

基于异构图归纳学习的恶意域名检测研究

梁建鹏, 莫秀良, 王鹏翔, 王焕然, 王春东

引用本文

梁建鹏, 莫秀良, 王鹏翔, 王焕然, 王春东. 基于异构图归纳学习的恶意域名检测研究[J]. 计算机科学, 2025, 52(12): 358-366.

LIANG Jianpeng, MO Xiuliang, WANG Pengxiang, WANG Huanran, WANG Chundong. [Research on Malicious Domain Detection Based on Heterogeneous Graph Inductive Learning](#) [J]. Computer Science, 2025, 52(12): 358-366.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于前馈PID的应急救援四旋翼无人机安全控制研究](#)

Research on Emergency Rescue Quadcopter UAV Safety Control Based on Feedforward PID
计算机科学, 2025, 52(11A): 241200203-9. <https://doi.org/10.11896/jsjcx.241200203>

[基于知识图谱嵌入的异构图欺诈用户检测](#)

Fraud User Detection Based on Heterogeneous Information Network with Knowledge Graph Embedding
计算机科学, 2025, 52(11A): 250400085-7. <https://doi.org/10.11896/jsjcx.250400085>

[轻量级航空宽带通信系统安全认证协议](#)

Lightweight Aeronautical Broadband Communications System Security Authentication Protocol
计算机科学, 2025, 52(11A): 241200183-7. <https://doi.org/10.11896/jsjcx.241200183>

[基于多模态数据融合的公害网站识别方法研究](#)

Research on Public Nuisance Website Identification Method Based on Multi-modal Data Fusion
计算机科学, 2025, 52(11A): 241100171-10. <https://doi.org/10.11896/jsjcx.241100171>

[一种基于改进D-S证据的智慧水利网络安全态势评估方法](#)

Security Situation Assessment Method for Intelligent Water Resources Network Based on Improved D-S Evidence
计算机科学, 2025, 52(6A): 240600051-6. <https://doi.org/10.11896/jsjcx.240600051>

基于异构图归纳学习的恶意域名检测研究

梁建鹏¹ 莫秀良¹ 王鹏翔² 王焕然³ 王春东⁴

1 天津理工大学计算机科学与工程学院 天津 300384

2 白俄罗斯国立大学 明斯克 220030

3 哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001

4 天津公安警官职业学院 天津 300382

(2947279118@qq.com)

摘要 当前基于图神经网络的恶意域名检测技术需要依赖领域专家进行元路径选择,才能将异构图转换为同构图进行直推式学习。这种方法难以利用图中丰富的拓扑信息,不具有良好的扩展性和泛化能力。对此,提出一种基于异构图归纳学习的恶意域名检测技术。首先,利用元路径生成算法构建以域名、主机和域名注册信息为节点的异质信息网络。其次,为克服直推式训练方式下的模型在真实网络中适用能力差的问题,使用归纳式图神经网络 HeteroGAT 来学习由训练样本构成的异构图的通用结构,并利用基于自编码器的域名特征表示来提升检测性能。最后,在公开数据集上将所提算法与机器学习和深度学习方法进行对比。实验结果显示,所提出的方法取得了更优的性能指标,且在训练样本较少的条件下依旧能够有效处理数据不平衡问题,具有良好的鲁棒性。

关键词: 网络安全; 恶意域名; 归纳式学习; 异构图; 元路径

中图分类号 TP393.08

Research on Malicious Domain Detection Based on Heterogeneous Graph Inductive Learning

LIANG Jianpeng¹, MO Xiuliang¹, WANG Pengxiang², WANG Huanran³ and WANG Chundong⁴

1 School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China

2 Belarusian State University, Minsk 220030, The Republic of Belarus

3 College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China

4 Tianjin Public Security Police Profession College, Tianjin 300382, China

Abstract Current malicious domain detection techniques based on graph neural networks rely on domain experts for meta-path selection to convert heterogeneous graphs into homogeneous graphs for direct learning. This approach struggles to leverage the rich topological information within the graph and lacks good scalability and generalization capabilities. For this issue, this paper proposes a malicious domain detection technique based on inductive learning from heterogeneous graphs. Firstly, it constructs a heterogeneous information network with nodes representing domains, hosts, and domain registration information using a meta-path generation algorithm. Secondly, to address the model's poor applicability in real networks under direct training, it utilizes the inductive graph neural network HeteroGAT to learn the general structure of the heterogeneous graph formed by training samples and enhances detection performance through an autoencoder-based domain feature representation. Finally, it compares the proposed algorithm with machine learning and deep learning methods on public datasets. Experimental results demonstrate that the proposed method achieves superior performance metrics and effectively handles data imbalance even with a limited number of training samples, showing strong robustness.

Keywords Network security, Malicious domain detection, Inductive learning, Heterogeneous graph, Meta-path

1 引言

域名系统 (Domain Name System, DNS) 是互联网中重要的基础设施,主要功能是将域名映射为 IP 地址,网络中的各

种活动都与其关联紧密。尤其是越来越多的攻击者也借助域名系统进行各种网络攻击活动,如伪造与合法网站相似的恶意域名诱骗用户点击,或将恶意软件托管在这些域名下,并通过垃圾邮件等方式引诱用户下载和执行。现如今,诈骗、

到稿日期:2024-10-17 返修日期:2025-06-10

基金项目:国家自然科学基金重点项目(61931019)

This work was supported by the State Key Program of National Natural Science Foundation of China(61931019).

通信作者:莫秀良(moxiuliang@163.com)

赌博、钓鱼等网络黑色产业已经严重威胁到了互联网的安全和稳定,因此,高效、准确地检测出恶意域名已经成为网络安全领域的重要研究内容。

随着一些规避检测技术(如 Domain-flux^[1])被攻击者广泛使用,早期的黑名单技术已经不能及时应对现实网络中产生的大量恶意域名和复杂多变的网络环境。因此,学者开始转向基于统计分析的机器学习方法,通过从 DNS 流量和域名字符中提取关键特征来训练恶意域名分类器,但是这种方式忽略了域名之间以及域名与主机、注册信息等实体之间的拓扑结构,以至于模型检测性能往往不佳。基于图的检测方法通过从不同视角抽取结构信息构造主机-域图^[2]、域名-IP图^[3]甚至异质信息网络^[4-5]来解决这一问题。由于攻击者难以篡改域名系统中各实体之间的固有关联关系,因此基于图的检测方法在恶意域名识别上具有良好的检测效果。

