基于图论算法的网络通信异常节点识别*

桂丹萍[1]，费　扬[2]
（1. 闽南科技学院　通识教育学院，福建　泉州　362300；2. 上海交通大学　电子信息与电气工程学院，上海　200240）

摘　要：针对网络通信中异常节点的识别，传统的基于规则和签名的方式，或是只参考局部图形特征的方法，在识别网络中的关键用户时都存在局限性。提出了一种基于图论算法的异常节点检测方法。首先，通过线下采集的真实局域网数据集生成图网络；利用网络的多个图形特征来定位异常节点，分析其可能存在的异常行为；其次在网络公开数据集上进行实验，以验证检测的效果；最后的测试结果证明，本方法可以在网络通信中有效地定位异常节点，高效便捷，实用性佳。

关键词：图论算法；异常检测；图形网络生成；特征分析
中图分类号：TP393．1　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097－1788.2023.07.007
引用格式：桂丹萍，费扬．基于图论算法的网络通信异常节点识别［J］．网络安全与数据治理，2023，42(7)：43－48.

Identification of abnormal nodes in network communication based on graph theory algorithm

Gui Danping[1]，Fei Yang[2]
（1. School of General Education，Minnan Science and Technology University，Quanzhou 362300，China；
2. School of Electronic Information and Electrical Engineering，Shanghai Jiao Tong University，Shanghai 200240，China）

Abstract：The traditional methods of identifying abnormal nodes in network communication，which rely on rules and signatures，or methods that only use partial graphical features，are limited when identifying key users. An anomaly node detection algorithm based on graph theory is proposed in this paper. Firstly local area network datasets collected offline are used to build a graph network；multiple graph features are analyzed to locate abnormal nodes in the network and analyze their potential abnormal behavior；secondly，experiments are conducted to test the detection effect on public network datasets. As a result of the final test results，it has proven to be efficient，convenient，and practical in locating abnormal nodes in network communication.

Key words：graph theory algorithm；abnormal detection；graph network generation；graph feature analysis

0　引言

随着信息化时代的到来，网络安全问题开始在全球大量的局域网中出现，不法分子利用网络结构的漏洞对网络内部的信息、设备甚至用户进行攻击，引发网络异常，以达到窃取信息、瘫痪网络等效果。为了提高网络安全保障能力，需要利用大量数据进行网络安全监测、风险评估和威胁画像构建。在这个网络安全检测的全过程中[1-6]，分析网络通信节点的可靠性是非常重要的一环。因此，网络安全的研究和应用变得至关重要。在真实环境中，尤其是在存有大量网络节点的内部网络中，很难预知并自动检测可疑节点。如何对通信网络流量实现安全监控并构建网络节点的威胁画像，很大程度上依赖于对于网络异常节点的正确识别。

在网络安全领域，异常检测是一个重要的研究方向。异常检测旨在识别和分类网络中的异常行为，从而判断可能存在的如端口扫描、拒绝服务等攻击行为。传统的异常检测方法主要是基于规则和签名的方法，即通过预先定义的规则和特征来检测节点行为。然而，这些方法无法有效应对未知通信模式和变化的节点行为。网络异常节点识别则是对网络中通信用户的识别与定位，尤其是对异常参与者的检测，在整个异常检测中起到重要作用。目前，针对异常节点识别问题的研究方法有许多，包括基于局部特性的方法，比如只参考度中心性，但没有考虑节点位置及周围节点的影响，其分类效果并不理

*基金项目：闽南科技学院校级科研项目（MKKYTD202304）

想[7]；再如考虑介数中心性、接近中心性，即基于全局特性的方法，不适合节点与边数众多的网络[8]；近年来，也有科研工作者利用层次分析法或是反馈机制来对网络通信节点进行风险评估[9-13]。上述方法分别从不同角度衡量节点的重要性，但是仍然存在不足：比如现有的网络拓扑结构多样，此类方法仅考虑网络的局部结构或全局结构；难以预先设定参数值；或是分辨率低，在网络中识别关键节点的能力较差，故其对其安全影响的研究尚不深入。

