面向未知网络威胁的网络要地自适应防御模型

郝宵荣[1,2]，刘波[1]，周鼎[2]，曹玖新[1]，张进[2]
（1.东南大学网络空间安全学院，江苏 南京 211189；2.紫金山实验室，江苏 南京 211111）

摘　要：针对未知网络威胁的隐匿性和渗透性等特点，提出了一种基于拟态防御理论的新型自适应防御模型。该模型引入拟态伪装机制，创新性地提出基于子网伪装的动态重构策略，通过动态调整子网的拓扑结构，扰乱攻击路径的生效过程，自适应阻止未知威胁的扩散，实现对网络要地的保护。该模型包括输入代理、可重构子网、调度控制层和策略裁决层，输入代理将业务流传输至可重构子网，策略裁决层构建强化学习驱动的智能决策模型，感知可重构子网的状态并生成防御策略；调度控制层根据防御策略动态调整子网连接，自适应地干扰攻击路径并阻止未知威胁的扩散。实验结果表明，与同类防御方法相比，所提模型能在有限步数内显著提高未知网络威胁防御成功率。

关键词：未知威胁；动态异构冗余；强化学习；拟态防御；自适应防御
中图分类号：TP302
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025037

Adaptive defense model for critical assets against unknown network threats

HAO Xiaorong[1,2], LIU Bo[1], ZHOU Ding[2], CAO Jiuxin[1], ZHANG Jin[2]
1. School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China
2. Purple Mountain Laboratories, Nanjing 211111, China

Abstract: To address the stealthy and penetrative characteristics of unknown network threats, a novel adaptive defense model based on mimic defense theory was proposed. The model introduced a mimic disguise mechanism and proposed a dynamic reconstruction strategy using subnet camouflage. By dynamically adjusting subnet topologies, it disrupted attack path and protected critical assets. The model included input proxy, reconfigurable subnet, scheduling control layer, and policy decision layer. The input proxy forwarded traffic to reconfigurable subnet. A reinforcement learning-based decision model in the policy decision layer perceived reconfigurable subnet states to generate defense strategies. Subnet connections were dynamically adjusted by the scheduling control layer to adaptively interfere with attack paths and prevent unknown threat diffusion. Experiments show that the proposed model achieves higher success rate in blocking unknown threats within limited steps compared to existing methods.

Keywords: unknown threat, dynamic heterogeneous redundancy, reinforcement learning, mimic defense, adaptive defense

收稿日期：2024–12–31；修回日期：2025–02–18
通信作者：刘波，bliu@seu.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2022YFB3104300）；国家自然科学基金资助项目（No.62472092, No.62172089）；江苏省网络与信息安全重点实验室基金资助项目（No.BM2003201）；教育部计算机网络与信息集成重点实验室基金资助项目（No.93K-9）
Foundation Items：The National Key Research and Development Program of China (No.2022YFB3104300), The National Natural Science Foundation of China (No.62472092, No.62172089), The Jiangsu Provincial Key Laboratory of Network and Information Security (No.BM2003201), The Key Laboratory of Computer Network and Information Integration of Ministry of Education of China (No.93K-9)

0　引言

未知网络威胁是指网络上存在的未被发现或记录的新型威胁，具有高度的隐蔽性和扩散性[1]。网络要地通常指防御方需要保护的，具有重要价值的目标[2-3]。在本文中，“网络要地”特指网络中那些具有重要商业价值或社会价值的核心服务器和数据中心。相比其他网络区域的设备，“网络要地”拥有高级特权和核心机密，更容易受到攻击者的青睐。近年来频发的网络安全事件表明[4]，高级攻击者通常采用分阶段策略，利用先进的技术和资源潜伏在与攻击目标相连的子网中。企图通过其他系统作为跳板进行不断渗透和扩散，进而隐蔽地到达“网络要地”，获取敏感数据或造成破坏性的攻击。例如：攻击者可能攻击组织的多个子网（如向网络要地提供服务和资源的多个业务子网等），试图通过多路径策略静默地到达“网络要地”。图1给出了一个企业内部的核心数据面临的网络威胁，不同类型的攻击者可能通过网络中的不同部门子网渗透到内网的数据中心，达到窃取机密文件和核心数据的目的。

如何防御未知网络威胁的扩散已经成为网络空间安全面临的严峻问题。有研究工作利用一些具有明显异常的特征检测未知攻击，如识别未知域名[5-6]、分析孤立的超文本传输协议通信流量[7-8]以及检测恶意软件的异常应用程序编程接口调用[9-10]等。然而，“网络要地”作为网络中最关键的资产，需要防御者为其配备高级的防御工具，于是一些研究尝试采用基于图模型分析[11]、传染病传播模型[12-13]和深度强化学习[14]等方法检测未知网络威胁。尽管这些研究在检测未知网络威胁中已经获得了高准确率，但仅仅检测威胁还远不能满足安全管理的需求，如何高效地对检测到的威胁进行有效响应更为重要。因此，针对未知网络威胁进行自适应防御受到研究领域的高度重视。

