

doi:10.13682/j.issn.2095-6533.2026.05.011

融合异构图网络与大语言模型推理的恶意域名检测方法

王闪闪^{1,2}, 马昊^{1,2}, 李群³, 潘岳^{1,2}, 陈贞翔^{1,2}

(1. 济南大学 信息科学与工程学院, 山东 济南 250022; 2. 济南大学 山东省泛在智能计算重点实验室(筹), 山东 济南 250022;
3. 济宁职业技术学院 电子信息工程系, 山东 济宁 272007)

摘要: 针对域名系统中基于黑名单与静态特征的方法难以有效应对复杂攻击手段的问题, 提出一种融合异构图网络与大语言模型推理的恶意域名检测方法。该方法先构建包含域名、互联网协议、规范名称及注册者等实体的异构信息网络, 再在此基础上构建基于元路径的多头图注意力网络, 动态量化不同语义空间的邻居权重, 并通过多层聚合挖掘节点的结构化关联。最后, 引入大语言模型作为推理引擎, 计算结构与语义特征的匹配度, 输出决策置信度评分, 实现对潜在恶意域名的推理与分类, 并最终根据证据生成判决报告。实验结果表明, 与主流图模型及纯文本方法对比, 所提方法对恶意域名检测的准确率达到 92.58%, 冷启动检测准确率为 79.22%, 证实了该方法的有效性。

关键词: 域名检测; 异构信息网络; 图注意力网络; 大语言模型; 协同推理

中图分类号: TN393.08; TP183

文献标志码: A

文章编号: 2095-6533(2026)05-0115-11

Malicious domain detection method combining heterogeneous graph networks and large language model reasoning

WANG Shanshan^{1,2}, MA Hao^{1,2}, LI Qun³, PAN Yue^{1,2}, CHEN Zhenxiang^{1,2}

(1. School of Information Science and Engineering, University of Jinan, Jinan 250022, China;

2. Shandong Key Laboratory of Ubiquitous Intelligent Computing, University of Jinan, Jinan 250022, China;

3. Department of Electronic Information Engineering, Jining Polytechnic, Jining 272007, China)

Abstract: Traditional methods for malicious domain detection based on blacklists and static features in the domain name system struggle to effectively address complex attack techniques. To address this issue, a malicious domain detection method integrating a heterogeneous graph network with large language model reasoning is proposed. Firstly, a heterogeneous information network including domain names, internet protocol, canonical names, and registrants is constructed. On this basis, a meta-path-based multi-head graph attention network is constructed to dynamically quantify the weights of neighboring nodes across different semantic spaces, and to mine the structured associations among nodes through multi-layer aggregation. Finally, a large language model is introduced as the inference engine to measure the matching degree between structural and semantic features, output a decision confidence score, perform reasoning and classification of potentially malicious domain names, with a final evidence-based judgment report generated. The proposed method was evaluated against mainstream graph-based and text-based methods. Experimental results indicate that the proposed method achieves an accuracy of 92.58% in malicious domain detection and an accuracy of 79.22% in cold-start detection, which demonstrate

收稿日期: 2026-01-14

基金项目: 国家自然科学基金项目(62502182); 山东省自然科学基金项目(ZR2023QF096); 泰山学者工程项目(tstp20240828, tsqzn20221146)

通信作者: 陈贞翔, zxchen@ujn.edu.cn

引文格式: 王闪闪, 马昊, 李群, 等. 融合异构图网络与大语言模型推理的恶意域名检测方法[J]. 西安邮电大学学报, 2026, 31(5): 115-125.

WANG Shanshan, MA Hao, LI Qun, et al. Malicious domain detection method combining heterogeneous graph networks and large language model reasoning[J]. Journal of Xi'an University of Posts and Telecommunications, 2026, 31(5): 115-125.

the effectiveness of the proposed method.

Keywords: domain name detection; heterogeneous information network; graph attention network; large language model; collaborative reasoning

作为互联网的基础设施,域名系统(Domain Name System, DNS)承担着域名解析的关键职能,其流量数据也因此成为检测恶意域名的核心情报来源^[1]。然而,正是鉴于DNS在网络通信中的不可或缺性,其也频频成为网络攻击者进行恶意攻击的首选。在网络钓鱼、恶意软件等各类网络犯罪活动中,域名都扮演着不可或缺的角色^[2]。例如,攻击者可以通过控制域名维持命令与控制信道的稳定性、利用虚假域名诱骗用户泄露敏感信息等^[3]。因此,有效检测恶意活动所涉及的域名始终是网络安全研究的热点^[4]。

随着安全防御技术的进步,攻击者也在不断演进其规避策略。早期的网络攻击往往使用固定的域名和互联网协议(Internet Protocol, IP)地址,这使得基于规则的方法十分有效^[5-6]。然而,为了对抗黑名单机制,攻击者通过快速轮换域名背后的IP地址记录,使得单一IP的封锁失效。为此,研究人员通过分析被动DNS^[7-8]、顶级域名服务器流量^[9]和非存续域名(Non-Existent Domain, NXDomain)流量^[10-11]中字符模式、间隔等特征构建分类器,为未标记域名分配信誉分数^[12]。然而,尽管这些方法在静态评估中表现出色,但基于特征的系统很容易被对手设计的攻击所规避。如文献[13]指出攻击者能够混淆域名的字符分布特征,以逃避检测。文献[14]进一步证明了攻击者还会通过操控请求间隔或生存时间(Time To Live, TTL)值扰动时间模式,从而降低模型的识别效果。相比之下,域名之间的关联关系则难以被攻击者刻意伪造或频繁改变。因此,将DNS场景建模为图,挖掘域之间的关联,可对域名进行检测^[15-16]。但受同构网络模型限制,这类方法既无法充分挖掘DNS的复杂关联、造成大量信息闲置,又仅能处理静态场景,而难以应用于现实动态环境。为了克服孤立分析的缺陷,DeepDom模型^[17]通过构建异构信息网络(Heterogeneous Information Network, HIN),将客户端、域名、IP地址等多种实体及其查询、解析、注册等复杂关系进行统一建模。但这类方法高度依赖关联特征,面对新恶意域名易因关联稀疏出现冷启动问题,且深度学习模型普遍缺乏可解释性^[18],限制了其在安全运维中的效能,需明确判定依据与恶意类别才能制定针对性策略。

结合上述问题与DNS的动态特性,理想的恶意域名检测框架应满足:能够处理新域名并解决冷启动问题;在保证检测率的同时尽可能扩大检测覆盖范围;可对恶意域名进行细粒度分类,并深入分析其恶意成因。因此,本文拟提出一种融合异构图网络与大

语言模型(Heterogeneous Graph Attention Network-Large Language Model, HGAT-LLM)的恶意域名检测方法。该方法通过分析恶意域名的结构特征,将图注意力网络(Graph Attention Network, GAT)作为证据检索器,在网络拓扑中提取出与恶意域名相关的结构化路径。然后将这些路径转化为证据语言,输入到大语言模型(Large Language Model, LLM)中作为其推理判断的证据。LLM借助证据和域名的语义特征,对域名的恶意性以及恶意类型进行综合判断,并将恶意的原因呈现出来。最后,与当前同类型工作进行对比,测试每条元路径对HGAT-LLM方法的贡献程度,以及其冷启动效果。