与此同时，图神经网络在异常检测中也得到了越来越广泛的应用[14]，因其消息传递机制而具有的高度表达能力，图神经网络已被用于高效直观地检测图中的异常。但由于图神经网络的算法复杂度较高，需要大量的计算资源和数据支持，因此如何有效地利用图神经网络进行异常检测的研究仍然是一个挑战性问题。目前基于 GNN 的图异常检测仍然具有多维度的问题与挑战：比如用于检测图异常的可解释图神经网络较少，样本的极度不平衡，无法充分利用图中可用的信息以识别图形异常等[15]。

因此，本文将基于图论算法的多种高级图特征来进行网络通信节点的识别，尤其是异常节点的识别，旨在提出一种基于图论算法的异常检测方法，通过构建网络拓扑结构和节点特征实现异常检测。该方法不仅可以便捷地进行异常检测，还可以有效减少计算复杂度，提高算法的效率和可扩展性，增强网络的安全保障能力。本研究还将基于线下采集的数据集构建的异常通信节点检测模型，应用在网络公开数据集中，通过对公开数据集的测试结果的分析验证，证明该方法在网络异常节点识别具有良好的效果。

1　基于图论的网络通信异常节点识别方法

本文关于网络通信异常节点识别工作将基于表 1 中涉及的图形指标开展，这些指标分别基于节点、边、子图三种粒度开展，对图网络中节点的多种特性进行了衡量。

本文的总体流程结构如图 1 所示，在对图特征进行了预研选取后，异常节点识别工作由两部分组成：第一阶段指代基于线下自采集数据集进行的网络通信节点识别与可靠性分析，其中借助端口扫描等信息实现了可能异常场景的分析；第二阶段则是指建立在网络公开数据集 LANL 上的工作，采取了一些算法优化策略，利用类似于基于线下数据集的模型，分三步实现了网络通信节点识别，并进行了效果验证。

表 1　本文中出现的图特征[16]总览

图形特征	种类	输入	描述
度中心性	中心性	图/子图	表示节点的归一化程度
介数中心性	中心性	图/子图	一个节点作为两个其他节点之间最短路径的桥梁的次数，除以它们之间的最短路径总数
紧密中心性	中心性	图/子图	输入图中该节点与所有其他节点间平均最短路径长度的倒数
偏心率	距离	子图	是所考虑子图中一个节点到所有其他节点的最大距离
中心	距离	子图	节点偏心率是否等于子图半径
半径	距离	子图	每个考虑子图中最小的偏心率
节点尺寸	形状	子图	节点所属的子图中的总节点数
边尺寸	形状	子图	节点所属的子图中的总边数