然而,现有的基于图的恶意域名检测方法只考虑了节点之间的关联关系而忽略了节点本身特征信息在关系推理中的作用,并且需要人工选择元路径将异构图转换为同构图进行直推式学习。这种方法具有一定的局限性:首先,图中节点的特征往往包含丰富的上下文信息,这对模型理解 DNS 信息的动态变化至关重要。其次,人工选择元路径过于依赖领域专家知识,且同构图转换也忽略了不同类型的点边信息,导致直推式学习训练下的模型难以捕捉复杂的图结构和动态变化的模式,导致模型的可扩展性和泛化能力较差。

本文创新性地提出了一种基于异构图归纳学习的恶意域名检测方法来解决上述问题。该方法通过元路径生成算法 GTN^[6](Graph Transformer Network)识别原始图中未连接节点之间的有用连接来生成新的图结构,然后使用异构图注意力网络 HeteroGAT(Heterogeneous Graph Attention Network)来处理不同节点类型和边类型之间的异质信息,并使用稀疏自编码器对域名特征进行增强表示。直推式学习容易导致模型过拟合,从而降低其泛化能力,因此,本文引入了归纳式学习方法,其优势在于从训练数据中总结一般性规律,而不仅仅是记忆具体的训练样本,这使得基于归纳式学习的模型能够发现数据中的潜在模式,具有更好的泛化能力。当数据分布发生变化或出现新类型数据时,模型仍能保持较好的预测性能。因此,本文采用归纳式学习的方法进行模型训练。在训练阶段,模型从已标记的训练样本构成的异构图中学习通用模式,并利用这些模式在测试阶段对未标记样本构成的异构图进行恶意域名检测。通过这种方式,训练出的模型不仅能够有效识别潜在的恶意域名,还能够新的、未见过的恶意域名上保持良好的泛化性能。

本文的主要贡献如下:

1)提出了一种基于元路径生成算法的异构图构建方法,利用 HeteroGAT 处理图中异质信息,并将全局图结构信息和局部节点特征相结合,构建域名异构图关系模型。

2)提出了一种基于归纳式学习的恶意域名检测方法,通过从训练样本构成的已知图中学习图的通用结构和节点特征表示,使模型能够对未知图的恶意域名节点进行检测,具有更好的泛化能力和扩展性。

3)基于公开数据集,利用提出的检测模型进行了大量

实验,证明了模型不仅具有较高的检测精度,更具有良好的鲁棒性能。

2 相关工作

目前已经提出了许多针对恶意域名的检测方法,大致可以分为基于特征的检测方法和基于图的检测方法两类。

基于特征的恶意域名检测方法的研究成果颇多。Bilge 等^[7]提出了一个低误报率且实时有效的恶意域名检测系统,该系统通过被动分析大规模 DNS 流量,提取了 4 类不同特征,并利用这些特征构建机器学习分类器来区分良性域名和恶意域名。Palaniappan 等^[8]提出了一种主动 DNS 分析方法,通过将域名的词汇特征、DNS 特征和 Web 特征融合在一起,并训练逻辑回归分类器来检测恶意域名。Liu 等^[9]引入层次聚类改进了传统 EasyEnsemble 算法,使其采样方式更加合理,有效解决了实际 DNS 流量内数据不平衡问题。Park 等^[10]提出一种基于自编码器的无监督学习方法来检测由域名生成算法创建的恶意域名,减少了对数据标注的依赖,具有很高的应用价值。Ren 等^[11]将卷积神经网络和长短期记忆神经网络结合,用于提取域名序列特征,并引入注意力机制给提取的深层域名信息分配权重,以检测 DGA(Domain Generation Algorithm)生成的恶意域名。Wei 等^[12]提出了一种基于增强嵌入特征的超图学习方法,首先利用域名空间统计特征结合决策树构建超图结构来捕捉域名之间的高阶关联关系,并利用超图结构对域名嵌入特征进行编码和增强来进一步挖掘域名字符之间隐藏的高阶关系,特别适用于少样本场景。Yuan 等^[13]提出了一种将双向递归神经网络、注意力机制和胶囊网络相结合的联合神经网络模型,该模型通过提取域名深层语义信息来实现高效的恶意域名检测。

基于特征的检测方法只考虑了域名的本身特征,并没有关注域名之间的关联关系,易被攻击者通过改变恶意域名的字符特征等方式逃避检测,因此越来越多的学者开始探索基于图的检测方法。Khalil 等^[3]提出了一种基于全局图结构分析的恶意域名检测方法,通过域名与 IP 之间的解析关系构建域名图,克服了基于域名局部特征易被攻击者规避的问题。Sun 等^[4]提出一种鲁棒的恶意域名检测系统,该系统首次将 DNS 场景建模为包括主机、域名、IP 及其关系的异质信息网络,并设计出基于元路径转导分类的方法,可以在仅使用一部分标记样本的情况下检测出大量恶意域名。Sun 等^[5]将 DNS 场景表示为由主机、域名等实体组成的异质信息网络,并提出一种新的图卷积网络方法。该方法通过元路径引导的随机游走来处理复杂的结构信息,同时支持在大规模网络上进行增量学习。Lei 等^[14]通过构建多个二部图捕捉域名与主机、IP 等实体之间的关联,并利用单模式投影将二部图转换为同构图来提取域名之间的隐藏关系,进而利用图嵌入技术来生成有效的特征表示。该方法可以应用于恶意域名不断变化的实时检测场景下。Li 等^[15]分析了域名、主机和 IP 地址之间的交互行为,提取了域名的时序特征,并首次建模离散时间动态图来检测恶意域名。Zhang 等^[16]提出一种基于图对比学习的恶意域名检测方法,通过域名和 IP 地址之间的映射关系以及域名之间的字符相似性构造多个二部图来进行无监督

学习,在无标记样本的训练模式下依旧能够有效检测出大量的恶意域名。Wang 等^[17]探索异构图中不同邻居节点的重要性和不同元路径的重要性,从而设计出一种层次注意力机制来检测隐蔽恶意域名。但上述模型的性能过于依赖领域专家知识对于元路径的选择,因此本文选择元路径生成的方式构建新的图结构来适应不断动态变化的网络环境。