有研究者从博弈论的角度分析和防御未知网络威胁，因为未知网络威胁的防御可视为一场攻防博弈，防御者可以在与攻击者的交互过程中自适应调整和优化防御策略。这类方法通常把攻击方和防御方建模为博弈的双方，攻击方利用先验知识、信号传递等信息，推演未知网络的攻击路径，防御方则试图阻断这些攻击路径。文献[15]通过求解攻防博弈的最优平衡策略，最终实现防御收益的最大化。研究方法可以分为静态博弈和动态博弈2种方法，静态博弈通过让攻防双方同时采取策略来减缓攻击的扩散[16]，而动态博弈考虑攻防双方行动的顺序，更好地模拟了实际防御中的迭代优化过程[17]。然而，博弈论方法通常假设攻击者的策略是已知的，而在现实世界中防御者是无法准确预知攻击者行为模式的。

强化学习被引入攻防研究领域，用于未知网络威胁防御策略的优化。强化学习不依赖于预设的攻击者策略，而是通过与环境的不断交互来学习最优防御策略。例如，研究者应用Q-学习、深度Q网络（DQN, deep Q-network）、双重深度Q网络（DoubleDQN, double deep Q-network）以及演员−评论家等方法感知网络环境的状态并学习未知网络威胁防御策略。防御策略包括通信网络与电网系统中的恶意进程终止[18]、防御间隔识别[19]以及最优防御资源分配策略[20-21]等。强化学习虽然能在不依赖攻击者策略的情况下提供防御策略，但其在训练过

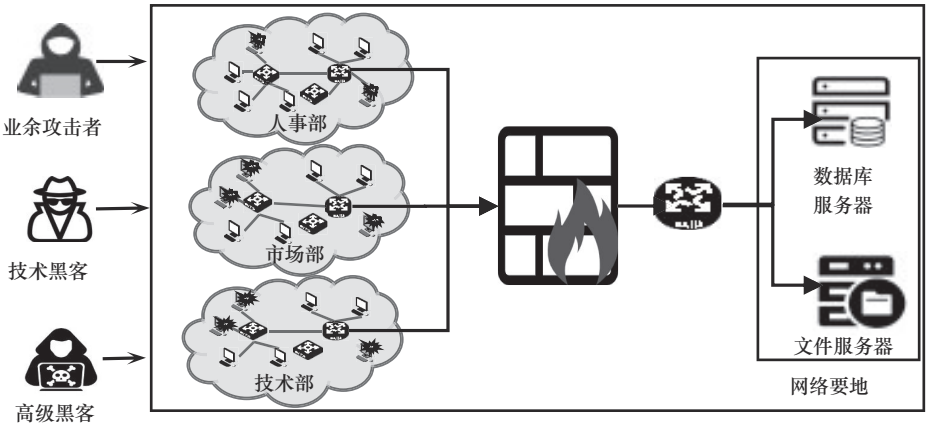


图1　攻击者从多子网入侵“网络要地”

程中往往依赖相对稳定的网络环境，缺乏面对网络环境频繁变化时的灵活性。

为应对复杂变化的网络环境，防御者需要能在面对不断变化的网络威胁时，灵活地调整其防御策略，并通过多样化的拓扑配置增强应对能力。拟态防御的思想为该研究提供了新的思路。拟态防御是由 Wu[22]借鉴生物学的免疫机制和拟态现象提出的一种安全防御理念。拟态防御导入拟态伪装机制，其期望是通过伪装目标软硬件系统，在提供高可用、高可靠和高可信服务的同时欺骗攻击者，造成攻击者认知上的困境[23]。受此启发，本文面向“网络要地”受到的未知威胁，提出了一种新型防御方法，该方法通过伪装异构子网，阻止攻击路径形成，间接地保护“网络要地”。为抵御未知的攻击路径，建立智能决策模型动态感知网络拓扑变化并重构子网，以达到攻击面的动态调整和拟态伪装，从而迅速适应不同类型的攻击。本文设计的自适应防御模型能实时感知网络环境，并迭代产生防御动作以使子网本身形态迭代更新，使攻击者难以抵达“网络要地”。

基于上述考虑，本文提出了一种面向未知网络威胁的网络要地自适应防御模型。模型的具体流程是输入代理接收正常业务流并输入可重构子网，可重构子网的状态通过调度控制层传递给策略裁决层，策略裁决层针对当前子网状态中可能潜伏的攻击路径，利用深度学习模型生成相应的子网调整动作，并向调度控制层提供子网重构策略。由调度控制器对异构子网进行响应和重构，从而实现攻击面的动态调整，阻碍攻击者对网络要地的进一步攻击。本文的主要贡献如下。

1) 创新性地提出了基于子网重构的网络要地自适应防御模型，丰富了拟态防御的理论应用。迁移拟态防御的目标对象（软硬件系统）到网络层级，通过干扰或破坏潜在攻击路径，动态重构和调整子网结构，提高了对复杂网络环境下未知网络威胁扩散的防御能力。