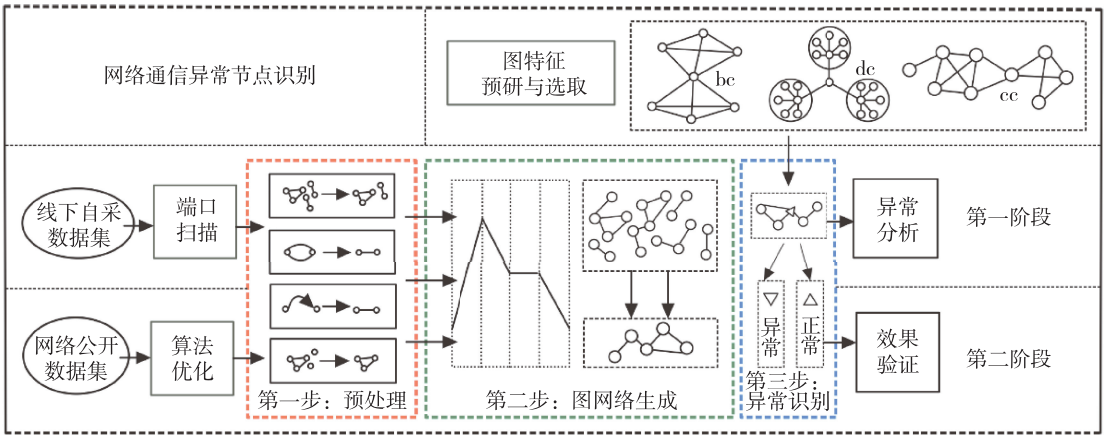

图 1　文章总体框架图示

1.1 基于线下自采数据集的异常节点识别工作

在对线下自采集数据集进行异常检测模型之前，首先需要对源数据进行数据预处理，从多个时间长度为一天的 JSON 流量数据集中提取必须字段，如协议、源 IP 地址、目的地 IP 地址、端口号等，进而进行去除空行和冗余重复数据等必要的处理。在去除重复数据的过程中分别对 < scrip, srcport, dstip, dstport > 四元组以及 < srcip, dstip > 二元组去重，同时对数据集中的数据进行处理和清洗，将数据转换成图网络中的节点和边。整体的过程大致过程如图 2 所示。

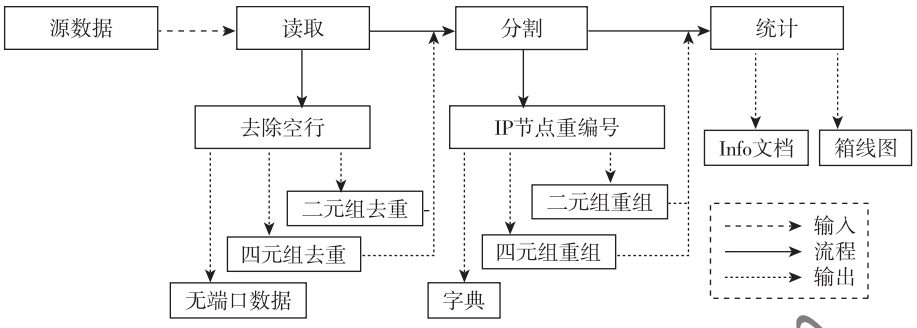

图 2 数据预处理基本流程

预处理之后的数据集内包括了生成图结构的相关信息，每个流量都由 IP、端口、协议、字节等多个特征标识。从每一个网络流量数据集中提取流量分散图。每个图中将 IP 地址和端口号的组合作为节点，边缘指示流量是否在节点之间交换。按照这些定义，网络流量数据集的每一行都可以看作是一条边。每一行都包含有关源和目标之间交换的数据信息，都可以简化为源实体（IP 和端口）、目标实体（IP 和端口）的四元组，并可以通过重新编号后保留的字典对应到原数据集中的边缘特征。

用于定位异常节点行为的参数字段[17]包括时戳、协议、消息类型、目标网址等。在对异常图特征的提取过程中，同时参考了端口扫描的结果（图 3）。通过结合对原始数据的横向和纵向综合扫描，也可以定位到异常的节点行为模式。

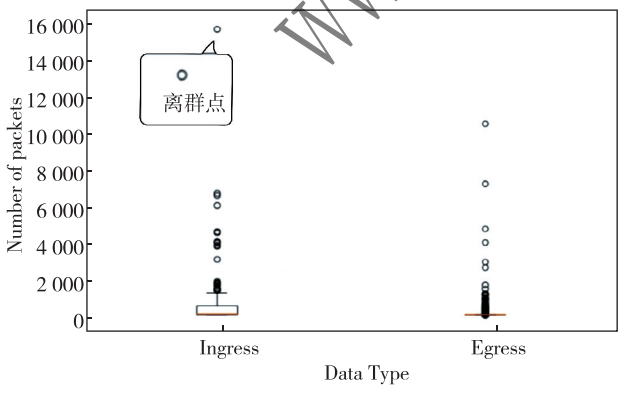

图 3 在某一通信时段端口扫描时发现节点端口数量离群值

1.2 基于网络公开数据集的异常节点识别工作

为了类似的选取合适的网络环境，本文选择了由洛斯阿拉莫斯国家实验室公司公开的多源网络安全事件数据集 LANL，它记录了公司内部计算机网络内的五个来源收集的连续 58 天的去标识化事件数据。部分信息已经脱敏处理。所使用的数据源包括来自个人计算机和集中式 Active Directory 域控制器服务器的基于 Windows 的身份验证事件，以及一组在 58 天内表现出不良行为的定义明确的红队事件。