1 域名特征及差异分析

选取公开白名单Alexa Top1M(Amazon Web Services)组成良性域名,共75万条,12万条恶意域名来自公共黑名单Kaggle Malicious URLs Dataset和爬虫收集的数据。为验证细粒度推理效果,从中随机选取1万余条域名,补充其IP、规范名称(Canonical Name, CNAME)记录、所属网段、注册者及域名类型信息,构建实验数据集。经分析发现域名呈现类别聚集特征,据此将样本划分为4个互斥类别:良性、恶意软件、钓鱼网站和未明确分类的其他恶意域名,分别占65.4%,14.5%,18.0%和2.1%。实验数据的域名类型分布情况如图1所示。

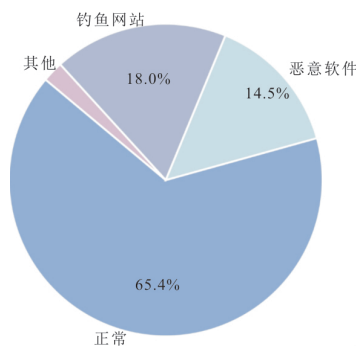


图1 域名类型分布情况

尽管攻击者会采用灵活手段规避检测,但受限于大规模基础设施的高昂维护成本,其网络资源(如IP地址)的复用难以避免。已有研究表明,托管过恶意域名的IP所托管的其他域名具有较高潜在风险^[19-20],且同一IP托管的域名往往是相同类型,域名与IP之间存在强关联性。数据集中恶意域名和良性域名所属IP分布情况如图2所示。

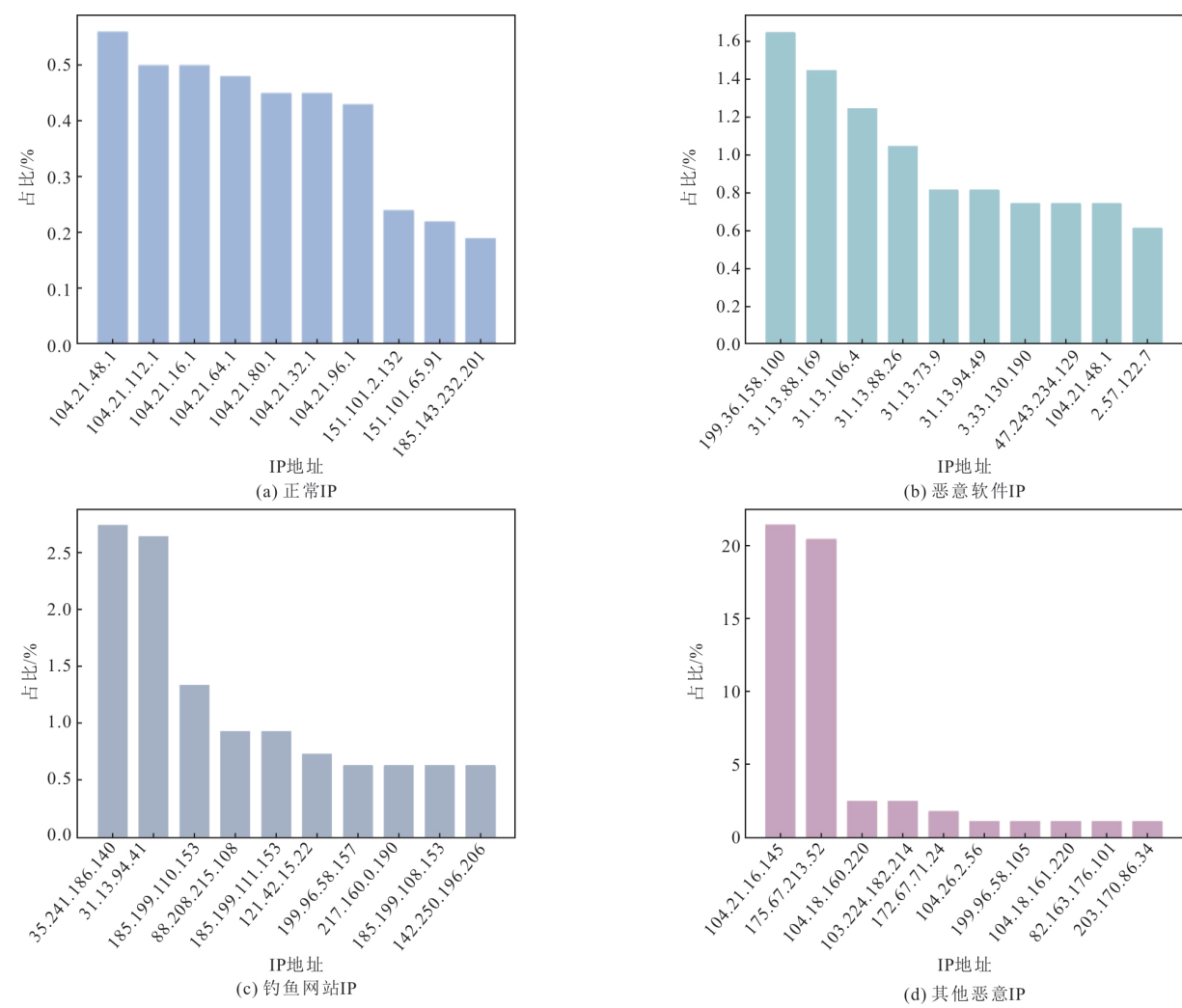


图 2 域名 IP 托管情况

通过分析域名注册信息查询服务 (Who Is, WHOIS) 注册信息,发现攻击者倾向于批量注册价格低廉或审核宽松的注册商,从而根据注册者注册域名的模式和关联特征检测恶意域名^[21]。恶意域名注册者注册分布情况如图 3 所示。

由图 3 可以看出,注册者倾向于注册多个恶意域名,这也验证了挖掘其关联的直觉。

2 HGAT-LLM 恶意域名检测方法

HGAT-LLM 恶意域名检测方法共包含 3 个主要模块:异构信息网络构建与元路径挖掘、基于元路径的图注意力网络,以及基于检索增强的 LLM 协同推理引擎。异构信息网络构建与元路径挖掘模块将 DNS 相关数据构建包含多种实体关系的 HIN 以对 DNS 场景进行形式化建模,并根据数据分析中攻击者资源聚合的特点,提取 4 种类型的元路径,从不同角度展示域名之间的深层关联。基于元路径的图注意力网络模块通过元路径引导图注意力网络训练、路径评分和 Top-K 筛选,得到结构化证据路径。基于检索增强的 LLM 协同推理引擎模块则对证据路径进行词法特征分析并构建提示词,利用 LLM 结合结构证据和域名语义特征判断该域名是否恶意,从而得出细粒度的分类结果并