2) 提出了强化学习模型驱动的威胁防御与智能决策模型，实现对未知网络威胁的自适应智能裁决。构建了基于深度 Q 网络学习的智能决策模型，该智能决策模型针对可重构子网设计时间敏感策略选择多元异构随机探索和异构 Q 网络，确立重构子网的防御动作。

3) 实验结果表明，与随机防御、DQN 防御算法和双重 DQN 防御算法相比，本文模型在 3 个公开数据集上均具有较高的防御奖励，且模型收敛速度较快，能够有效防御未知威胁。

1　相关工作

学术界对未知威胁的防御研究主要涉及动态异构冗余（DHR, dynamic heterogeneous redundancy）架构理念，以及博弈论和强化学习等方法，下面从这几方面展开相关工作的总结。

1.1　基于拟态防御构造的未知威胁防御

拟态防御理论为传统网络防御提供了新的突破性思路，尤其是在面对未知的高级威胁时，能够使目标对象自身的构造场景变化[23]，有效提升目标对象的安全防护能力。拟态防御发明的拟态构造，即动态异构冗余架构，赋予了目标软硬件系统可重组、可重构、软件可定义等多样化功能[23]。DHR 架构的抽象模型主要由输入、异构执行体、裁决器、执行体调度器和输出模块组成，如图 2 所示。其中输入模块包括输入信息和输入

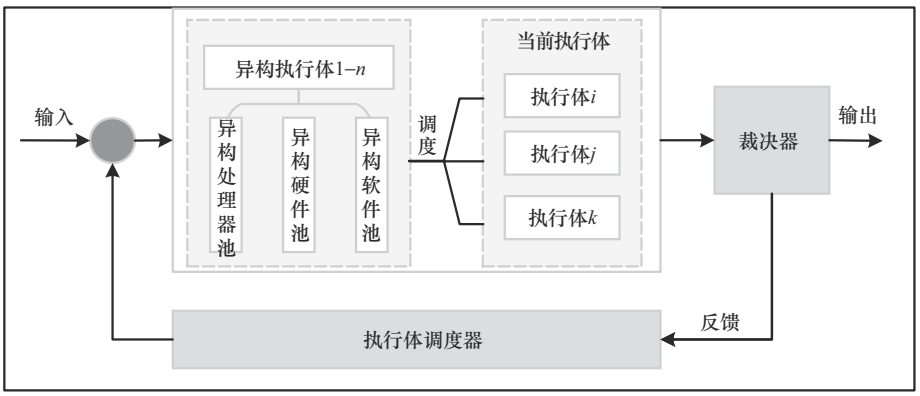


图2　DHR 架构

代理。输入代理将用户输入复制多份，并分发给异构执行体模块。异构执行体模块的输出结果通过裁决器模块进行裁决处理，根据裁决的结果选择输出或使用执行体调度器模块对异构执行体模块进行清洗和调度。基于拟态构造的威胁防御研究主要集中在防御框架和功能算法方面，如基于DHR的资源对抗模型[24]、动态调度算法[25]以及自适应的安全机制[26]，用于提升目标软硬件系统安全性、可靠性与抗攻击能力。其研究涉及多个领域，如软件定义网络（SDN, software defined network）[27]、区块链[28]、无线通信[29]等，这些方法通过提高目标软硬件系统的异构性、冗余性和动态调度能力，提高系统的防御性能，目的是防御已知攻击和未知攻击。然而，已有研究主要关注点是如何对单个目标对象实现可重组、可调度和可重构的拟态构造，当被保护的网络核心部件不是一个单一的目标对象时（如被保护的是一个子网），或者实现异构冗余面临巨大代价时，就无法按照严格的DHR架构去构造异构执行体。因此，如何根据实际应用需求灵活地实现拟态防御，尤其是如何从网络层面实现动态重构，以应对多变的网络环境和攻击情景还是一个有待研究的问题。

1.2 基于博弈论的未知威胁防御

为突破传统防御对已知攻击特征库的依赖，博弈论将攻击与防御抽象为策略集合的交互，而非具体攻击特征，可以为防御未知网络威胁扩散提供有效的思路。文献[16]构建了一个零和博弈防御模型，通过求解鞍点策略来选择适当的响应动作，其目标是确定哪些节点与网络断开连接时会减慢攻击者的行动，使可疑行为者尽可能地远离敏感组件。文献[30]通过构建非零和博弈模型和非协作博弈模型，从基于签名的方法、基于异常检测的方法和基于蜜罐的方法中选择防御策略以有效对抗攻击者。文献[31]考虑了博弈双方的动作可能存在非理性的情况，提出了一种基于累积前景理论的高级持续威胁检测博弈方案，并结合概率加权失真以及对主观攻击者和防御者的框架效应找到了最佳防御策略。考虑到博弈双方的交互是动态的，文献[17]提出了动态博弈框架。该框架允许系统实时观察和收集每个阶段的信息，并实施自适应策略，使系统能够在不同阶段调整其安全响应。文献[32]考虑到双方信息的不确定性，构建了双方不完整信息的动态博弈框架，为防御者提供针对多阶段隐蔽威胁的防御策略。基于博弈论的方法将攻防对抗的过程建模到理想的环境中，该建模方法未能充分考虑现实环境中信息的隐蔽性和动态性，限制了博弈论在真实复杂网络环境中的应用。