表 2 多源网络安全事件数据集描述

统计信息名称	数量
时间/天	58
大小/GB	12
用户	12 425
计算机	17 684
进程	62 974
总计事件	1 648 275 307
红队事件	704

在使用 LANL 作为异常节点识别的图网络数据集之前，同样需要将其进行预处理。由于在给出的基本事实——红队数据集中，只有出现在攻击事件边中的节点才被定义为攻击，若此节点出现在其他被定义为正常事件的边中则不定义为攻击节点。因此，分割并选取了含有所有红队攻击事件的数据集，共计 85 个小时（LANL 数据集总计 1 392 小时）。使用处理线下自采数据集的类似手段，将 LANL 分割得到的小数据集清洗去重，以红队事件数据集作为基本事实打上标签，将脱敏的 IP 地址@ 域看作一个节点重新编号，以便图网络的生成。

预处理之后的数据按小时分割，并指定了其中含有

异常标签的时间段 85 个，代表含有红队事件的 85 小时。对这些数据集批量处理，各自生成图网络，并将边的标签对应至节点，从而完成图的初步创建。类似定义了图特征函数，来进行每个快照内的网络异常节点的识别。

而在具体的识别过程中，需要将已知红队事件提供的标签从边集对应至节点集，考虑到流量数据的时效性与通信事件的独立性，对异常节点的定义有如下规则：

（1）只在目标小时快照内有效。

若事件（a，c）在快照 A 内被定义为红队事件，则 a，c 节点的标签为 1；出现在其他快照时段内的同一事件不符合这一标准。

（2）只在目标通信事件内有效。

若事件（a，c）已经被定义为异常，即 a 节点与 c 节点异常，但在通信事件（a，b）或（b，c）中，a 节点和 c 节点不被定义为异常。

（3）只在定义的通信方向上有效。

若事件（a，c）已经被定义为异常，即 a 节点与 c 节点异常，但在通信事件（c，a）中，a 节点和 c 节点不被定义为异常。

以上规则在本次 LANL 数据集的实验中普遍适用，由于数据集内容大部分脱敏处理，此设定将异常节点严格定义在已知红队事件内。在此基础之上，定义各类高级图特征的离群值数量，进行基本统计方法的校验，将检测到的离群值计入异常节点集。

在衡量图网络的异常程度的算法中做了一些优化。由于图网络的大小在计算中往往成为决定计算时间的关键要素，往往需要相应地调整目标图的尺寸，针对于可能含有异常节点的最小图开展分析检测，从而减小运算时间，提高检测效率。首先，除了调用相关图算法库（如 NetworkX）来生成最大子图，切分图网络的策略还有：按小时划分流量信息，监测流量峰值时段，即缩短目标快照时间跨度；或是定义过滤规则来筛除大部分无意义节点：比如为检测红队事件针对 NTLM 协议进行目标数据的筛除。其次，在公开数据集 LANL 上的计算也借助了并行工作模块，实际实验时可以快速帮助分割时长达到 58 天的数据集。

2　实验结果与分析

本文实验设计分为两部分，首先是在线下自采集数据集上进行异常节点的识别定位，并对检测到的异常节点分析构建可能的异常场景；之后对网络公开数据集进行测试，利用已知基本事实来检验节点识别的效果。

2.1　线下自采数据集的异常节点分析

通过图论相关算法与高阶图特征进行了提取检测，采用数值量化和可视化手段进行定位。最终在各个日期的数据集内定位异常节点，截取相关流量数据集，从而对流量数据进行分析。用于识别异常值的较为直接的统计方法是直方图和其他统计检验，例如箱线图离群值等，可以更加完整地识别出异常节点的行为模式，并可以进一步分析可能导致异常的行为，如恶意端口扫描，垃圾邮件攻击、DoS 等。利用脚本分析的内容包括请求网址获取分析、时间戳分布分析等。以此，可以相应得到节点在某一时段内的行为模式与活跃程度，进而将其归类异常与否。同时也根据路由网关、负载等字段信息，判断其是否属于自动响应或无效通信流量。比如使用 UNIX 时间戳字段在每一小时进行时戳划分，从而得到如图 4 所示的流量数据的时间分布。从中可以确定流量峰值时段并进行端口扫描等分析手段，辅助分析某一节点的其他字段信息，如协议、路由信息、危险等级等，进而排除其流量来自防火墙转发或是其他默认行为模式的可能性，辅助衡量节点行为模式的异常程度。