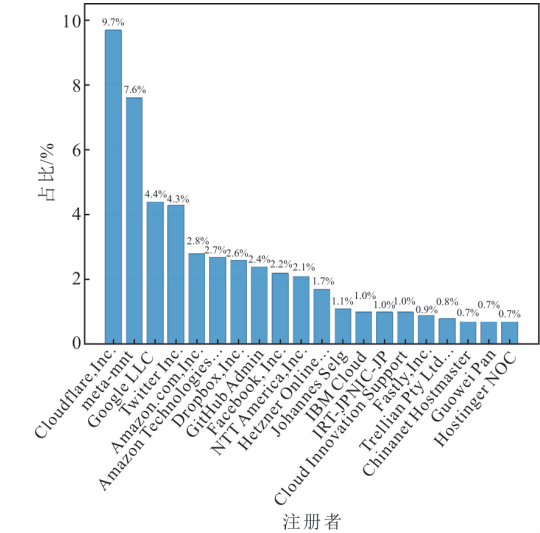


图 3 恶意域名注册者注册分布情况

提供可解释的证据判决报告。HGAT-LLM 恶意域名检测方法架构如图 4 所示。

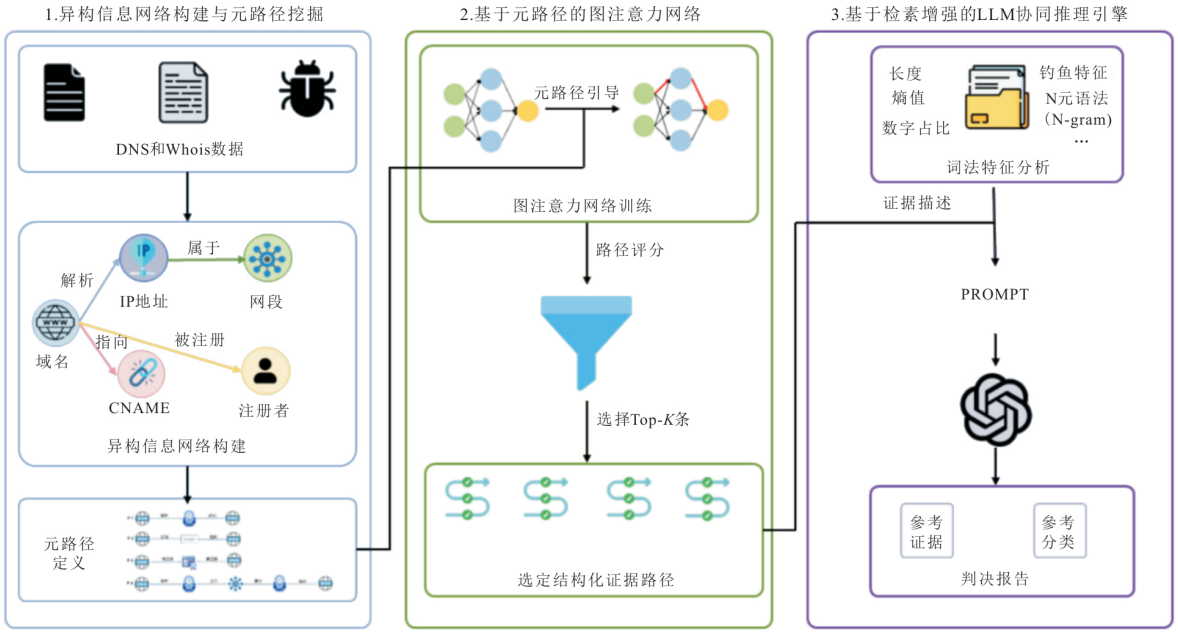


图 4 HGAT-LLM 恶意域名检测方法架构

2.1 异构信息网络构建与元路径挖掘

异构信息网络构建与元路径挖掘模块基于域名特征及差异分析,从域名与攻击者视角将非结构化 DNS 日志转换为结构化 HIN,并完成数据标注,为后续图模型训练做准备。根据域名的关联关系, DNS 基础设施实体关系如图 5 所示,其揭示了域名在注册、查询及解析全生命周期的可追溯操作痕迹。

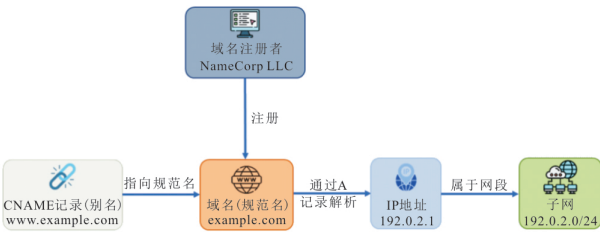


图 5 DNS 基础设施实体关系

基于对上述交互行为的分析,从原始 DNS 数据流中抽取出域名、IP 地址、注册者名称、网段以及 CNAME 记录共 5 种核心实体。为刻画实体之间的交互逻辑,构建实体间的正向关系,包括域名解析 IP、IP 属于网段及域名被注册者注册等,并通过邻接矩阵表示之间的关系。

1) R_1 (域名 $\xrightarrow{\text{解析}}$ IP 关系)。构建域名-IP 解析矩阵 $\mathbf{M}_{\text{res}}$,若域名 i 解析到 IP 地址 j ,则 $\mathbf{M}_{\text{res}}(i, j) = 1$,否则为 0。

2) R_2 (域名 $\xrightarrow{\text{指向}}$ CNAME 记录关系)。生成域名 CNAME 指向矩阵 $\mathbf{M}_{\text{cname}}$,若域名 i 指向 CNAME 记录 j ,则 $\mathbf{M}_{\text{cname}}(i, j) = 1$,否则为 0。

3) R_3 (域名 $\xrightarrow{\text{被注册}}$ 注册者关系)。构建域名-注册者注册矩阵 $\mathbf{M}_{\text{reg}}(i, j)$,若域名 i 由注册者 j 注册,则 $\mathbf{M}_{\text{reg}}(i, j) = 1$,否则为 0。

4) R_4 (IP $\xrightarrow{\text{属于}}$ 网段关系)。生成 IP-网段归属矩阵 $\mathbf{M}_{\text{net}}$, $\mathbf{M}_{\text{net}}(i, j) \in \{0, 1\}$ 表示 IP 地址 i 是否属于网段 j 。

上述 4 种邻接矩阵联合定义了 HIN 的拓扑结构,隔离了不同语义的边的同时也构成了后续 GAT 中多通道消息传递与元路径计算的基础。下面构建对应的反向关系,这一双向拓扑设计旨在支持 GAT 进行双向消息传递,从而捕获更丰富的结构化特征。同时,确立用于监督学习的基准真值至关重要。通过依托安全检测平台 VirusTotal 应用程序编程接口 (Application Programming Interface, API)对域名进行标签获取与清洗。首先,批量查询目标域名的历史扫描记录,获取数 10 家主流安全引擎的判定结果。当超过 3 家及以上的引擎将其标记为恶意时,该域名才被判定为异常;全部引擎标记为良性时,则判定为良性。对于 1~2 家引擎报警的域名,考虑到误报风险,不予纳入。对异常域名提取威胁分类关键字,标准化映射为恶意软件、钓鱼网站和其他恶意域名。最终所有样本归入良性、恶意软件、钓鱼网站和其他恶意域名,并映射为数值索引 $\{0, 1, 2, 3\}$ 作为 GAT 训练的监督信号。

基于实体关系,将 DNS 场景形式化建模为 HIN。HIN 形式上可定义为 $G = (V, E, T, X)$, V 表示节点集合, E 表示边集合, $T = \langle T_V, T_E \rangle$ 表示