1.3 基于强化学习的未知威胁防御

在强化学习方法中，防御智能体与网络环境实时交互，可以针对已知攻击和未知攻击生成防御策略[33]。当前，研究者们主要利用强化学习方法生成不同类型的防御策略。例如，从终止恶意进程的防御策略角度，文献[18]构建了分层强化学习模型，该模型使用一个高级控制智能体来训练2个子智能体，分别对抗具有/不具有网络内部先验知识的2个攻击智能体。从优化防御动作实施时间间隔的角度，文献[19]主要解决了雾计算环境中雾节点易受未知网络威胁的问题。利用心理学及行为科学的“前景理论”成果研究攻防双方的主观性，并利用强化学习方法在不了解未知网络威胁模型的情况下选择检测攻击的最优时间间隔，防御雾环境中的未知网络威胁。文献[34]建立了基于深度强化学习的高级持续威胁防御模型，给出防御未知攻击的时间间隔。采用演员-评论家方法优化电网中数据采集与监视控制系统的防御策略，确定防御分配中央处理器的最佳数量、时间间隔和资源调度。在网络环境状态频繁变化的情况下，基于强化学习的未知威胁防御的训练过程可能无法收敛或产生有效的防御策略。这就要求在设计强化学习驱动的防御模型时，需要考虑到网络环境的快速变化。

针对未知威胁的防御，基于邬江兴[23]提出的拟态防御构造，研究者们探讨了多种针对系统和设备安全防御的动态异构冗余架构模型，从防御架构上提供了理论支持。而基于博弈论的未知威胁防御主要致力于解决攻防之间的对抗策略，为了更加接近于实际应用，研究者们利用强化学习从不同角度探索如何针对网络威胁行为实施防御策略。综上所述，现有研究在未知威胁防御方面进行了丰富的探索，但是当面临隐藏于复杂网络环境中的未知攻击时，如何设计出兼具智能决策能力和自适应调整能力的防御模型仍是当前面临的重要挑战，而已有研究尚不能完全解决此问题。

2　模型设计

为了构建有效的网络防御模型，首先需要明确网络防御面临的已知环境、攻击目的和假设条件。所以本节首先描述网络威胁环境，接着引出所设计的防御模型和防御策略。

2.1　威胁环境

本文研究的未知网络威胁主要指那些隐藏于动态网络之中具有一定攻击目的且有攻击路径的网络攻击。

1) 已知环境。假设攻击者已经进入内网，并控制了内网中至少一台主机。

2) 攻击目的。到达目标节点并窃取机密数据。在到达目标节点之前，攻击者会先侦察并收集有价值的内网信息，利用侦察的信息不断损害沿途的其他中间节点。这些有价值的内网信息由攻击者通过扫描工具进行收集，包含内网的活动主机、开放的端口以及网络存在的一些漏洞等。

3) 假设条件。为了准确地描述防御模型，给出下列合理性假设。① 假设攻击者是低调且谨慎的，尽可能地混入正常行为中以主动隐藏其恶意行为。② 假设攻击者具备高级技术，能够随着时间动态改变其攻击策略。当攻击者利用受损主机连接某台主机失败后，优先利用该台受损主机继续连接相邻的主机。若无法连接，则考虑重新选取新的受损主机进行扫描和连接。③ 假设攻击者只能观察到内网的部分节点。④ 假设系统记录的行为日志等信息是未经过篡改的完整数据，攻击者的攻击行为记录没有被清除且被准确记录，这些数据记录了网络随时间发生变化的状态。⑤ 假设网络内部受损主机的位置和数量未知且攻击者保留的痕迹信息与正常行为记录能够被正确区分。受损主机和正常主机建立的连接可以被阻断，而且受损主机可以被重新恢复且不会留有后门。

2.2　面向未知网络威胁的自适应防御模型

未知网络威胁具有复杂性和不可预测性，而具有重要信息的局部子网容易成为攻击的目标并被入侵。本文借鉴拟态防御的思想，设计了面向未知网络威胁的自适应防御模型，如图3所示。整个模型分为输入代理、可重构子网、调度控制层和策略裁决层4个部分。输入代理把当前的正常业务流送入可重构子网。然后，调度控制层的网络观察器收集观察到的可重构子网的初始网络状态并送到策略裁决层，经过智能裁决层产生对应子网的防御动作（重构子网的边列表）并送到调度控制层，由调度控制层的网络重构控制器接受防御策略指令和实施防御动作，得到重构后的网络状态，网络观察器观察该状态，并将防御奖励输入策略裁决层迭代优化智能决策模型。