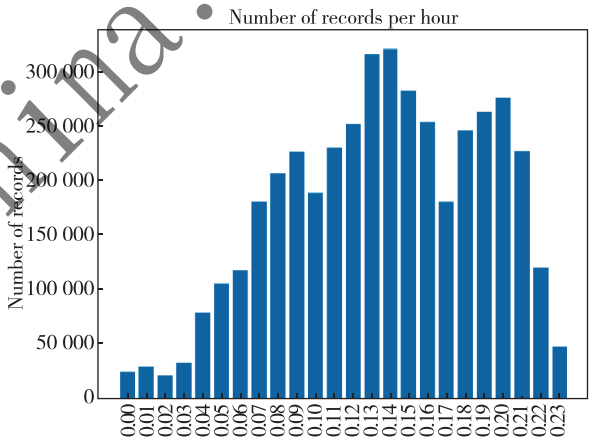

图 4　2 月 12 日流量数据在时间上的分布

例如在某一日期内的流量数据中，通过时间戳分析得到其中一个异常节点的数据仅在 4 小时内处于高度活跃，期间动态占用了大量端口，且其目标请求网址是同一企业邮箱账户，故判定该节点存在异常行为，且通过对其他字段信息的进一步的人工分析（如流量负载为零等）推测其可能存在垃圾邮箱发送或是暴力破解账号的行为。

实验结果表明，基于线下自采数据集的异常节点检测能够有效地定位到网络中的异常节点，在不需要任何先验知识的情况下自动检测异常节点。同时还能够对潜在异常节点的通信行为进行分析，从而得出关于这些节点的一些结论。

2.2　网络公开数据集的异常检验效果

在 85 个已知含有异常节点的数据集上，进行了异常

节点的识别与数据集信息的统计。异常识别的结果如表 3 所示，该表数据代表多次实验的平均结果。

表 3　在 LANL 上的平均实验结果

红队事件数（边数）	红队节点数	异常节点识别数	检测红队事件数	含异常节点快照数
706	299	4 250	45	33

已知红队事件共计 706 次（即 706 条通信边），相应定义了 299 个红队事件节点。本次异常节点识别共得到约 4 250 个异常节点，其中检测到已知红队事件的节点 45 个，分布在 33 个快照小时内。

作为对比，2022 年提出的图神经网络模型 Pikachu[18] 在 LANL 数据集上同样以 1 小时作为快照时间以检测图形网络中的异常。由于数据集的不平衡，该模型选取平均曲线下面积作为参考指标，在 LANL 的实验结果中，该项指标达到了 94%，此外并未公布其平均检测准确率与最终异常节点的成功识别数。而在类似的研究中，还有数个利用图神经网络算法的模型在 LANL 上开展了实验，但由于其研究并非针对 LANL 中的身份认证事件开展，或是更多针对异常边进行研究，因此将本文结果与类似的图神经网络模型在 LANL 数据集上的检测结果进行比较并不公平。

故针对本文的研究结果，实验单独进行了分析。基于最低限度的异常节点定义规则，本次异常识别成功有 40% 的快照内检测到了已知红队攻击事件。其中，对 9661 号节点的异常识别频次达到 20 次，而该节点在已知红队事件中属于核心通信节点，在 94.9% 的通信异常事件中出现，如图 5 所示。