网络模式,由节点类型集合 T_V 与边类型集合 T_E 构成, $X = \{x_i | v_i \in V\}$ 表示所有节点的特征集合。通过映射函数 $\phi: V \rightarrow T_V$ 和 $\varphi: E \rightarrow T_E$, 网络中的每个节点和每条边分别映射到对应的类型空间。当且仅当网络满足 $|T_V| + |T_E| > 2$ 时,称 G 为异构信息网络。 T 作为 G 的网络模式,描述了 HIN 中节点与边类型构成的语义结构。在恶意域名检测场景, HIN 可实例为 $T_V = \{\text{域名}, \text{IP}, \text{注册者}, \text{网段}, \text{CNAME}\}$, $T_E = \{\text{解析}, \text{指向}, \text{被注册}, \text{属于}\}$ 。构建图时,映射函数 ϕ 将真实世界中提取的实体(如 example.com)赋予相应的类型标签,而映射函数 φ 将具体的解析动作识别为边类型,从而严格遵循网络模式 t 的语义约束。特征集合 X 显式地表达了各类型节点自身的固有属性向量。最终,将多源 DNS 关系数据形式化建模为异构信息网络 G , 为后续元路径的邻居聚合与特征提取定义了合法的类型约束。

2.2 基于元路径的图注意力网络

在异构信息网络中,元路径被广泛用于刻画节点类型之间的语义连接^[22]。假设 $R_1, R_2, \dots, R_L$ 表示一系列允许的关系类型,其中 L 表示路径长度,则元路径 P 可被定义为其复合关系,即 $P = R_1 R_2 \dots R_L$,从某一类型 V_1 出发,遵循 R_1 ,经过 V_2 ,继续通过 $R_2 \dots$ 最终到达 V_{L+1} 。该元路径为 $V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \dots \xrightarrow{R_L} V_{L+1}$,若上下文无歧义,也可简写为 $V_1 V_2 \dots V_{L+1}$ 。当 V_1 与 V_{L+1} 类型相同时,该元路径称为对称元路径。HGAT-LLM 能够根据元路径捕捉 HIN 中不同类型节点之间的高阶语义关系,元路径示例如图 6 所示。

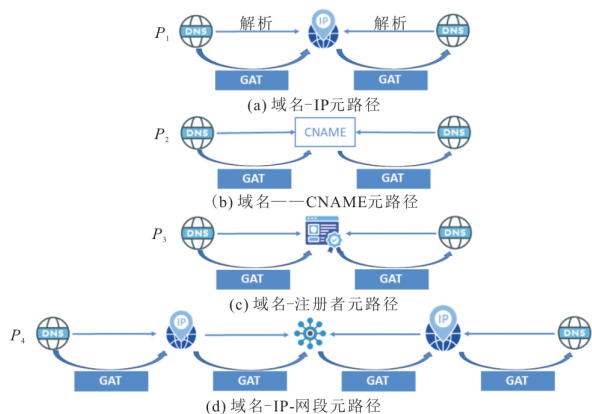


图 6 元路径示例

图 6(a)中 P_1 揭示了攻击者资源聚合特征,互联网中的 IP 资源相对稳定,一段时间内解析到同一 IP 地址的域名往往属于同一类别;图 6(b)中 P_2 描述了良性域名的 CNAME 域名不太可能是恶意的,

反之亦然;图 6(c)中 P_3 表示若一个域名的注册账户还注册了其他恶意域名,则会增加该域名的可疑性程度。图 6(d)中 P_4 表示若两个域名解析到的 IP 地址属于同一网段,则也会增加二者的关联。因此,选择这 4 种关系作为模型参考的元路径。

为实现端到端的恶意域名检测, HGAT-LLM 摒弃了依赖人工定义路径权重乘积的启发式评分规则,构建了基于元路径的注意力网络 MetaPath-GAT。该网络可直接在异构图结构上进行消息传递,其核心在于利用图注意力机制动态量化邻居节点在不同语义空间下的重要性,自动学习不同路径对恶意性判定的贡献权重。具体而言, HGAT-LLM 首先将长距离的元路径分解为异构图中具体的单跳邻接关系。以图 6(a)中元路径 P_1 为例, HGAT-LLM 将其分解为域名 $\xrightarrow{\text{解析}}$ IP 和 IP $\xrightarrow{\text{被解析}}$ 域名两个连续的原子步骤,并利用图注意力机制在这些局部关系上分别计算注意力系数^[23]为

$$\alpha_{i,j} = \frac{e^{\text{LeakyReLU}(\mathbf{a}^T [\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_j])}}{\sum_{m \in \mathcal{N}_i} e^{\text{LeakyReLU}(\mathbf{a}^T [\mathbf{W}\mathbf{h}_i \parallel \mathbf{W}\mathbf{h}_j])}} \quad (1)$$

式中: $\mathbf{h}_i$ 和 $\mathbf{h}_j$ 分别表示节点 i 和邻居节点 j 的特征向量; $\mathbf{W}$ 表示可训练的权重矩阵,用于对节点特征进行线性变换; $\mathbf{a}$ 为注意力向量; $\parallel$ 表示拼接操作; LeakyReLU 为非线性激活函数; m 表示节点 i 的邻居集合 $\mathcal{N}_i$ 中的任一邻居节点。

注意力系数 α_{ij} 反映了在当前上下文语境下,特定的关联实体对于判定目标域名性质的关键支撑作用。以域名 $\xrightarrow{\text{解析}}$ IP 这一单跳关系为例,向量 $\mathbf{h}_i$ 代表目标域名节点的高维特征表征(如长度、熵值等词法特征),而 $\mathbf{h}_j$ 代表其关联的 IP 节点的特征表征。集合 $\mathcal{N}_i$ 不再是泛化的全图邻居,而是特定关系下的邻居子集(如域名 i 历史上曾解析过的所有 IP 地址集合)。 α_{ij} 量化了在域名 i 解析过的众多 IP 中, IP 对 j 该域名恶意的决定性贡献度。由此, HGAT-LLM 能够从复杂的 DNS 关系拓扑中,自动筛选并聚焦于最具威胁指示性的关键实体。

为捕捉不同维度的语义特征, HGAT-LLM 在卷积层中引入多头注意力机制,注意力头 $K=4$ 个表示子空间并行学习节点间的关联强度并融合起来,让模型更加关注不同位置间的关系,捕捉长距离依赖^[24]。多头聚合后的节点特征更新规则^[23]为

$$\mathbf{h}'_i = \parallel_{k=1}^K \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^{(k)} \mathbf{W}^{(k)} \mathbf{h}_j \right) \quad (2)$$

式中: $\alpha_{ij}^{(k)}$ 表示在第 k 个注意力头下,邻居 j 对节点 i 的注意力系数, $\alpha_{ij}^{(k)}$ 越大,说明邻居 j 是 i 的关键证据;