3 恶意域名检测方法

基于异构图归纳学习的恶意域名检测方法架构如图 1 所示。训练阶段包括特征分析、异构图构建和检测模型。

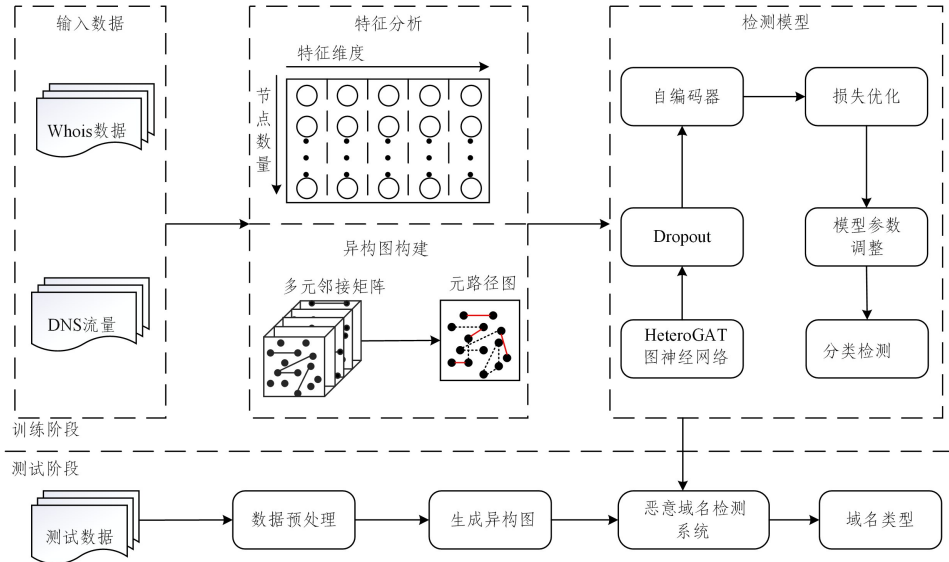


图 1 基于异构图归纳学习的恶意域名检测方法架构

Fig. 1 Architecture of malicious domain detection based on heterogeneous graph inductive learning

3.1 特征分析

3.1.1 域名字符统计特征

为挖掘恶意域名与良性域名之间的差异,提取域名字符统计特征,包括域名长度、域名中可读单词数量、域名中特殊字符和数字的数量、域名顶级域、域名子域数量以及域名中数字数量占字符总数的比例等。

3.1.2 域名时序统计特征

本文以一个月为时间窗口,分别设置每天和每天固定小时段作为时间间隔,以提取域名时序统计特征。由于被感染的恶意主机往往会查询相似的恶意域名,且在访问次数和访问时间上具有一定的规律性。因此,全天被划分为以下两种类型时间段:[0, 4), [4, 24) 和 [0, 12), [12, 24)。通过分析 DNS 流量中的主机请求和域名解析类型,提取以下 3 类时序数据。

1) 主机请求时序特征:时间段内访问域名的主机总数量、域名被主机访问的小时总数、域名被主机访问的总次数、域名被主机各小时访问次数的最高值以及所在时间、域名被主机各小时访问次数的最高值占总访问次数的比例、域名被主机各小时访问次数的最高值和最低值的差值、时间段内域名被主机访问次数的均值和方差等。

2) 域名解析时序特征:域名解析的不同资源记录类型总数、域名解析所有资源记录类型的总次数、域名解析一类资源记录类型最高次数等。

在特征分析模块,从 DNS 流量和 Whois 数据中提取出多元域名特征。在异构图构建模块,构建主机-域名和域名-注册邮箱之间的多种邻接矩阵,并使用元路径生成算法 GTN 创建新的异构图结构。在检测模型部分,首先将特征矩阵和邻接矩阵同时输入 HeteroGAT 中对节点特征进行增强表示,然后选择出域名节点并输入稀疏自编码器模块^[18]中进行特征提取,最后输入全连接网络中进行训练。训练结束后,模型可以在测试阶段识别出由未知域名以及与其相关联的主机和域名注册邮箱组成的新的异构图中的恶意域名。

3) 主机的网段时序分布特征:时间段内访问域名的主机所在的网段总数量、访问域名最多的主机所在的网段以及访问的总次数、访问域名不同主机所在的网段的数量均值和方差等。

为从整体上挖掘域名的时序特征规律,本文额外统计上述 3 类时序特征在时间窗口内的总体分布情况以及域名被主机请求的总天数和域名被主机请求的最长持续天数,挖掘高度隐蔽的恶意域名。

3.1.3 域名注册信息特征

恶意攻击者往往会使用相同的信息注册大量恶意域名,提取出以下域名注册信息特征:域名注册时间、域名结束时间、域名更新时间、域名管理者邮箱、域名注册邮箱、域名技术邮箱、域名注册所在国家、域名 Whois 服务器、域名 NS 服务器数量,以及域名注册信息更新次数等。

本文为更加有效地反映主机、域名注册邮箱与域名之间的关联关系,设置主机和域名注册邮箱两类节点的特征为其所关联的所有域名节点特征的均值。考虑到节点特征数据值范围过大或者过小可能对模型训练产生影响,因此进行归一化处理,具体计算过程如下:

$$\text{normalized_value} = \frac{\text{value} - \min}{\max - \min} \quad (1)$$

其中, value 为原始数据值, \min 是数据值中的最小值, \max 是数据值中的最大值。

3.2 异构图构建

3.2.1 多元邻接矩阵

异构图是具有多种节点类型和边类型的图数据结构。给定一个异构图 $G=(V,E)$,其中 V 表示一组节点的集合, E 表示一组边的集合。因此 G 可以由两个映射函数表示,即节点类型的映射函数 $\varphi:V \rightarrow T$,以及边类型的映射函数 $\alpha:E \rightarrow R$ 。其中 T 是不同节点类型对象的集合, R 是不同边类型对象的集合。

由于攻击者在固定的时间段内重复使用相同的受感染主机访问恶意域名,因此查询记录之间存在规律性和周期性,使用边类型 R_1 表示主机对域名访问的关系。为使模型从域名的角度分析被访问的频率和规模,使用边类型 R_2 表示域名被主机访问的关系。同时,攻击者为节约成本会使用相同的域名注册信息注册大量域名,导致恶意域名之间存在资源共享的关系,因此使用边类型 R_3 表示注册邮箱与域名的注册关系。另外,使用边类型 R_4 表示从域名指向其注册邮箱的关系,如果某个域名被确认与恶意活动相关,可以通过 R_4 追踪到其关联的注册邮箱,进而分析该邮箱注册的其他域名,发现更多潜在的恶意域名。