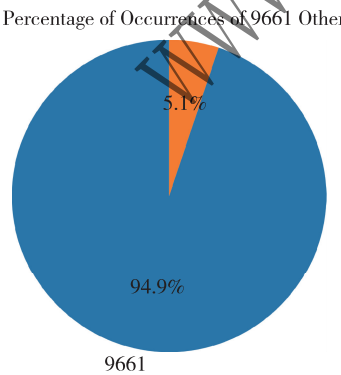

图 5　节点 9661 出现在红队事件集中的占比

而其余节点在红队数据集中的出现频次大多为个位数，且大量异常节点只出现过一次，因此在网络通信节点识别工作中，其作为异常节点被识别检测到的次数也只有一次，计入检测红队事件数。

而针对所有出现的 299 个异常节点，未能达到很高的检测覆盖率。对于这一结果的分析，可能有下列原因：首先是由于对异常节点的规则限制，出现在限制的 85 个快照之外的事件不被统计，这可能导致样本不足，从而导致检测到的异常节点较少；另外，由于对红队事件的定义在 LANL 数据集中未公开说明，其范围限制在攻击事件维度，与异常事件的定义存在差异，也可能导致检测效果有待提升。总体而言，在 LANL 公开数据集上对异常节点的检测效果较好，能够实现对多数关键节点的识别，并覆盖大多数异常事件。

3　结论

本文确定了以图论算法为数学支撑的一系列图特征检测方法，简洁实用地设计了在实际企业局域网内进行异常节点分析的系统。在线下自采数据集中，利用多种图特征的计算，实现了全面的异常检测。并辅以端口扫描检测、关键字段分析等策略推测其异常场景，涵盖了常见的硬件故障、通信故障、攻击事件等，实现了有效的可靠性分析。而在公开数据集 LANL 上，本文利用类似的方法检测并识别其中的异常节点，并基于红队事件的基本事实进行效果检验。通过本课题的异常节点识别方法发现了其中大量的异常节点，检测结果能够覆盖其中超过 90% 的异常事件，且具有良好的可用性。

总体来说，本文实现了由线下自采数据集到公开数据集的应用，为网络通信异常节点的良好识别提出了基于图论算法的便捷检测方法，为解决多种网络通信安全问题提供了思路。

参考文献

[1] CHEN W, YEUNG D Y. Defending against TCP SYN flooding attacks under different types of IP spoofing [C]//In International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies IEEE. 2006: 38 – 38.

[2] YAO G, BI J, XIAO P. VASE: filtering IP spoofing traffic with agility [J]. Computer Networks, 2013, 57 (1): 243 – 257.

[3] ALI M, NADEEM M, SIDDIQUE A, et al. Addressing sinkhole attacks in wireless sensor networks – a review [J]. International Journal of Scientific and Technology Research, 2020, 9 (8): 406 – 411.

[4] MUTALEMWA L C, SHIN S. Secure routing protocols for source node privacy protection in multi-hop communication wireless networks [J]. Energies, 2020. 13 (2): 292.

[5] MAHONEY M V, CHAN P K. PHAD: packet header anomaly detection for identifying hostile network traffic [Z]. 2001 – 04 – 11.

[6] KARTHIGHA M, LATHA L, SRIPRIYAN K. A comprehensive survey of routing attacks in wireless mobile ad hoc networks