$W^{(k)}$ 为权重矩阵,通过拼接操作融合 K 个注意力头的特征,形成更新后的节点 h_i' 。

通过堆叠多层图注意力层, HGAT-LLM 可沿元路径逐层传递与聚合,无需显式计算全图路径矩阵即可捕获整条元路径的高阶语义特征。聚合后,域名 i 在第 m 条特定的元路径 Φ_m 下,均会获得一个专属的高阶语义表征 $z_i^{\Phi_m}$ 。针对不同元路径的恶意性判定效力差异,引入语义级注意力机制自动学习第 m 条元路径的权重系数 β_m ,并通过加权求和的方式,将来自 p 条不同元路径的特征深度融合,生成节点 i 最终的高维嵌入向量为^[25]

$$z_i = \sum_{m=1}^p \beta_m \cdot z_i^{\Phi_m} \quad (3)$$

最后,高维嵌入向量 z_i 通过一个线性分类器被映射为 4 个类别的非归一化置信度分数,通过 Softmax 函数转化为最终的预测概率表示为^[26]

$$\hat{y}_i = \text{Softmax}(W_{\text{cls}} \cdot z_i + b) \quad (4)$$

式中:参数 W_{cls} 和 b 构成线性分类器,负责将高维特征压缩映射到类别空间。 $\hat{y}_i$ 值越高,表明该域名属于对应恶意类别的可能性越大。

在训练阶段,为解决真实场景中恶意样本稀缺导致的类别不平衡问题, HGAT-LLM 采用加权交叉熵损失函数^[27]将训练集中类别 c 的样本数量 N_c 的倒数作为类别权重 $w_c \propto 1/N_c$,并将损失函数定义为

$$\mathcal{L} = - \sum_{i \in \mathcal{D}_{\text{train}}} \sum_{c=1}^C w_c \cdot y_{i,c} \cdot \log(\hat{y}_{i,c}) \quad (5)$$

式中: $\mathcal{D}_{\text{train}}$ 表示训练样本集合; C 表示域名种类个数,此处为 4; $\hat{y}_{i,c}$ 表示模型预测样本 i 属于类别的概率, $y_{i,c}$ 为真实标签(0 或 1)。通过引入权重系数 w_c ,强制模型在优化过程中关注少数类样本。

HGAT-LLM 使用 Adam(Adaptive Moment Estimation)优化器(学习率为 0.01)对模型进行 200 轮的训练。为防止过拟合并确保泛化能力,将数据集划分为独立的训练集、验证集与测试集。在训练过程中引入早停机制:每轮迭代后,实时监控其在验证集上的 F1 分数,仅当 F1 分数指标达到历史新高时才保存当前的模型参数,以此作为后续 LLM 协同推理与证据链提取的最优基座模型。

2.3 基于检索增强的 LLM 协同推理引擎

在 HGAT-LLM 中, LLM 并非简单的分类器,而是语义推理与决策仲裁的核心组件。MetaPathGAT 虽然能够有效捕捉异构网络拓扑关联并量化恶意概率,但纯图嵌入表示往往面临双重局限,即难以充分解析域名的词法特征,且由于对显式关联的高度依赖,导致其在检测新恶意域名时会陷入冷

启动困境。因此,引入 LLM 作为语义推理引擎,旨在通过结构化证据与词法语义的深度融合,弥补单一图模型的局限性。

不同于直接利用 LLM 进行预测, HGAT-LLM 采用检索增强生成思想(Retrieval Augmented Generation, RAG),提示词构建阶段采用动态拓扑证据映射策略,将 MetaPathGAT 筛选的 4 条高置信度证据链及其预测标签,作为外部知识结构化地嵌入到提示词中。为量化决策的可信度,引入多维证据一致性置信度建模机制。在执行思维链(Chain-of-Thought, CoT)推理时,引导 LLM 遵循 3 步逻辑验证:1)验证结构化证据的可靠性;2)分析语义特征;3)计算拓扑证据支持度 S_t 与语义特征支持度 S_s 的匹配程度,量化输出决策置信度评分

$$Q = \omega \cdot \text{sim}(S_t, S_s) + (1 - \omega) \cdot \hat{y}_{\text{max}} \quad (6)$$

式中: S_t 由 MetaPathGAT 输出的四分类概率 $\hat{y}_i$ 定义; S_s 通过提取 LLM 输出层的词元概率并进行映射得到; $\text{sim}(S_t, S_s)$ 为采用余弦相似度计算 S_t 与 S_s 的空间夹角,量化两模态证据在逻辑层面的契合程度; ω 为平衡权重系数,旨在平衡模态对齐程度与图模型原始置信度; $\hat{y}_{\text{max}}$ 表示图结构模型对域名预测的初始概率。

当 Q 值较高时,表明两种模态证据达成一致;当两者冲突时, LLM 作为推理裁决器,根据评估结果修正图模型的结构化偏差。若 Q 低于预设阈值,模型将触发置信度门控的拒判机制,输出选项“E. unknown”而非强行分类。针对冷启动问题, HGAT-LLM 设计了鲁棒的检测机制。通过 LLM 的预备知识库同时结合域名的语义特征对域名进行检测。这种置信度量化策略有效降低了模型幻觉风险^[28],提升了判决的严密性。LLM 提示词示例如图 7 所示。

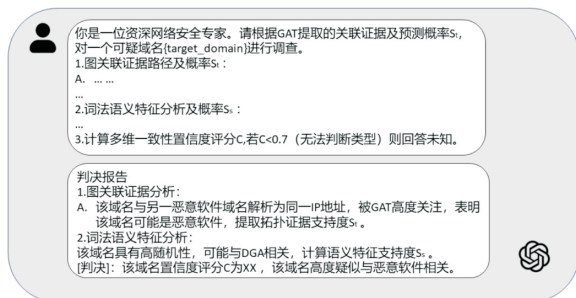


图7 LLM 提示词示例

3 实验结果与分析

3.1 实验环境及参数

实验在一台高性能计算服务器上完成,服务器

搭载 8 张 NVIDIA Tesla V100S-PCIE 显卡(单卡显存 32 GB,总显存 256 GB)。协同推理引擎选用 Qwen2.5-32B 大语言模型,单样本平均推理时间约为 0.85 s,整体吞吐量约为 6.5 样本/s,推理期间图形处理器利用率稳定在 80% 以上。然而,考虑到 DNS 检测的高并发特性,当前秒级的推理延迟难以满足实时监测任务,轻量化模型压缩方案将是未来的重要研究方向。

为验证模型泛化能力并规避数据泄露风险,采用分层抽样策略,按 7 : 1 : 2 将数据集划分为互斥的训练集 V_{tr} 、验证集 V_{val} 与测试集 V_{te} 。针对异构图中易发生的跨集信息泄露问题,实验采用归纳式学习范式,并实施子图物理隔离。定义训练子图为 $g_{tr}=(V_{tr},\epsilon_{tr})$,其中 $\epsilon_{tr}=\{(u,v)\in\epsilon|u,v\in V_{tr}\}$ 。模型训练阶段构建仅包含 V_{tr} 内部连边的训练计算图,强制掩码并移除所有跨集的边,确保消息传递不接触验证集和测试集的信息,从架构层面消除前瞻偏差。以下实验结果均无人工干预。

3.2 图结构分析

为刻画 DNS 场景下多实体间的关联关系,图 8 展示了包含域名、IP、注册者、网段与 CNAME 等实体异构关系。鉴于全图数据规模较大,难以直观呈现,随机采样 30% 节点及其邻接关系构建子图。域名、IP、注册者、网段与 CNAME 分别为红、紫、蓝、黄和绿。对图 8 拓扑结构的分析揭示两类关键特征:首先,域名与 IP 地址之间呈高密度的多对多映射关系。既存在单 IP 托管多域名的资源复用现象,也包含单域名解析至多 IP 的分布模式,这类底层关联是捕捉攻击团伙特征的基础。其次,CNAME 记录虽稀疏但置信度高,域名通过同一 CNAME 节点建立关联时,二者类别一致的概率显著提升,表明 CNAME 关联更能反映域名背后的实质性业务同源性。

