系。其时间复杂度与图的规模成正比，为$O(N+M)$；在预训练阶段，我们生成三种掩码策略，由于全局感知和局部感知掩码的复杂度均为$O(N^2)$，掩码预训练阶段的总时间复杂度为$O(N^2)$；在图自编码器训练阶段的复杂度约为$O(K\cdot L\cdot(M\cdot d+N\cdot d^2))$；最后异常检测阶段的时间复杂度为$O(N\log N)$。

我们在E3-Theia数据集上评估了上述四个阶段的实际时间与内存开销，结果如表6所示。我们的框架通过精心设计的离线预计算策略，实现了训练效率的显著提升。在掩码预训练阶段，三种掩码策略可以预生成并缓存，只需在训练前计算一次，后续训练过程可直接复用这些预生成的掩码，避免了重复计算的开销。通过将掩码预生成与图自编码器训练解耦，我们的框架实现了训练过程的模块化和高效化。在保持高检测精度的同时，该框架的离线预计算设计使其具备在真实世界APT检测场景中部署所需的计算效率，充分证明了其实际部署价值。

表6 时间空间开销分析

Table 6 Time and space overhead analysis

阶段	时间开销/s	峰值内存占用/MB
图构建	578	2 412
预训练	1 251	1 958
训练	540	1 152
异常检测	921	2 201

为了验证本文方法在实际生产环境中的适用性，本节在统一硬件环境下对比了本文框架与基线模型MAGIC的训练时长、异常检测时间和内存占用，实验数据集选自E3-Trace数据集，实验结果如表7所示。

本模型通过适度的时空性能投入换取了显著的安全性收益，引入层次化拓扑掩码机制虽然在预处理阶段具有$O(N^2)$时间复杂度，但通过设计掩码预生成与训练解耦的采样机制，将计算压力有效地转移至离线阶段，使得在线训练过程中的渐近时间复杂度保持

在$O(V)$水平,确保了系统在实际部署中的运行效率。而检测时间优于MAGIC主要得益于检测阶段采用的自适应k-NN算法在局部密度计算上的优化,有效避免了MAGIC中由于全图搜索带来的时间波动。

表7 训练与推理时间空间开销对比

Table 7 Comparison of training and inference time and space overhead

项目	Ours	Magic
训练时间开销/s	1 262	922
检测时间开销/s	921	1 023
峰值内存占用/MB	2 201	1 844

5 结论

在实际网络环境中,APT攻击呈现出链式、多阶段及高隐蔽等特征,导致溯源图规模庞大、类型繁多且拓扑结构复杂,如何高效挖掘溯源数据的隐含模式并在预训练阶段有效保留图结构信息,在图表征学习中兼顾局部细节与全局拓扑,是当前APT威胁检测研究中的关键难点。针对上述问题,本文引入层次化拓扑掩码策略,旨在预训练阶段能够保留溯源图关键的拓扑结构信息。其次,引入节点级上下文一致性损失,对编码器施加多视角一致性约束,并使用自适应异常检测算法以适配异构实体密度,区分出不同类型的良性与异常节点。在多个公开数据集上的实验结果表明,所提出方法在保持高检测准确率的同时显著降低了误报率,验证了本方法在复杂溯源图环境和实际APT检测场景中的有效性和实用价值。

参考文献

[1] Aminu M, Akinsanya A, Oyedokun O, et al. A review of advanced cyber threat detection techniques in critical infrastructure: Evolution, current state, and future directions[J]. Iconic Research and Engineering Journals, 2024, 8(2): 74-87.

[2] Nabi N, Rahman M M, Ghosh S K, et al. Machine learning-based anomaly detection for cyber threat prevention[J]. Journal of Primeasia, 2025, 6(1): 1-8.

[3] Li Zhenyuan, Chen Q A, Yang Runqing, et al. Threat detection and investigation with system-level provenance graphs: A survey[J]. Computers & Security, 2021, 106: 102282.

[4] Jia Zian, Xiong Yun, Yuhong Nan, et al. MAGIC: Detecting advanced persistent threats via masked graph representation learning[C]//33rd USENIX Security Symposium. Berkeley: USENIX Association, 2024: 5197-5214.

[5] Bilot T, El Madhoun N, Al Agha K, et al. Graph neural networks for intrusion detection: A survey[J]. IEEE Access, 2023, 11: 49114-49139.

[6] Anjum M M, Iqbal S, Hamelin B. ANUBIS: A provenance graph-based framework for advanced persistent threat detection[C]//Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. New York: ACM, 2022: 1684-1693.

[7] Hossain M N, Milajerdi S M, Wang Junao, et al. SLEUTH: Real-time attack scenario reconstruction from COTS audit data[C]//26th USENIX Security Symposium. Berkeley: USENIX Association, 2017: 487-504.

[8] Milajerdi S M, Eshete B, Gjomemo R, et al. POIROT: Aligning attack behavior with kernel audit records for cyber threat hunting[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2019: 1795-1812.