由不同边类型构成的邻接矩阵集合 A 包括 A_1, A_2, A_3, A_4, A_5 。其中 A_1 表示主机访问域名的邻接矩阵,若主机 h 访问域名 d ,则 $A_1(h,d)=1$,其余设置为 0; A_2 表示域名被关联主机访问的邻接矩阵,若域名 d 被主机 h 访问,则 $A_2(d,h)=1$,其余设置为 0; A_3 表示域名注册邮箱与被注册域名之间的邻接矩阵,若域名注册邮箱 m 注册域名 d ,则 $A_3(m,d)=1$,其余设置为 0; A_4 表示域名与注册邮箱之间的邻接矩阵,若域名 d 被注册邮箱 m 注册,则 $A_4(d,m)=1$,其余设置为 0;此外,为保留异构图的本身性质,增加单位矩阵 A_1 。

直接从真实 DNS 流量和 Whois 数据中提取原始信息会存在大量的冗余信息。因此,本文增加一个裁剪操作来提高模型的性能和实用性。通过以下过滤规则对原始 DNS 流量和 Whois 数据进行裁剪。

1)在时间窗口内查询域名次数少于 3 次的主机被认为是不活跃主机。恶意域名活动通常涉及频繁的主机查询域名的复杂行为模式,被不活跃主机查询的域名更可能是偶然的查询。将这些查询记录剔除可以使模型更加关注于具有显著行为活动的主机和域名,从而提高模型检测的可靠性。

2)在时间窗口内只关联到一个域名的域名注册邮箱被认为是不活跃邮箱。关联单个域名的域名注册邮箱的关联信息有限,不利于图中信息的传播;而关联多个域名的域名注册邮箱通常表示更复杂的资源关联,有助于提高模型检测效果。

3.2.2 元路径生成过程

元路径是由不同边类型组成的一个节点序列,定义了节点之间的复合关系。例如,一个元路径 P 可以被表示为 $P=(t_1 \xrightarrow{r_1} t_2 \xrightarrow{r_2} \dots \xrightarrow{r_{i-1}} t_i)$,其中 t_1, t_2 等表示元路径上不同节点的类型, r_1, r_2 等表示元路径上不同的边类型。

与人工选择元路径不同,GTN 算法的目的是识别原始图上未连接节点之间的有用连接来生成新的图结构,其具体过程可分为软选择生成加权邻接矩阵和矩阵相乘生成新的图结

构。元路径生成过程如图 2 所示。

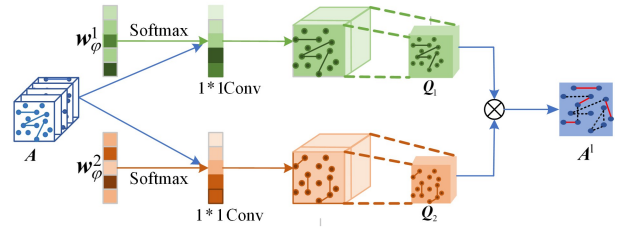


图 2 元路径组合过程

Fig. 2 Process of meta-path combination

软选择的核心是通过卷积操作为不同的边类型分配权重,从而生成一个综合各种边类型信息的加权邻接矩阵 Q 。具体计算过程如下:

$$Q = F(A; \text{Softmax}(W_\varphi)) \quad (2)$$

其中, A 为邻接矩阵集合; $W_\varphi = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$ 为卷积核的参数,作用是给每种边类型赋予一个权重系数; Softmax 函数的作用是将 W_φ 转换为概率分布,确保每种边类型的权重非负并且总和为 1; F 函数的作用是通过 1×1 的卷积操作将权重分配到不同边类型对应的邻接矩阵上,确保最终生成的 Q 矩阵是多种边类型邻接矩阵的加权求和表示。

在生成加权邻接矩阵后,GTN 算法通过矩阵相乘生成新的图结构,计算过程如下:

$$A' = Q_1 * Q_2 \quad (3)$$

其中, Q_1 和 Q_2 是通过软选择过程生成的, A' 为聚合一次元路径的邻接矩阵。矩阵相乘为组合 Q_1 和 Q_2 之间的连接来生成新的图结构,体现出节点间更复杂的关系,能够帮助模型捕获更深层次的信息。

3.3 恶意域名检测

3.3.1 异构图注意力网络

异构图注意力网络 HeteroGAT 主要由 3 个关键部分组成。1)异构图输入: HeteroGAT 能够处理包含多种节点类型和边类型的图结构; 2)异构注意力层: HeteroGAT 为不同类型的节点和边设计了独立的注意力机制; 3)节点表示更新: HeteroGAT 通过多层次结构捕捉异构图中的深层信息,从而逐层更新节点特征。因此, HeteroGAT 的特点在于异构注意力机制,它不仅关注节点和边的类型差异,还能自适应地根据邻居节点的特征和重要性调整聚合权重。

本文选用两层架构的 HeteroGAT 来处理异构图中的不同类型信息。首先针对每一种边类型构建对应邻接矩阵 A^r 表示特定节点类型之间的连接关系,其中 A^r 可以看作 A' 的一个子矩阵,并根据上述特征分析模块为每种节点类型构建初始化特征表示 h_i^0 。其次在每层 HeteroGAT 中,目标节点对通过每种边类型连接的邻居节点进行单独的注意力权重计算和信息聚合。最后将不同类型的结果聚合得到最终的节点特征表示。具体过程如下。

在第一层 HeteroGAT 中,首先计算目标节点 i 与通过边类型 r 连接的邻居节点 j 之间的注意力权重系数,其计算过程如下:

$$e_{ij}^{(r)} = \text{LeakyReLU}(\alpha_r^T [\mathbf{W}_r \mathbf{h}_i^{t_i} \parallel \mathbf{W}_r \mathbf{h}_j^{t_j}]) \quad (4)$$

其中, \mathbf{W}_r 表示边类型 r 对应的特征变换矩阵,用于对邻居节

点 j 的特征进行变换; α_r 是注意力权重向量; \parallel 代表特征向量的拼接操作; $LeakyReLU$ 代表激活函数; t 表示不同的节点类型。

为更加公平地反映每个邻居节点对目标节点的贡献,使用 $Softmax$ 函数对注意力权重系数进行归一化,计算过程如下:

$$\alpha_{ij}^{(r)} = Softmax_j = \frac{\exp(e_{ij}^{(r)})}{\sum_{k \in N_i^{(r)}} \exp(e_{ik}^{(r)})} \quad (5)$$