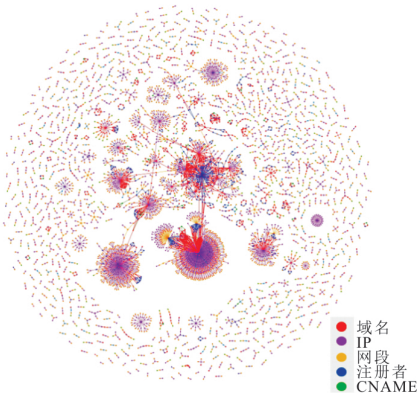


图 8 实体异构关系

3.3 HGAT-LLM 指标表现

为全面评估 HGAT-LLM 性能,实验除关注检

测准确率外,重点关注精准率、F1 得分、召回率以及受试者工作特征曲线下面积(Area Under the Receiver Operating Characteristic Curve, AUC)值。在恶意域名这类高隐蔽性少数类检测场景中,这些指标能更敏锐地反映模型降低误报、漏报的实际效能,客观评估其对威胁的检测能力。

在独立测试集上进行二分类的性能评估, HGAT-LLM 分类性能稳健:整体准确率为 92.58%,恶意域名识别的精确率、召回率、F1 分数分别达 93.64%、94.48%、94.06%,AUC 值为 94.30%。这说明在类别不平衡场景下, HGAT-LLM 可通过有效聚合底层异构网络拓扑特征,维持对正负样本的高效判别与稳健排序能力。

3.4 HGAT-LLM 细粒度分类表现

尽管 HGAT-LLM 在二分类检测中表现良好。但为测试模型在恶意域名细粒度分类场景下的潜力,测试了其四分类性能。混淆矩阵如图 9 所示,可以看出, HGAT-LLM 对良性域名的召回率为 93.61%,表明其误报风险较低。恶意软件、钓鱼网站及其他恶意域名误判为良性的比例分别为 4.97%、4.31%、6.35%,说明多数恶意域名被成功检出。但在恶意软件与钓鱼网站之间容易混淆,误判比分别为 15.99% 和 17.22%。即便如此, HGAT-LLM 在 3 类恶意域名上的召回率仍分别达到 67.88%、72.08% 与 77.78%,验证了其具备一定的细粒度恶意域名分类能力。

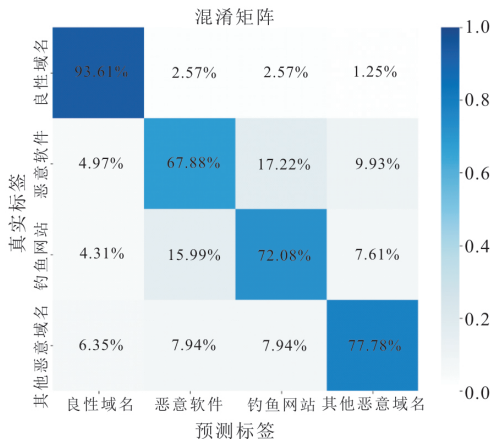


图 9 混淆矩阵

3.5 同类相关工作对比

为验证所提方法在恶意域名检测任务中的综合优势,选取以下 3 种代表不同技术路线的主流方法作为基线模型进行对比实验:文献[7]方法利用被动 DNS 流量,分析域名的查询行为与时间模式检测恶意域名,提取时间特征、DNS 响应、TTL 值和域名拼写 4 类统计特征,将这些特征输入机器学习

分类器,从而自动识别出有恶意行为的域名;文献[17]方法通过构建域名的异构关联图,利用可扩展的异构图卷积网络(Scalable and Heterogeneous Graph Convolutional Network, SHetGCN)聚合邻居节点的关键信息,实现对恶意域名的识别;文献[24]方法将域名视为字符序列,通过嵌入层将离散字符转换为包含位置信息的稠密向量,利用一维卷积层捕捉字符间的局部关联特征结合残差网络(Residual Network, ResNet)增强特征传递效率,最后用多头注意机制捕获关键信息,直接输出该域名的恶意概率。

在相同实验环境下,4种方法在域名二分类任务结果上精准率、召回率和F1得分指标对比如表1所示。

表 1 4种方法指标对比				%
方法	精准率	召回率	F1 得分	
所提方法	93.64	94.48	94.06	
文献[7]	89.87	81.19	85.30	
文献[17]	90.62	96.43	93.43	
文献[24]	95.65	86.51	90.85	

由表1可以看出,所提方法在综合检测性能上优于各对比方法。具体而言,相比于基于特征机器学习的文献[7]方法,所提方法的F1得分实现了8.76%的显著提升。这一性能跨度验证了单靠人工定义的统计特征难以检测复杂的攻击模式,而融合异构图结构信息能有效捕捉攻击者在基础设施层面的隐蔽关联,证明了图结构关联的必要性。与基于图结构DeepDom的文献[17]方法相比,所提方法的精准率提升了3.02%。DeepDom虽然擅长利用关联挖掘潜在威胁,但容易产生误报,而HGAT-LLM引入大语言模型作为协同推理引擎,利用LLM强大的语义理解能力对可疑目标进行二次逻辑研判,有效剔除了误报样本,提升了检测的纯度。尽管端到端深度学习的文献[24]方法在精确率上略有优势,但召回率仅为86.51%,低于所提方法,这表明仅依靠字符特征容易导致对伪装性强的恶意域名发生漏报^[29]。相比之下,所提方法关注微观字符特征并利用宏观图谱关联,构筑了更严密的检测防线。

为评估图结构跨信息集泄露对性能的影响,设计归纳式与转导式学习的对比实验,结果如表2所示。相比转导式学习,归纳式学习实施了物理隔离,可确保消息传递被限制在子图内部。可以看出,在实施子图物理隔离之后,各项指标出现了约1%的回落,表明归纳式范式的必要性。

表 2 不同范式对比实验结果				%
范式	精准率	召回率	F1 得分	
归纳式学习	93.64	94.48	94.06	
转导式学习	94.47	95.82	95.14	

综上所述,所提方法成功克服了单一模态方法的局限性,有效解决了传统方法在应对动态、长尾恶意域名时漏报率高的痛点,为构建智能化网络防御体系提供了有力支撑。

3.6 消融实验

为测试HGAT-LLM相较单GAT(Only GAT)模型和单LLM(Only LLM)模型带来的提升,设计消融实验对域名二分类和四分类进行测试。3种模型检测准确率对比如表3所示。可以看出,HGAT-LLM双模型推理对恶意域名检测准确率为92.58%,检测范围为96.24%,相较Only GAT模型检测域名恶意性成功率提升近10%,对恶意域名分类成功率提升5.57%,同时相比Only LLM模型分别提升13.16%和10.19%,体现了对域名细粒度分类的较高准确率与检测范围。

表 3 3种模型检测准确率对比			%
模型	二分类准确率	四分类准确率	
Only GAT	82.65	75.00	
Only LLM	79.42	70.38	
HGAT-LLM	92.58	80.57	

为探究不同元路径在恶意域名检测中的贡献度,统计了各元路径在聚合邻居信息时的平均注意力强度。不同元路径贡献度对比如表4所示。

表 4 不同元路径贡献度对比	
元路径	注意力强度
P_1 (IP 关联)	0.443 3
P_2 (CNAME 关联)	0.783 2
P_3 (注册者关联)	0.501 6
P_4 (网段关联)	0.406 6