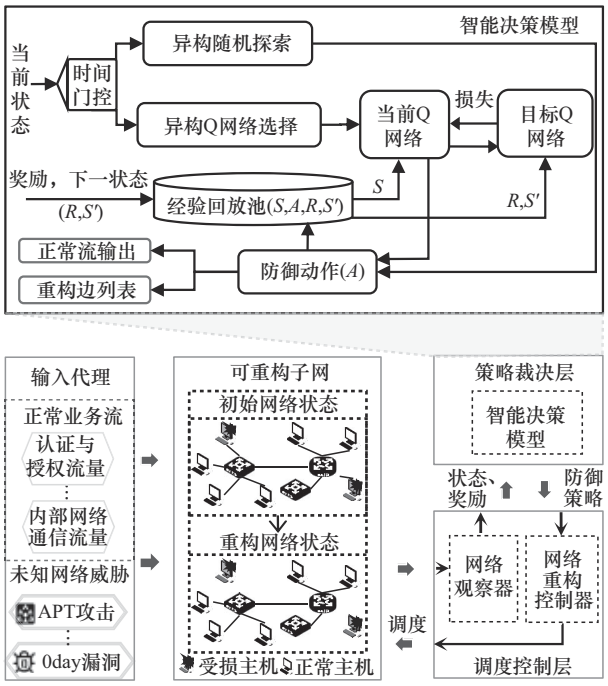


图3　面向未知网络威胁的自适应防御模型

输入代理负责将网络中的正常业务流分发给可重构子网。可重构子网除了执行正常业务流外，还会受到未知网络威胁的攻击。调度控制层包含网络观察器和网络重构控制器，主要提供对网络的实时观察和部署的能力。网络观察器（如传统入侵检测系统和流量计）。监视网络流量并提供网络状态数据及时反馈奖励给策略裁决层。网络重构控制器从策略裁决层获取调度子网重构策略指令，如软件定义网络控制器。策略裁决层基于强化学习建立智能决策模型，也称防御智能体。防御智能体采用深度Q网络实现。防御智能体根据观察到的网络状态（S）经过时间门控机制选择对网络进行异构随机探索或利用异构深度Q网络探索，异构随机探索完成后，输出防御动作（A），深度Q网络探索需要先选出合适的深度Q网络，并利用该网络学习输出防御动作。防御动作给出需要重构的连接（边）列表或者正常的连接列表。执行完防御动作后，观察新的网络状态（S'）和奖励（R）并存到经验回放池中，

同时之前的网络状态（S）和采取的防御动作也会被存到经验回放池中。当前Q网络根据当前网络状态预测Q值，目标Q网络根据奖励和新的网络状态计算得到目标Q值，通过降低预测Q值和目标Q值之间的损失来完成对防御智能体的训练，从而优化网络选取的防御动作。

2.3 策略裁决层：基于强化学习的智能决策模型

策略裁决层是整个自适应防御模型中的核心部分。在策略裁决层中，设计基于强化学习的智能决策模型实时感知网络状态并防御动态环境中的复杂未知网络威胁。本节首先简要介绍深度Q网络的概念，接着给出如何从时间、异构多融合策略方面设计防御模型去自适应地应对网络中的未知威胁。

2.3.1 深度Q网络

强化学习可以通过指导训练对象在与环境的交互过程中，学习如何在给定状态下采取有效的防御动作以最大化累积奖励$R=\sum_{i=0}^{\infty}r_i$。深度Q网络由Google DeepMind团队[35]提出，是强化学习算法的一种实现方式。深度Q网络中的Q表示质量的意思，用于衡量防御智能体在某个状态下采取某个防御动作的质量，即采取这个防御动作后能够获得的预期累积奖励。深度Q网络使用深度神经网络（如图3中的当前Q网络）将环境的当前状态输入神经网络中估计每个（状态，动作）对的函数值$Q(\boldsymbol{s},a)$。为提高学习的稳定性，深度Q网络引入了目标Q网络。目标Q网络每隔一段时间更新一次参数，可以有效缓解训练过程的不稳定性。目标Q网络的值函数表示为

$$Q^{\pi}(\boldsymbol{s},a)=r+\gamma Q^{\pi}(\boldsymbol{s}',\pi(\boldsymbol{s}')) \tag{1}$$

其中，r表示当前状态下执行防御动作得到的奖励，γ表示折扣因子，$\boldsymbol{s}'$表示新的状态，π表示防御策略。给定当前状态$\boldsymbol{s}$，最大化奖励构建防御策略，表示为

$$\pi(\boldsymbol{s})=\arg\max_{a}Q(\boldsymbol{s},a) \tag{2}$$

当前Q网络预测的Q值和目标Q网络计算的目标Q值的差为

$$d=\left(r+\gamma\max_{a}Q(\boldsymbol{s}',a)\right)-Q(\boldsymbol{s},a) \tag{3}$$