[9] Milajerdi S M, Gjomemo R, Eshete B, et al. HOLMES: Real-time APT detection through correlation of suspicious information flows[C]//2019 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE, 2019: 1137-1152.

[10] Liu Yushan, Zhang Mu, Li Ding, et al. Towards a timely causality analysis for enterprise security[C]//25th Annual Network and Distributed System Security Symposium. Reston: The Internet Society, 2018.

[11] Hassan W U, Guo Shengjian, Li Ding, et al. NoDoze: Combatting threat alert fatigue with automated provenance triage[C]//26th Annual Network and Distributed System Security Symposium. Reston: The Internet Society, 2019.

[12] Zhong Meihui, Lin Mingwei, Zhang Chao, et al. A survey on graph neural networks for intrusion detection systems: Methods, trends and challenges[J]. Computers & Security, 2024, 141: 103821.

[13] Rehman M U, Ahmadi H, Hassan W U. Flash: A comprehensive approach to intrusion detection via provenance graph representation learning[C]//2024 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE, 2024: 3552-3570.

[14] Cheng Zijun, Lv Qiujian, Liang Jinyuan, et al. Kairos: Practical intrusion detection and investigation using whole-system provenance[C]//2024 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE, 2024: 3533-3551.

[15] Chen Tieming, Dong Chengyu, Lv Mingqi, et al. APT-KGL: An intelligent APT detection system based on threat knowledge and heterogeneous provenance graph learning[J]. IEEE Transactions on Dependable and Secure Computing, 2022.

[16] Zengy J, Wang Xiang, Liu Jiahao, et al. SHADEWATCHER: Recommendation-guided cyber threat analy-

sis using system audit records[C]//2022 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE, 2022: 489-506.

[17] Wang Su, Wang Zhiliang, Zhou Tao, et al. THREATRACE: Detecting and tracing host-based threats in node level through provenance graph learning[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 3972-3987.

[18] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C/OL]//Proceedings of the 5th International Conference on Learning Representations, 2017. https://researchr.org/publication/KipfW17.

[19] Zhu Hongyin, Li Yakun, Liu Luyang, et al. RETRACTED: Pre-training graph autoencoder incorporating hierarchical topology knowledge[J]. Expert Systems with Applications, 2025, 265: 125976.

[20] Pasquier T, Han Xueyuan, Goldstein M, et al. Practical whole-system provenance capture[C]//Proceedings of the 2017 Symposium on Cloud Computing. New York: ACM, 2017: 405-418.

[21] Keromytis A D. Transparent computing engagement 3 data release[EB/OL]. (2018-09-01). https://github.com/darpa-i2o/Transparent-Computing.

[22] The streamspot dataset[EB/OL]. (2022-09-17). https://github.com/sbustreamspot/sbustreamspot-data.

[23] Wget dataset[EB/OL]. (2022-09-17). https://dataverse.harvard.edu/dataverse/unicorn-wget.

[24] Darpa transparent computing dataset[EB/OL]. (2024-10-08). https://github.com/darpa-i2o/Transparent-Computing.

[25] Manzoor E, Milajerdi S M, Akoglu L. Fast memory-efficient anomaly detection in streaming heterogeneous graphs[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 1035-1044.

[26] Han Xueyuan, Pasquier T F J M, Bates A, et al. Unicorn: Runtime provenance-based detector for advanced persistent threats[C]//27th Annual Network and Distributed System Security Symposium. Reston: The Internet Society, 2020.

[27] 郑锐, 汪秋云, 林卓庞, 等. 一种基于威胁情报层次特征集成的挖矿恶意软件检测方法[J]. 电子学报, 2022, 50(11): 2707-2715.
Zheng Rui, Wang Qiuyun, Lin Zhuopang, et al. Cryptojacking malware hunting: A method based on ensemble learning of hierarchical threat intelligence feature[J]. Acta Electronica Sinica, 2022, 50(11): 2707-2715. (in Chinese)

[28] Hamilton W L, Ying R, Leskovec J. Inductive representation learning on large graphs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 1025-1035.

[29] Jiang Baoxiang, Bilot T, El Madhoun N, et al. ORTHRUS: Achieving high quality of attribution in provenance-based intrusion detection systems[C]//34th USENIX Security Symposium. Berkeley: USENIX Association, 2025: 7173-7192.

[30] Paccagnella R, Datta P, Hassan W U, et al. Custos: Practical tamper-evident auditing of operating systems using trusted execution[C]//27th Annual Network and Distributed System Security Symposium . Reston: The Internet Society, 2020.

[31] 冷涛, 蔡利君, 于爱民, 等. 基于系统溯源图的威胁发现与取证分析综述[J]. 通信学报, 2022, 43(7): 172-188.
Leng Tao, Cai Lijun, Yu Aimin, et al. Review of threat discovery and forensic analysis based on system provenance graph[J]. Journal on Communications, 2022, 43(7): 172-188. (in Chinese)

[32] Tang Xianfeng, Li Yandong, Sun Yiwei, et al. Transferring robustness for graph neural network against poisoning attacks[C]//Proceedings of the 13th International Conference on Web Search and Data Mining. New York: ACM, 2020: 600-608.