GAT能够根据元路径的判别力自适应地分配权重。由表4可以看出, P_2 虽关系稀疏,但获得了最高的平均注意力强度0.783 2。这一结果与安全专家的先验知识一致,表明共享CNAME记录往往是识别同源恶意攻击的最强结构化证据。 P_3 以0.501 6的权重次之,印证了攻击者复用同源身份批量注册基础设施的行为模式。相比之下, P_1 与 P_4 的注意力强度分别为0.443 3和0.406 6。尽管IP和网段层面的特征容易受到良性资源共享带来的噪声干扰,导致权重略低,但其关联关系众多仍为不可或缺的证据。上述结果证实,HGAT-LLM摆脱了传统人工定义路径权重的局限,能够精准捕提高价值的

恶意线索。

3.7 LLM 推理引擎案例检测

LLM 作为 HGAT-LLM 最终决策推理模块,旨在弥补单一图结构特征存在的视野盲区^[30]。为验证 LLM 的独立研判能力,对 LLM 的判断进行量化分析。实验将 GAT 输出的 Top-4 证据链按置信度降序依次设为选项 A~D,另增设选项 E 代表证据不足、不予判定的兜底情形。结果显示,尽管具有最高置信度的选项 A 获得了 88.40% 的采纳率,但仍在剩余 11.60% 的样本中触发了交叉验证机制,选项 B、C、D 的采纳率分别为 2.62%、2.84% 和 2.38%;另有 3.76% 的样本触发拒判选项 E。这表明,LLM 能够突破底层网络拓扑的局限,执行实质性的多维语义审查。

为验证 LLM 的统计显著性,实验分析 1 903 条有效测试样本(不含因证据不足导致 LLM 未决策的样本),决策转移矩阵如表 5 所示。

表 5 决策转移矩阵

GAT 预测	LLM 判断	样本数量	占比/%
正确	正确	165 3	86.86
正确	错误	28	1.47
错误	正确	178	9.35
错误	错误	44	2.31

由表 5 可知,GAT 判断错误但 LLM 修正正确的样本为 178 条,GAT 判断正确但 LLM 误改的样本为 28 条。在 GAT 初始预测错误的 222 个样本中,LLM 纠偏率为 80.18%(95% 置信区间:74.42%~84.86%);在 GAT 预测正确的 1 681 个样本中,LLM 干扰率为 1.67%(95% 置信区间:1.14%~2.43%)。由此说明了 LLM 的引入有效矫正了图模型在复杂节点上的逻辑偏差,提升了系统整体的检测上限。

LLM 修正 GAT 预测结果的证据链示例如图 10 所示。GAT 依据图结构关联判定 A 选项概率最高,但 LLM 通过分析域名文本中的语义矛盾,综合判断后修正为 B 选项,最终成功识别该钓鱼网站。该过程证实了 LLM 能够利用词法解析与外部知识推理对图模型的结构化偏差进行有效校准。

A.目标域:sanada-manu.co.jp 证据链路径:域1 解析 ip1 属于网段属于⁻¹ ip2 解析⁻¹→域2 证据链:sanada-manu.co.jp → 157.7.107.85 → 157.7.106.0 - 157.7.107.255 → 157.7.107.42 → setona.main.jp 类别:malware 分类概率:[0.9019, 0.0503, 0.0031, 0.0447]

B.目标域:sanada-manu.co.jp 证据链路径:解析 ip1 属于网段属于⁻¹ ip2 解析⁻¹→域2 证据链:sanada-manu.co.jp → 157.7.107.85 → 157.7.106.0 - 157.7.107.255 → 157.7.107.99 → bestfive.main.jp 类别:phishing 分类概率:[0.0164, 0.8648, 0.0109, 0.1079]

E.unknown

图 10 证据链示例

3.8 冷启动检测

针对冷启动难题,基于原始数据集 10% 的孤立节点构建隔离测试子集,旨在杜绝图结构与语义信息的跨集泄露。实验对比了基于字符模型的文献[7]方法、基于 DeepDom 图结构的文献[17]方法和所提方法针对冷启动问题的表现,结果如表 6 所示。

表 6 3 种方法的冷启动检测结果 %

方法	二分类准确率	四分类准确率
文献[7]	84.21	60.53
文献[17]	55.79	35.71
所提方法	79.22	70.35

表 6 结果表明,文献[17]方法二分类与四分类的检测准确率为 55.79% 和 35.71%。这主要归因于 DeepDom 高度依赖图结构关联,在面对新域名时检测效率大幅下滑。所提方法与之相比得益于 LLM 强大的文本理解能力,能够从域名的字符组合、词法等特征中提取深层语义线索,填补了拓扑信息缺失^[30-31]的空白。在完全没有任何图关联辅助的情况下,所提方法对恶意域名的二分类检测正确率仍高达 79%,虽然低于基于字符模型的文献[7]方法,但对恶意类型的多分类成功率高于另外两个方法。这一结果证明了 HGAT-LLM 不仅擅长利用图结构特征,在面对新攻击时,依然具备可靠的能力。

4 结语

随着网络攻击日趋复杂,传统单一视角的检测方法在准确率与可解释性上均面临挑战。融合异构图网络与大语言模型推理的恶意域名检测方法首先构建了 DNS 场景下的异构信息网络,利用基于元路径的 GAT 挖掘域名各实体间关联,生成结构化证据链,再通过 LLM 融合域名语义特征实现对恶意域名的精准定性。实验结果表明,所提方法在恶意域名检测任务中取得 92.58% 的准确率,整体性能优于传统图模型与基于特征的方法;在冷启动场景下,检测准确率仍达到 79.22%,说明该方法对结构信息不足的样本具有一定适应能力。

所提方法虽然在恶意域名分类检测与冷启动场景中表现良好,但仍存在优化空间。首先异构信息网络尚未纳入证书指纹和自治系统号等深层基础设施信息,此外,LLM 推理带来的计算开销较高,导致模型推理速度慢于轻量化模型。因此,未来工作将集中在两方面:一是扩展异构信息网络 HIN 的

拓扑结构,引入更多类型的实体与关联关系,构建更稠密的信息网络;二是探索轻量化推理策略,提升大规模场景下恶意域名的检测效率。通过上述改进,旨在构建一个高准确性、广覆盖率与强鲁棒性的智能化恶意域名检测体系。