训练Q网络的损失函数通常使用Hubber损失函数，Hubber损失函数对离群值更具有稳健性。于是，损失函数表示为

$$L=\begin{cases}\dfrac{1}{2|B|}\displaystyle\sum_{(s,a,r,s')\in B}d^2, & |d|\leqslant 1\\ \dfrac{1}{|B|}\displaystyle\sum_{(s,a,r,s')\in B}|d|-0.5, & \text{其他}\end{cases} \tag{4}$$

通过有效地训练深度Q网络，最终得到针对特定网络环境的最佳防御策略。为了限制未知网络威胁的进一步扩散、减少攻击带来的破坏和保护网络要地，本文通过断开主机之间的连接这样一个离散的防御动作实现重构异构子网。

2.3.2 智能决策模型

网络环境通常是复杂且动态变化的，攻击者的行为并不总是固定的，防御者需要根据网络状态和威胁环境实时调整防御策略。为了实现自适应防御，防御智能体（即智能决策模型）应在不同时段和不同网络环境中产生最优防御策略以应对未知网络威胁的扩散。如图3所示的智能决策模型部分，在不同时段下，智能决策模型采用时间敏感的策略在异构随机探索机制和异构Q网络机制之间做出选择。一旦选中某种机制，智能决策模型将针对当前异构子网状态产生对应的防御策略。

1) 时间敏感的策略选择机制

时间敏感的策略选择机制目的是在时间上实现自适应防御，主要用于选择是否要根据历史经验发现网络中存在的威胁行为。网络环境与攻击模式可能随时间变化，因此过早依赖历史经验可能导致防御未知网络威胁策略失效。反之，适时借鉴过去的成功经验，并尝试探索新的网络行为模式则有助于提升防御效率和精准性。因此，基于最近2次防御的时间间隔，判断是否需要探索和学习新的网络行为模式，以便于在面对新的网络行为模式的攻击时能够做到快速响应，实现时间上的自适应防御。

为了能在合适的时间段选择是否要探索新的网络行为模式，时间门控机制是一个用于策略选择的好方法。让具有历史经验的异构Q网络防御策略生成发生在门开时间，门关时间则采用异构随机探索策略生成新的防御行为模式，该模式能够进一步被异构Q网络学习存入历史经验中。时间门控机制采用时间门控函数T实现。时间门控函数表示为

$$T=\sin\left(\frac{\pi(t_l-t_0)}{\text{steps}+1}\right) \tag{5}$$

其中，t_0表示初始防御时间，t_l表示上一次重构网络连接的时间，steps表示防御步数，从0步开始计

算。如果T的结果为0，则选择异构随机探索策略生成来阻止恶意连接。

2) 异构随机探索策略生成

为了提高整体防御的有效性，希望防御系统尽可能从内网中的高风险区域探索多种类型的未知威胁行为模式。因此，考虑设计异构随机探索策略生成模块生成多个异构的防御策略，并选择出当前的最优防御策略作为最终防御策略。异构随机探索策略生成模块如图4所示。给定当前的网络状态，防御智能体将产生3种随机防御策略，包括完全随机的策略生成、基于密度的随机策略生成和基于条件概率的随机策略生成。当3种随机防御策略中的任意2种一致时，将选择2种一致的防御策略实施。否则，当3种随机防御策略都不一致时，以相同的概率任意选择一种防御策略实施。

异构随机探索策略生成实现的核心在于3种异构策略的设计。完全随机的策略生成首先随机选择网络的一个源主机，之后选择某个时刻t下源主机u从某个端口p连接的目标主机v。生成的防御策略是从网络中阻断并移除选择的潜在恶意连接，即(t,u,v,p)。基于密度的随机策略生成是从稀疏密度连接中随机选出可能的恶意连接。稀疏密度连接是指某条连接所在的子图G_e，其密度小于全图G的密度。全图是基于每个固定时间窗口Δt内，所有源主机到目标主机之间连接构建的有向图。子图由全图中任意一条边的2个端点及其邻居节点之间的边构成，描述该连接的局部结构。全图的密度计算方法如式(6)所示。

$$d_G = \frac{m}{n(n-1)} + \varepsilon \tag{6}$$

其中，m为全图中的实际边数，n为全图中的节点数，$\varepsilon \leqslant 0.01$是一个松弛因子，表示一个小的稀疏限定范围值。子图的密度d_{G_e}与全图的密度计算方法相似，但没有松弛因子。基于条件概率的随机策略生成计算源节点鲜少有可能连接的节点。计算已知源节点的条件下连接到某个节点的条件概率，若条件概率小于一个阈值，则随机选择该范围内的连接进行阻断。条件概率为源节点u和目标节点v同时出现的概率$p(u,v)$与源节点u出现的概率$p(u)$的比值。具体计算式如式(7)所示。

$$p(v|u) = \frac{p(u,v)}{p(u)} = \frac{\dfrac{m_{uv}}{m_G}}{\dfrac{\sum\limits_{j \in \Gamma(u)} m_{uj}}{m_G}} = \frac{m_{uv}}{\sum\limits_{j \in \Gamma(u)} m_{uj}} \tag{7}$$