其中, $N_i^{(r)}$ 表示由边类型 r 连接到目标节点 i 的邻居节点集合。

其次,目标节点 i 聚合通过边类型 r 连接的不同邻居节点信息的过程如下:

$$h_i^{1:r} = \sigma(\sum_{k \in N_i^{(r)}} \alpha_{ij}^{(r)} W_r h_j^0) \quad (6)$$

其中, σ 是 ReLU 激活函数, W_r 是边类型 r 的特征变换矩阵。

最后,将每种节点类型 t 在每种边类型 r 上的聚合结果进行合并,其计算过程如下:

$$h_i^1 = f_{\text{aggregate}}(h_i^{1:r}) \quad (7)$$

其中,聚合函数执行均值操作, r 是与节点类型 t 相关联的边类型。

第二层 HeteroGAT 重复第一层的操作,只是输入的是第一层的输出特征表示 h_i^1 。经过两层 HeteroGAT 的更新,最终目标节点 i 的特征表示为 h_i^2 ,其聚合了来自不同邻居节点的信息。

3.3.2 稀疏自编码器

本文加入了稀疏自编码器模块,以进一步挖掘域名特征之间的复杂非线性关系。该模块由数据输入层、编码器、隐藏层、解码器和数据重构层组成,其中,稀疏性约束机制能够使模型在处理未知数据时表现得更加稳健。图 3 是稀疏自编码器的结构。

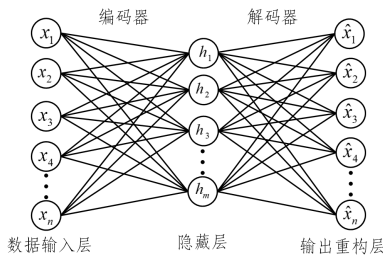


图 3 稀疏自编码器结构

Fig. 3 Structure of sparse autoencoder

由于本文最终只关注域名节点分类,因此提取出域名节点特征 h_{domain}^H 输入稀疏自编码器。在编码器部分将域名节点特征 h_{domain}^H 映射为隐藏层节点特征表示 h_{domain}^E ,计算过程如下:

$$h_{\text{domain}}^E = f_{\text{encoder}}(h_{\text{domain}}^H) \quad (8)$$

其中, f_{encoder} 是编码器的函数。

解码器部分将隐藏层节点特征表示重新映射为节点的重构特征表示 h_{domain}^V ,计算过程如下:

$$h_{\text{domain}}^V = f_{\text{decoder}}(h_{\text{domain}}^E) \quad (9)$$

其中, f_{decoder} 是解码器函数。

最后将经过 HeteroGAT 和稀疏自编码器增强的域名节点特征输入全连接层进行预测,模型参数在反向传播过程中使用损失函数进行优化。总损失函数为 $Loss$,计算过程如下:

$$Loss = classification_{\text{loss}} + sparsity_{\text{loss}} \quad (10)$$

其中, $classification_{\text{loss}}$ 和 $sparsity_{\text{loss}}$ 分别为交叉熵损失函数和稀疏性损失函数。 $classification_{\text{loss}}$ 的计算过程如下:

$$classification_{\text{loss}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(p_i) + (1-y_i) \log(1-p_i)] \quad (11)$$

其中, N 是样本数量, y_i 是样本的真实标签, p_i 是样本的预测概率。 $sparsity_{\text{loss}}$ 的计算过程如下:

$$sparsity_{\text{loss}} = \frac{1}{N} \times \sum_{i=1}^N \|x_i - \hat{x}_i\|^2 + \lambda \times \frac{1}{M} \sum_{i=1}^M |encoded_i| \quad (12)$$

其中, N 是输入样本的数量, x_i 是样本的原始输入数据, \hat{x}_i 是样本的重构输出, $\|x_i - \hat{x}_i\|^2$ 是均方误差, M 是隐藏层的维度, λ 是稀疏性惩罚项, $encoded_i$ 是样本在稀疏自编码器中最终输出的激活值。因此可以使用训练好的模型来检测未知异构图中的恶意域名节点。

4 实验评估

4.1 数据集

本实验选用奇安信实验室提供的开源数据集¹⁾。设置一个月作为时间窗口来提取 DNS 流量和 Whois 数据,共包括 13871 个良性域名和 840 个恶意域名。其中恶意域名来自 9 个不同的家族,体现出数据集中域名的多样性和广泛覆盖。考虑到真实网络环境中良性域名和恶意域名的数量存在较大差异,为使模型更好地应用于真实网络环境中,本文不对恶意域名进行数据增强处理。

4.2 实验设置

使用 Adam 优化器对模型进行优化,最大 epoch 数为 100,初始学习率为 0.005,稀疏自编码器隐藏层的维度为 16,稀疏性惩罚项为 1×10^{-4} ,dropout 为 0.5。在验证集中使用最优 loss 度量作为保存模型的条件并进行测试。所有实验在 NVIDIA 3090 上运行,深度学习框架基于 PyTorch 搭建,使用的编程语言为 Python 3.10。

本文采用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 值和马修斯相关系数(MCC)5 个指标来评估模型性能,计算过程如下:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (13)$$

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (16)$$

¹⁾ <https://datacon.qianxin.com/opendata>

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (17)$$

其中, TP 表示预测为恶意域名事实上也是恶意域名的数量, TN 表示预测为良性域名事实上也是良性域名的数量, FP 表示预测为恶意域名事实上却是良性域名的数量, FN 表示预测为良性域名事实上却是恶意域名的数量。

4.3 性能评估结果分析

本文从数据集中随机选择 $k\%$ 的域名作为训练样本对模型进行训练, 其中 k 的取值依次为 70, 50, 30 和 10。仍然需要一个验证集来调整模型的参数, 因此固定选择数据集中 10% 的域名作为验证样本, 剩余的域名作为测试样本来测试模型的实际性能表现。