参 考 文 献

- [1] Kountouras A, Kintis P, Lever C, et al. Enabling network security through active DNS datasets[C]//Research in Attacks, Intrusions, and Defenses. Cham: Springer International Publishing, 2016: 188-208.
- [2] Zhang S, Zhou Z, Li D, et al. Attributed heterogeneous graph neural network for malicious domain detection[C]//2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design. Dalian: IEEE, 2021: 397-403.
- [3] Patil D R. A framework for malicious domain names detection using feature selection and majority voting approach[J]. Informatica, 2024, 48(3):419-438.
- [4] Zhauniarovich Y, Khalil I, Yu T, et al. A survey on malicious domains detection through DNS data analysis[J]. ACM Computing Surveys, 2019, 51(4): 1-36.
- [5] 赵宏,常兆斌,王乐. 基于词法特征的恶意域名快速检测算法[J]. 计算机应用, 2019(1): 227-231.
Zhao Hong, Chang Zhaobin, Wang Le. Fast malicious domain name detection algorithm based on lexical features[J]. Journal of Computer Applications, 2019(1): 227-231.
- [6] Sheng S, Wardman B, Warner G, et al. An empirical analysis of phishing blacklists[C]//6th Conference on Email and Anti-Spam (CEAS 2009). Mountain View: ACM, 2009: 112-118.
- [7] Bilge L, Sen S, Balzarotti D, et al. Exposure: A passive DNS analysis service to detect and report malicious domains[J]. ACM Transactions on Information and System Security, 2014, 16(4): 1-28.
- [8] 樊昭杉,王青,刘俊荣,等. 域名滥用行为检测技术综述[J]. 计算机研究与发展, 2022, 59(11): 2581-2605.
Fan Zhaoshan, Wang Qing, Liu Junrong, et al. Survey on domain name abuse detection technology[J]. Journal of Computer Research and Development, 2022, 59(11): 2581-2605.
- [9] Antonakakis M, Perdisci R, Lee W, et al. Detecting malware domains at the upper DNS hierarchy[C]//20th USENIX Security Symposium (USENIX Security 11). San Francisco: USENIX Association, 2011: 1-16.
- [10] Schüppen S, Teubert D, Herrmann P, et al. FAN-CI: Feature-based automated NXDomain classification and intelligence [C]//27th USENIX Security Symposium (USENIX Security 18). Baltimore: USENIX Association, 2018: 1165-1181.
- [11] Antonakakis M, Perdisci R, Nadji Y, et al. From {throw-away} traffic to bots: Detecting the rise of {DGA-based} malware [C]//21st USENIX Security Symposium (USENIX Security 12). Bellevue: USENIX Association, 2012: 491-506.
- [12] Zhang J, Gao C, Gong L Y, et al. Malware detection based on multi-level and dynamic multi-feature using ensemble learning at hypervisor[J]. Mobile Networks and Applications, 2021, 26(4): 1668-1685.
- [13] Anderson H S, Woodbridge J, Filar B. DeepDGA: Adversarially-tuned domain generation and detection [C]//2016 ACM Workshop on Artificial Intelligence and Security. Vienna: ACM, 2016: 13-21.
- [14] Sidi L, Nadler A, Shabtai A. MaskDGA: A black-box evasion technique against DGA classifiers and adversarial defenses [EB/OL]. [2025-12-25]. <https://doi.org/10.48550/arXiv.1902.08909>.
- [15] Khalil I, Yu T, Guan B. Discovering malicious domains through passive DNS data graph analysis[C]//11th ACM on Asia Conference on Computer and Communications Security. Xi'an: ACM, 2016: 663-674.
- [16] Rahbarinia B, Perdisci R, Antonakakis M. Segugio: Efficient behavior-based tracking of malware-control domains in large ISP networks[C]//2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. Rio de Janeiro: IEEE, 2015: 403-414.
- [17] Sun X Q, Wang Z L, Yang J H, et al. Deepdom: Malicious domain detection with scalable and heterogeneous graph convolutional networks[J]. Computers & Security, 2020, 99: 102057.
- [18] 雷霞,罗雄麟. 深度学习可解释性研究综述[J]. 计算机应用, 2022, 42(11): 3588-3602.
Lei Xia, Luo Xionglin. Review on interpretability of deep learning[J]. Journal of Computer Applications, 2022, 42(11): 3588-3602.
- [19] Khalil I, Guan B, Nabeel M, et al. Killing two birds with one stone: Malicious domain detection with high accuracy and coverage [EB/OL]. [2025-12-25]. <https://doi.org/10.48550/arXiv.1711.00300>.
- [20] Bilge L, Kirda E, Kruegel C, et al. EXPOSURE: Finding malicious domains using passive DNS analysis[C]//18th Annual Network and Distributed Sys-

tem Security Symposium (NDSS'11). San Diego: The Internet Society, 2011: 1-17.

[21] Hao S, Thomas M, Paxson V, et al. Understanding the domain registration behavior of spammers[C]//2013 Conference on Internet Measurement Conference. Barcelona: ACM, 2013: 63-76.

[22] Sun Y Z, Han J W, Yan X F, et al. PathSim: Meta path-based top-K similarity search in heterogeneous information networks[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003.

[23] Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks[EB/OL]. [2025-12-25]. <https://doi.org/10.48550/arXiv.1710.10903>.

[24] 曾尚文, 陈丽芳. 基于深度学习的端到端恶意软件检测方法研究[J]. 计算机科学与应用, 2025, 15(5): 548-557.

Zeng Shangwen, Chen Lifang. Research on end-to-end malware detection method based on deep learning[J]. Computer Science and Application, 2025, 15(5): 548-557.

[25] Wang X, Ji H Y, Shi C, et al. Heterogeneous graph attention network[C]//The World Wide Web Conference. San Francisco : ACM, 2019: 2022-2032.

[26] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[27] Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library[C]//Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Vancouver: Curran Associates, Inc. , 2019: 8026-8037.

[28] Mavromatis C, Karypis G. GNN-RAG: Graph neural retrieval for efficient large language model reasoning on knowledge graphs[C]//Findings of the Association for Computational Linguistics: ACL 2025. Vienna: Austria Association for Computational Linguistics, 2025: 16682-16699.

[29] 付志远, 陈思宇, 陈骏帆, 等. 大语言模型安全的挑战与机遇[J]. 信息安全学报, 2024, 9(5): 26-55.

Fu Zhiyuan, Chen Siyu, Chen Junfan, et al. Challenges and opportunities of large language model security[J]. Journal of Cyber Security, 2024, 9(5): 26-55.

[30] 付豪, 龙春, 宫良一, 等. 基于语义图学习的恶意域名检测技术[J]. 计算机研究与发展, 2025(9): 2298-2314.

Fu Hao, Long Chun, Gong Liangyi, et al. Malicious domain detection technology based on semantic graph learning[J]. Journal of Computer Research and Devel-

opment, 2025(9): 2298-2314.

[31] 曹荣荣, 柳林, 于艳东, 等. 融合知识图谱的大语言模型研究综述[J]. 计算机应用研究, 2025, 42(8): 2255-2266.

Cao Rongrong, Liu Lin, Yu Yandong, et al. Review of large language models integrating knowledge graph[J]. Application Research of Computers, 2025, 42(8): 2255-2266.

[作者简介]



王闪闪(1988—),女,山东菏泽人,博士,济南大学讲师,主要研究方向为人工智能、网络空间安全等。E-mail: ise_wangss@ujn.edu.cn



马昊(2002—),男,山东济南人,济南大学硕士研究生,主要研究方向为网络黑灰产行为分析、自然语言处理等。E-mail:ma_hao@stu.ujn.edu.cn



李群(1993—),男,山东济宁人,硕士,济宁职业技术学院讲师,主要研究方向为智能数据处理等。E-mail:ujn_liqun@qq.com



潘岳(1992—),男,山东菏泽人,博士,济南大学讲师,主要研究方向为图像处理等。E-mail:ise_pany@ujn.edu.cn



陈贞翔(1979—),男,湖南郴州人,博士,济南大学教授、博士生导师,主要研究方向为网络行为测量、智能数据安全等。E-mail:zxchen@ujn.edu.cn

[责任编辑:祝 剑]