其中，$p(u,v)$表示连接(u,v)出现的数量m_{uv}与全图所有连接数量m_G的比值。这里的全图与上文的全图一致，都是一段时间窗口Δt内所有源主机到目标主机的连接记录构成的有向图。$p(u)$表示源节点连接的所有直接邻居$\Gamma(u)$的总连接数出现在全图的比例。

3) 异构Q网络选择和策略生成

异构Q网络选择和策略生成用来学习未知网络威胁在可重构子网中的行为模式。异构随机探索策略生成已通过初步探索存储了一定的经验。存储的经验用于Q网络的训练和更新，用以学习更精准的防御决策动作。为了能同时满足多种不同拓扑模式的需要，此处同样采用DHR架构中异构冗余的设计思想，设计了3种异构的深度Q网络，分别是简单线性Q网络（即深度Q网络常用的神经网络）、

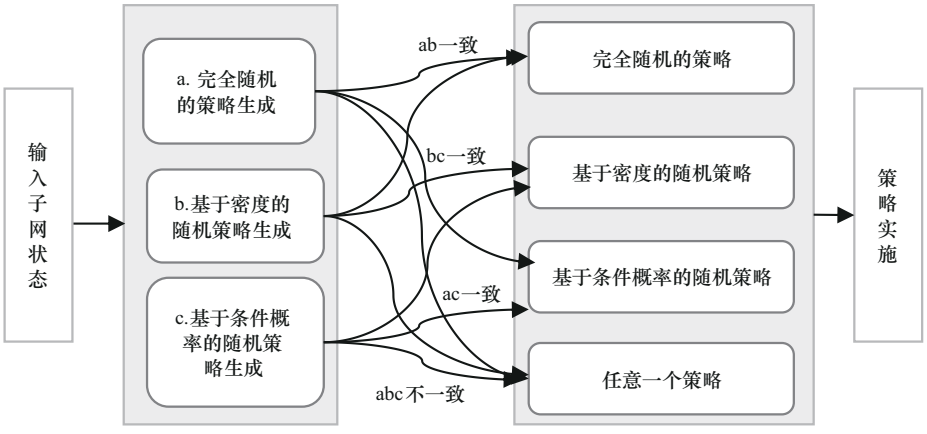


图4　异构随机探索策略生成模块

序列感知Q网络和结构感知Q网络。每种Q网络产生防御策略后会保存每次的防御奖励r。计算Q网络最近2次防御奖励的分数，选出得分最高的一个深度Q网络来生成防御策略重构受损的异构子网。实施防御策略后由观察器观察到新的奖励并保存用于下一次的Q网络选择。观察器观察到新的奖励和网络状态被用于训练Q网络。异构Q网络选择和策略生成模块的实现思路如图5所示。

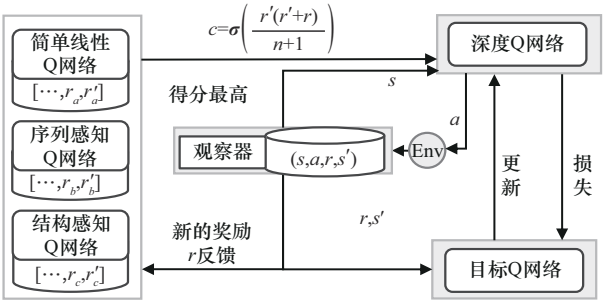


图5　异构Q网络选择和策略生成模块的实现思路

显然，最近防御总是能获得高奖励的Q网络是作为当前防御未知网络威胁的最佳策略生成对象，异构Q网络的选择需要结合每个Q网络最近2次防御奖励的分数，选出得分最高的作为当前防御未知网络威胁的策略生成模块。式(8)给出了计算得分的函数。

$$c=\boldsymbol{\sigma}\left(\frac{r'(r'+r)}{n+1}\right) \tag{8}$$

其中，r'为Q网络上一次的防御奖励，r为Q网络前2次的防御奖励，n为Q网络总共防御的次数，$\boldsymbol{\sigma}$为softmax函数，限定得分c在0~1范围内。

考虑到复杂的网络环境中存在不同级别的未知网络威胁，低级别的威胁行为模式较为简单，高级别的威胁行为模式较为复杂，因此设计的深度Q网络不仅需要有简单线性Q网络能够快速应对低级威胁，同时也要有复杂的序列感知Q网络和结构感知Q网络来分析不同行为模式的高级威胁。结合2.3.1节的基础知识，下面将详细介绍3种深度Q网络、目标Q网络的设计以及防御策略的生成。