表 1 列出了在各个训练样本比例下的评估指标, 其中准确率均达到 99% 以上。当训练样本占比为 70% 和 50% 时, 各个评估指标达到最优的两种情况。其中当训练样本占比为 70% 时, 召回率、精确率和 F1 值分别是 93.86%, 92.72% 和 93.29%; 当训练样本占比为 50% 时, 召回率、精确率和 F1 值分别是 93.96%, 92.89% 和 93.42%。可以看到, 训练样本占比为 50% 时各评估指标要略高于训练样本占比为 70% 时。这是因为两者均有较多的训练样本, 因此模型能充分学习到异构图的通用结构和节点特征表示。但当训练样本占比为 50% 时所构成的异构图和测试样本所构成的异构图在节点类型和边类型分布上具有更加相似的规律, 模型能够更好地将在训练样本上学习到的通用模式泛化到测试样本上, 因而性能表现更好。随着训练样本占比的逐渐减小, 各评估指标均略有下降。但即使训练样本占比为 10% 时, 模型依旧表现出很好的性能指标, 召回率、精确率和 F1 值均在 91% 以上, 表明本文方法具有很好的稳定性。

表 1 不同训练样本占比下的性能对比

Table 1 Performance comparison under different training sample proportions

样本占比	准确率	召回率	精确率	F1
70	99.25	93.86	92.72	93.29
50	99.21	93.96	92.89	93.42
30	99.09	93.34	90.78	92.04
10	99.05	92.62	91.14	91.88

考虑到实际应用中域名标注的成本, 本文将训练样本的比例设置为数据集域名总数的 50% 来进行实验分析。

为验证本文方法的不同模块对恶意域名的检测效果, 将本文方法与 3 种不同方法进行了对比实验, 结果如表 2 所列。

可以看出, 异构图中只包含主机和域名节点比只包含注册邮箱和域名节点有更高的精确率和更好的召回率。这是因为主机与域名之间的关联关系具有更加频繁的动态更新和实时变化, 这种动态信息更能揭示恶意域名集中区域, 有助于模型从更全面的视角识别出潜在的恶意域名。相比之下, 可能存在部分家族的恶意域名的注册信息与良性域名的注册信息极为相似的情况。因此, 单一依赖于域名注册邮箱不能充分挖掘恶意域名之间的资源关联, 从而造成更高的误报率。当

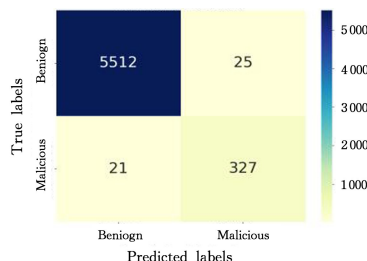
在异构图中同时考虑上述 3 类节点之间的关联时, 获得了最好的检测效果, 召回率、F1 值和 MCC 分别达到最高值 93.96%, 93.42% 和 93.01%。

表 2 不同方法性能对比

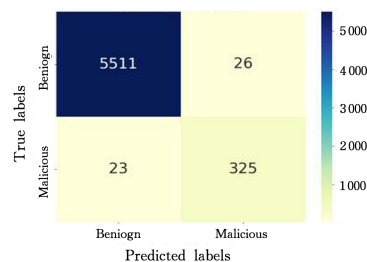
Table 2 Performance comparison of different methods

方法	准确率	召回率	精确率	F1	MCC
无注册邮箱节点	98.67	87.64	89.70	88.66	87.96
无主机节点	98.31	85.34	86.08	85.71	84.82
无稀疏自编码器	99.16	93.39	92.59	92.98	92.54
本文方法	99.21	93.96	92.89	93.42	93.01

图 4(a) 给出了当训练样本占比为 50% 时, 本文方法的混淆矩阵; 图 4(b) 给出了当训练样本占比为 50% 时, 未使用稀疏自编码器的混淆矩阵。从图 4 中可以看出, 与未使用稀疏自编码器的方法相比, 本文方法在假阳性和假阴性上均有所降低, 这表明稀疏自编码器在特征提取方面发挥了关键作用。由于本文选用的数据集中恶意域名数量较少且行为模式隐蔽, 稀疏自编码器的稀疏性约束机制有效地去除了冗余特征, 并强化了对关键特征的表示, 从而更加精准地识别出高度隐藏的恶意域名。此外, 稀疏自编码器还能够避免过拟合, 提高模型对未知域名的泛化能力。



(a) 本文方法混淆矩阵



(b) 无稀疏自编码器混淆矩阵

图 4 训练样本为 50% 时不同方法的混淆矩阵

Fig. 4 Confusion matrix of different methods with 50% training samples

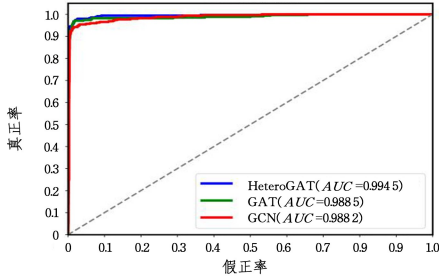
为验证 HeteroGAT 的有效性, 本文选择了 GCN^[19] (Graph Convolutional Network) 和 GAT^[20] (Graph Attention Network) 两种图神经网络方法进行对比分析。表 3 列出了在训练样本占比为 50% 时, 不同图神经网络方法的性能表现。实验结果表明, 相较于 GCN, GAT 具有更优的性能, 这主要得益于注意力机制能够有效挖掘恶意域名的关键特征。此外, 专为处理异构图结构设计的 HeteroGAT 在所有评估指标上表现更好, 主要归因于 HeteroGAT 能够更好地建模异构图中不同类型节点和边之间的复杂关系, 从而有效提升了检测性能。

表3 训练样本占比为50%时不同图神经网络方法的性能对比

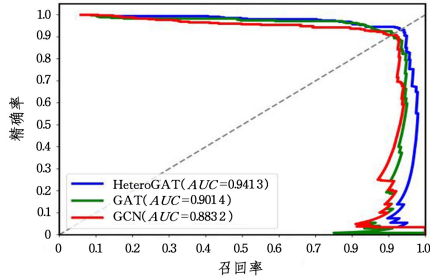
Table 3 Performance comparison of different GNN methods with 50% training sample proportion (%)

方法	准确率	召回率	精确率	F1	MCC
GCN	98.77	86.20	92.59	89.28	88.70
GAT	98.96	91.37	91.11	91.24	90.69
本文方法	99.21	93.96	92.89	93.42	93.01

为进一步验证 HeteroGAT 在处理不平衡数据集时的



(a) AUC-ROC 曲线



(b) AUC-PR 曲线

图5 训练样本占比为50%时不同图神经网络的 AUC-ROC 曲线和 AUC-PR 曲线

Fig. 5 AUC-ROC and AUC-PR curves of different GNN when the training sample proportion is 50%