[33] 仇晶, 陈荣融, 朱浩瑾, 等. 基于溯源图的网络攻击调查研究综述[J]. 电子学报, 2024, 52(7): 2529-2556.
Qiu Jing, Chen Rongrong, Zhu Haojin, et al. A survey of network attack investigation based on provenance graph[J]. Acta Electronica Sinica, 2024, 52(7): 2529-2556. (in Chinese)

[34] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[C/OL]//Proceedings of the 6th International Conference on Learning Representations, 2018. https://researchr.org/publication/VelickovicCCRLB18.

[35] Mikolov T, Chen Kai, Corrado G, et al. Efficient estimation of word representations in vector space[PP/OL]. V3. arXiv (2013-01-16)[2026-03-29]. https://arxiv.org/abs/1301.3781.

[36] Welsh D J A, Powell M B. An upper bound for the chromatic number of a graph and its application to timetabling problems[J]. The Computer Journal, 1967, 10(1): 85-86.

[37] Zhu Hongyin, Zeng Yi, Wang Dongsheng, et al. Brain knowledge graph analysis based on complex network theory[C]//Proceedings of International Conference on Brain Informatics and Health. Cham: Springer, 2016: 211-220.

[38] 李忠, 靳小龙, 庄传志, 等. 面向图的异常检测研究综述[J]. 软件学报, 2021, 32(1): 167-193.
Li Zhong, Jin Xiaolong, Zhuang Chuanzhi, et al. Overview on graph based anomaly detection[J]. Journal of Software, 2021, 32(1): 167-193. (in Chinese)

[39] Yang Fan, Xu Jiacen, Xiong Chunlin, et al. PROGRAPHER: An anomaly detection system based on provenance graph embedding[C]//32nd USENIX Security Symposium. Berkeley: USENIX Association, 2023: 4355-4372.
[40] Zhu Hongyin, Tiwari P, Zhang Yazhou, et al. SwitchNet: A modular neural network for adaptive relation extraction[J]. Computers and Electrical Engineering, 2022, 104: 108445.
[41] Xu Keyulu, Hu Weihua, Leskovec J, et al. How powerful are graph neural networks [JC/OL].]//Proceedings of the 7th International Conference on Learning RepresentationsarXiv preprint arXiv: 1810. 00826, 2018. https://researchr.org/publication/XuHLJ19.
[42] Hou Zhenyu, Liu Xiao, Cen Yukuo, et al. GraphMAE: Self-supervised masked graph autoencoders[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2022: 594-604.
[43] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.
[44] Shannon C E. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27(3): 379-423.
[45] Park J, Lee M, Chang H J, et al. Symmetric graph convolutional autoencoder for unsupervised graph representation learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 6518-6527.
[46] Wang Xiao, Liu Nian, Han Hui, et al. Self-supervised heterogeneous graph neural network with co-contrastive learning[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM, 2021: 1726-1736.
[47] Wu Wenhan, Hua Yilei, Zheng Ce, et al. Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition[C]//2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). Piscataway: IEEE, 2023: 224-229.
[48] Papadimitriou S, Kitagawa H, Gibbons P B, et al. LOCI: Fast outlier detection using the local correlation integral[C]//Proceedings 19th International Conference on Data Engineering (Cat. No. 03CH37405). Piscataway: IEEE, 2003: 315-326.
[49] Van Der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605.
[50] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2017: 6000-6010.

作者简介

刘尚东 男,1979年10月出生于甘肃省永靖县。现为南京邮电大学副教授、硕士生导师,主要方向为网络空间安全、人工智能、大数据等。
E-mail: lsd@njupt.edu.cn

杨易润 男,2001年12月出生于江苏省无锡市。现为南京邮电大学计算机学院、软件学院、网络空间安全学院硕士研究生,主要研究方向为APT检测与溯源。
E-mail: 1024041132@njuput.edu.cn

王一诺 女,2003年11月出生于江苏省连云港市。现为南京邮电大学计算机学院、软件学院、网络空间安全学院硕士研究生,主要研究方向为网络安全态势感知。
E-mail: 1025040932@njupt.edu.cn

杜宏煜 男,2000年11月出生于江苏省盐城市。现为南京邮电大学计算机学院、软件学院、网络空间安全学院博士研究生,主要研究方向为漏洞检测和网络空间威胁感知。
E-mail: 2024040408@njupt.edu.cn

汪文博 男,1998年6月出生于江苏省苏州市。现为南京邮电大学计算机学院、软件学院、网络空间安全学院在读博士,主要研究方向为计算机网络安全和机器学习等。
E-mail: wangwenbo@njupt.edu.cn

季一木 男,1978年9月出生于安徽省无为市。现为南京邮电大学计算机学院教授,博士生导师,主要研究方向为人工智能、云计算和大数据安全等。
E-mail: iym@njupt.edu.cn