[C] //2020 International Conference on Inventive Computation Technologies (ICICT), 2020: 396 - 402.
[7] YAN X, CUI Y, NI S. Identifying influential spreaders in complex networks based on entropy weight method and gravity law [J]. Chinese Physics B, 2020, 29 (4): 048902
[8] ULLAH A, WANG B, SHENG J F, et al. Identification of influential nodes via effective distance-based centrality mechanism in complex networks [J]. Complexity, 2021 (2021): 8403738.
[9] 林为伟, 施晓芳. 层次分析下的网络通信节点风险智能评估方法 [J]. 福建师大福清分校学报, 2021 (2): 127 - 132.
[10] 陈东洋, 郭进利. 基于图注意力的高阶网络节点分类方法 [J/OL]. 计算机应用研究, 2023, 40 (4): 1 - 7. http: //www. arocmag. com/article/02 - 2023 - 04 - 045. html.
[11] 陈妤, 秦威. 基于排序学习的复杂网络节点接近中心性近似排序 [J]. 计算机系统应用, 2022, 31 (11): 387 - 392.
[12] 王军. 网络节点数据可信度智能检测研究 [J]. 科技创新与应用, 2021, 11 (18): 46 - 48.
[13] 于洲. 复杂网络节点中心性度量及社团结构检测算法研究 [D]. 兰州: 兰州理工大学, 2021.
[14] KING I J, HUANG H H. Euler: detecting network lateral movement via scalable temporal link prediction [J]. ACM Transactions on Privacy and Security, 2022, 26 (3): 1 - 36.
[15] KIM H, LEE B S, SHIN W Y, et al. Graph anomaly detection with graph neural networks: current status and challenges [J]. IEEE Access, 2022 (10): 111820 - 111829.
[16] ZOLA F, SEGUROLA-GIL L, BRUSE J L, et al. Network traffic analysis through nodebehaviour classification: a graph-based approach with temporal dissection and data-level preprocessing [J]. Computers & Security, 2022 (115): 102632.
[17] YEN T F, OPREA A, OONARLIOGLU K, et al. Beehive: large-scale log analysis for detecting suspicious activity in enterprise networks [C] //29th Annual Computer Security Applications Conference (ACSAC'13), 2013: 199 - 208.
[18] PAUDEL R, HUANG H H, Pikachu: temporal walk based dynamic graph embedding for network anomaly detection [J]. NOMS 2022 - 2022 IEEE/IFIP Network Operations and Management Symposium, 2022: 1 - 7.

(收稿日期：2023 - 05 - 21)

作者简介：

桂丹萍 (1983 -)，女，讲师，主要研究方向：模糊数学及图像识别。

费扬 (2001 -)，通信作者，男，学士，主要研究方向：通信安全与数据分析。E-mail: 2019dfff@ sjtu. edu. cn。

(上接第 42 页)

[2] HUANG H P, ZHU P, XIAO F, et al. A blockchain-based scheme for privacy-preserving and secure sharing of medical data [J]. Computers & Security, 2020 (99): 102010.
[3] BEN S E, CHIESA A, GEN K D, et al. SNARKs for C: Verifying program executions succinctly and in zero knowledge [C]// Annual cryptology Conference. Springer, Berlin, Heidelberg, 2013: 90 - 108.
[4] GUO H, ZHANG Z F, XU J, et al. Accountable proxy reencryption for secure data sharing [J]. IEEE Transactions on Dependable and Secure Computing, 2021, 1 (18): 145 - 159.
[5] NIE X L, ZHANG A Q. Blockchain-empowered secure and privacy-preserving health data sharing in edge-based IoMT [J]. Security and Communication Neworks, 2022: 4 - 7.
[6] 郭凯阳, 韩益亮, 吴日铭. 基于 RLWE 的可撤销分层属性加密方案 [J]. 信息技术与网络安全, 2021, 40 (8): 9 - 16.
[7] Gao Jian, Zhou Fucai. An encrypted cloud email searching and filtering scheme based on hidden policy ciphertext-policy attribute-based encryption with keyword search [J]. Computer Science, 2022 (10): 8184 - 8193.
[8] YANG K, JIA X H. Expressive, efficient, and revocable data access control for multi-authority cloud storage [J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25 (7): 1735 - 1744.
[9] CUI H, ROBERT H D, WU G, et al. An efficient and expressive ciphertext-policy attribute-based encryption scheme with partially hidden access structures [C] //International conference on provable security, Springer, 2016: 19 - 38.
[10] XIONG H, ZHAO Y N, LIN P, et al. Partially policy-hidden attribute-basedbroadcastencryption with secure delegation in edge computing [J]. Future Generation Computet Systems, 2019 (97): 453 - 461.

(收稿日期：2023 - 05 - 31)

作者简介：

祁嘉琪 (2002 -)，女，本科，主要研究方向：网络安全与人工智能。

莫欣岳 (1991 -)，男，博士，讲师，主要研究方向：大数据与人工智能。

李欢 (1990 -)，通信作者，女，博士，讲师，主要研究方向：网络科学和人工智能。E-mail: lihuan@ hainanu. edu. cn。