4.4 与机器学习方法对比

为更好地展现所提方法的优势,将本文方法与机器学习方法进行比较,包括 KNN, SVM, RF(Random Forest)和 XGBoost等。将本文方法所使用的特征应用到机器学习方法中,并按照相同的比例随机选取训练集样本。考虑到所选用的数据集存在数据不平衡的情况,准确率很难展现模型性能,因此在表4中仅选择召回率和F1值作为评估指标。召回率反映不同方法识别出恶意域名的能力,F1值是召回率和精确

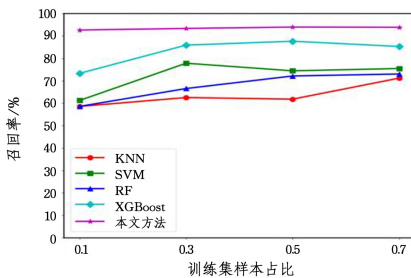
率的协调平均值。

从表4可以看出,在不同的训练样本占比下,本文方法的各个评估指标均高于机器学习方法,在所有机器学习方法中,XGBoost是性能表现最优的,其F1值为91.04%,召回率为87.64%。从图6也可以看出,随着训练样本占比的减少,机器学习方法的召回率和F1值均出现明显的下降趋势,而本文方法变化很小。当只有10%的训练样本时,机器学习方法的F1值也仅XGBoost维持在80%以上。相比之下,本文方法的F1值达到91.88%,这证明本文方法具有很好的鲁棒性。

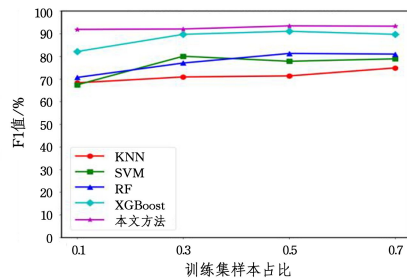
表4 与机器学习方法的性能对比

Table 4 Comparison of performance with machine learning methods (%)

训练样本占比	KNN		SVM		RF		XGBoost		本文方法	
	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1
70	71.17	74.84	75.46	78.85	73.01	80.95	85.28	89.68	93.86	93.29
50	61.78	71.31	74.43	77.78	72.13	81.23	87.64	91.04	93.96	93.42
30	62.50	70.86	77.82	80.00	66.53	77.01	85.89	89.68	93.34	92.04
10	58.55	68.27	61.21	67.37	58.55	70.64	73.30	82.08	92.62	91.88



(a) 不同方法召回率随训练集样本占比变化



(b) 不同方法 F1 随训练集样本占比变化

图6 不同方法的召回率和F1值对比

Fig. 6 Comparison of Recall and F1 scores for different methods

图7展示了当训练样本占比为10%时不同方法的性能表现。本文方法在准确率、F1值、召回率3个指标上均高于机器学习方法,在精确率上仅略低于XGBoost。然而XG-

Boost的精确率为93.25%,远高于其召回率73.30%;而本文方法的精确率为91.14%,召回率为92.62%,仅相差约1%。此外,本文方法的MCC达到91.38%,高于其他方法。在机

器学习方法中,XGBoost 和 RF 的 MCC 次优,而 SVM 的 MCC 最低,仅为 65.96%。这表明,本文方法在训练样本极少情况下依旧能够检测出大量恶意域名,并有效处理数据不平衡问题,可以更好地应用于实际网络环境中。

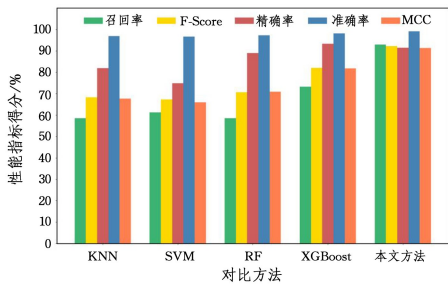


图7 训练样本占比为10%时各方法的性能比较

Fig. 7 Performance comparison of methods with 10% training sample proportion

4.5 与深度学习方法对比

将本文方法与先前工作进行比较。Yao 等^[21]提出了一种基于多元时序特征的恶意域名检测方法 LSTM-FCN。该方法将长短期记忆网络 LSTM 和全卷积神经网络 FCN 结合,从 DNS 流量中提取多元时序嵌入特征来建模域名的长期行为模式,并引入注意力机制实现对恶意域名的有效检测。Wang 等^[17]创造性地提出了一种新型的基于层次注意力机制的恶意域名检测方法 HANDOM。该方法通过使用异构注意力网络 HAN^[22]将域名统计特征和图的结构信息结合处理,依据人工选择的元路径,从异构图中提取多个只包含域名节点的同构图,并结合节点级的注意力和语义级的注意力实现对恶意域名的高效检测。本文方法与上述两种方法均选用相同的数据集进行实验分析,这样能够更准确地评估不同方法的实际性能。

表5列出了选取数据集中50%域名作为训练样本时不同方法的性能表现。

表5 不同检测方法的比较结果

Table 5 Comparison results with different detection methods

方法	准确率 (%)	召回率 (%)	精确率 (%)	F1 (%)
LSTM-FCN	98.95	93.14	71.41	81.48
HANDOM	99.00	94.81	90.78	92.75
本文方法	99.21	93.96	92.89	93.42

与 LSTM-FCN 方法相比,本文方法在4项评估指标上表现更优。这是因为 LSTM-FCN 关注了域名的时序特征,并利用 LSTM 提取规律,但未考虑域名之间的关联关系,导致其精确率和 F1 值较低,并且误报的恶意域名较多。本文方法在特征选择过程中,不仅融合了时序统计特征,还通过构建异构图神经网络挖掘域名之间的关联关系,避免了仅依赖于域名特征易被攻击者规避的风险,从而显著提高了恶意域名检测的精确率。

当训练样本占比为50%时,本文方法与 HANDOM 方法在不同评估指标上的性能差异较小。但在图构建方式上的差异较为明显。HANDOM 方法将数据集中所有域名与其相关的主机、域名注册信息构建为异构图,并基于元路径提取出

只包含域名节点的子图。然而,HANDOM 方法在将异构图转换为同构图时,失去了部分异构性。相比之下,本文方法通过元路径生成算法生成新的异构图结构,并使用 HeteroGAT 处理不同的节点类型和边类型,从而更好地挖掘异构图中的上下文信息,提升了检测性能。但当训练样本占比达到70%时,HANDOM 方法的 F1 值达到 94.62%,高于本文方法的 93.29%,这主要归因于两种方法中图的训练方式略有不同。HANDOM 方法将训练节点与测试节点合并在同一张图中,在模型训练阶段,训练节点不仅考虑了与测试节点的关联关系,还受到测试节点的特征影响。这使得模型在训练过程中提前获得了测试节点的某些通用模式,从而提升了检测性能。相比之下,本文方法通过将训练节点和测试节点分别构建为两个独立的异构图,避免了测试节点对训练过程的影响,减少了模型过度拟合测试数据的风险。因此,本文方法能够在面对未知数据时具有更好的泛化能力,但也丧失了测试节点与训练节点之间的某些重要关联关系,从而降低了模型性能。

4.6 计算复杂度与资源开销分析

在本文方法中,HeteroGAT 的计算复杂度相对较高,因此对其进行了分析。在单层 HeteroGAT 中,计算复杂度主要来源于两个方面:一是节点间的注意力计算,二是节点间的信息聚合与更新。因此,计算复杂度可表示为 $O(F \times E + F \times K \times N)$,其中 F 是节点特征输入维度, E 是边数, K 是节点特征输出嵌入维度, N 是节点数。在模型资源开销方面,当训练样本占比为10%时,显存占用为 3572 MiB。与之相比,当训练样本占比提升为50%时,显存占用达到 17927 MiB。可以得出,HeteroGAT 的计算复杂度和资源消耗主要受图中节点数和边数的影响,随着图的规模增大,注意力机制的计算和信息聚合过程将占用更多计算资源。因此,对于中小规模的异构图,本文方法具有良好的实际应用前景。

结束语 本文提出了一种基于异构图归纳学习的恶意域名检测方法,首先通过元路径生成算法构造包含3种不同类型节点的异质信息网络,并引入 HeteroGAT 和自编码器来增强模型性能,实现对恶意域名的高效检测。但本文方法在某些家族规模和活动范围较小的恶意域名检测上有一定的性能下降。此外,在处理较大规模的异构图时,本文方法可能面临较高的资源开销问题。因此,未来的工作将着重分析以下3点:1)考虑到域名系统的实时性,可以构建动态图,提取更为丰富的时序特征来增强检测性能;2)分析不同恶意域名所属的家族规模以及恶意活动的范围,并采用更先进的图神经网络算法提升检测性能;3)考虑使用稀疏邻接矩阵代替全连接矩阵表示异构图中的连接关系,以减少模型的资源开销。

参考文献

- [1] YADAV S, REDDY A K K, REDDY A L N, et al. Detecting algorithmically generated domain-flux attacks with DNS traffic analysis[J]. IEEE/ACM Transactions on Networking, 2012, 20(5):1663-1677.
- [2] MANADHATA P, YADAV S, RAO P, et al. Detecting malicious domains via graph inference[C]// Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop.

- 2014;59-60.
- [3] KHALIL I, YU T, GUAN B. Discovering malicious domains through passive DNS data graph analysis[C]// Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security. 2016;663-674.
- [4] SUN X, TONG M, YANG J, et al. {HinDom}: A robust malicious domain detection system based on heterogeneous information network with transductive classification[C]// 22nd International Symposium on Research in Attacks, Intrusions and Defenses(RAID 2019). 2019;399-412.
- [5] SUN X, WANG Z, YANG J, et al. Deepdom: Malicious domain detection with scalable and heterogeneous graph convolutional networks[J]. Computers & Security, 2020, 99:102057.
- [6] YUN S, JEONG M, YOO S, et al. Graph Transformer Networks: Learning meta-path graphs to improve GNNs[J]. Neural Networks, 2022, 153:104-119.
- [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] PALANIAPPAN G, SANGEETHA S, RAJENDRAN B, et al. Malicious domain detection using machine learning on domain name features, host-based features and web-based features[J]. Procedia Computer Science, 2020, 171:654-661.
- [9] LIU Z, ZENG Y, ZHANG P, et al. An imbalanced malicious domains detection method based on passive DNS traffic analysis[J]. Security and Communication Networks, 2018, 2018(1):6510381.
- [10] PARK K H, SONG H M, DO YOO J, et al. Unsupervised malicious domain detection with less labeling effort[J]. Computers & Security, 2022, 116:102662.
- [11] REN F, JIANG Z, WANG X, et al. A DGA domain names detection modeling method based on integrating an attention mechanism and deep neural network[J]. Cybersecurity, 2020, 3(1):4.
- [12] WEI J X, LONG C, FU H, et al. Malicious domain name detection method based on enhanced embedded feature hypergraph learning[J]. Journal of Computer Research and Development, 2024, 61(9):2334-2346.
- [13] YUAN J T, LIU Y P, YU L. A novel approach for malicious URL detection based on the joint model[J]. Security and Communication Networks, 2021, 2021(1):4917016.
- [14] LEI K, FU Q, NI J, et al. Detecting malicious domains with behavioral modeling and graph embedding[C]// IEEE 39th International Conference on Distributed Computing Systems(ICDCS 2019). IEEE, 2019:601-611.
- [15] LI Y, LUO X, WANG L, et al. DyDom: Detecting malicious domains with spatial-temporal analysis on dynamic graphs[C]// 2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys). IEEE, 2021:283-290.
- [16] ZHANG Z, ZHANG S F, YANG W, et al. Malicious Domain Name Detection Method Based on Graph Contrastive Learning[J]. Ruan Jian Xue Bao/Journal of Software, 2024, 35(10):4837-4858.
- [17] WANG Q, DONG C, JIAN S, et al. HANDOM: Heterogeneous attention network model for malicious domain detection[J]. Computers & Security, 2023, 125:103059.
- [18] NG A. Sparse autoencoder [J]. CS294A Lecture Notes, 2011, 72(2011):1-19.
- [19] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks[J]. arXiv:1609.02907, 2016.
- [20] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]// ICLR 2018. 2018.
- [21] YAO Y, FAN Z S, WANG Q, et al. Malicious Domain Detection Method Based on Multivariate Time-Series Features[J]. Netinfo Security, 2023, 23(11):1-8.
- [22] WANG X, JI H, SHI C, et al. Heterogeneous graph attention network[C]// The World Wide Web Conference. 2019:2022-2032.



LIANG Jianpeng, born in 2000, master, is a member of CCF(No. V7671G). His main research interest is malicious domain detection.



MO Xiuliang, born in 1969, postgraduate, associate professor. His main research interests include information security and network security.

(责任编辑:何杨)