① 简单线性Q网络。输入为可重构子网的状态，输出为状态动作值。简单线性Q网络的计算方法如式(9)所示。

$$Q(\boldsymbol{s},a)=\left[\mathrm{MLP}(\boldsymbol{s})\right]_a \tag{9}$$

其中，$\left[\mathrm{MLP}(\boldsymbol{s})\right]_a$表示从简单线性Q网络的输出中选择动作$a$对应的Q值，MLP表示使用多层感知机计算，多层感知机中的激活函数是ReLU，动作a为选择阻断的连接(t,u,v,p)，即从可重构子网中删除该条连接，$\boldsymbol{s}$为可重构子网的状态，由可重构子网的全局状态特征和动作源节点的具体特征组成，全局状态特征包含已分析边的数量，使用端口的统计。动作源节点的具体特征由源节点的id和源节点利用的协议编号构成。

② 序列感知Q网络。输入为可重构子网的状态，输出为序列感知Q网络生成的状态动作值。序列感知Q网络的计算方法如式(10)所示。

$$Q(\boldsymbol{s},a)=\left[\mathrm{ReLU}\left(\boldsymbol{W}_1\mathrm{GRU}(\boldsymbol{s})+\boldsymbol{b}_1\right)\right]_a \tag{10}$$

其中，$\mathrm{GRU}(\boldsymbol{s})$为门控循环神经网络学习的状态表示，目的是捕捉输入连续状态变化的依赖关系。$\boldsymbol{W}_1$为权重参数，$\boldsymbol{b}_1$为偏置项。由于防御者一步一步实施防御动作，产生多个连续的状态变化。每输入一段序列的状态变化都会更新GRU的隐藏状态。GRU又可以对前面多个序列状态“记忆”，因此GRU能够利用历史状态演化的信息对当前状态做出更好的动作预测。

③ 结构感知Q网络。输入为可重构子网的状态，输出为结构感知Q网络生成的状态动作值。结构感知Q网络利用胶囊网络实现。由于Q网络需要准确预测不同状态-动作对的价值，而状态特征往往具有复杂的多维特征关系。胶囊网络通过动态路由机制能更好地捕捉这些多维特征关系，提供更加精确的状态表示，从而提升Q值估计的精度。结构感知Q网络的计算方法如式(11)所示。

$$Q(\boldsymbol{s},a)=\left[\boldsymbol{W}_3 f_{\mathrm{Rout}}\left(f_{\mathrm{Cap}}\left(\mathrm{Conv1D}\left(\boldsymbol{W}_2\boldsymbol{s}'+\boldsymbol{b}_2\right)\right)\right)+\boldsymbol{b}_3\right]_a \tag{11}$$

其中，$\boldsymbol{s}'$表示将原始状态的特征维度进行扩展以适应卷积层。Conv1D表示一维卷积层，用于提取可重构子网状态特征的局部空间之间的依赖关系。f_{Cap}表示胶囊层，f_{Rout}表示动态路由层。$\boldsymbol{W}_2,\boldsymbol{W}_3$表示权重参数，$\boldsymbol{b}_2,\boldsymbol{b}_3$表示偏置项。卷积后的特征表示是$\boldsymbol{z}=\mathrm{Conv1D}\left(\boldsymbol{W}_1\boldsymbol{s}'+\boldsymbol{b}_1\right)$。胶囊层生成多个胶囊，每个胶囊表示为

$$\boldsymbol{u}_i=f_{\mathrm{Cap}}(\boldsymbol{z})=\mathrm{squash}\left(\mathrm{Conv1D}(\boldsymbol{z})\right)=\frac{\left\|\mathrm{Conv1D}(\boldsymbol{z})\right\|^2}{1+\left\|\mathrm{Conv1D}(\boldsymbol{z})\right\|^2}\frac{\mathrm{Conv1D}(\boldsymbol{z})}{\left\|\mathrm{Conv1D}(\boldsymbol{z})\right\|} \tag{12}$$

该函数表示对新的特征维度进行一维卷积后压缩，确保输出向量在一定范围内，且适用动态路由算法。动态路由算法通过迭代优化耦合系数来控制每个输入胶囊i对输出胶囊j的贡献度。输出胶囊表示为

$$\boldsymbol{v}_j = f_{\text{Rout}}\left(\boldsymbol{u}_j\right) = \text{squash}\left(\sum_i c_{ij}\boldsymbol{W}_{ij}\boldsymbol{u}_i\right) \tag{13}$$

其中，$\boldsymbol{W}_{ij}$为权重参数，c_{ij}为耦合系数，表示输入胶囊i对输出胶囊j的贡献度，计算式如式(14)所示。

$$c_{ij} = \frac{\exp\left(\boldsymbol{b}_{ij}\right)}{\sum_k \exp\left(\boldsymbol{b}_{ik}\right)} \tag{14}$$

其中，$\boldsymbol{b}_{ij}$是动态路由中的耦合概率，初始值为全0张量。更新方式为$\boldsymbol{b}_{ij} \leftarrow \boldsymbol{b}_{ij} + \boldsymbol{W}_{ij}\boldsymbol{u}_i\boldsymbol{v}_j$，随着更新迭代次数增加，正确的耦合关系逐渐加强。

④ 目标Q网络。与上述介绍的当前Q网络具有相同的网络架构，但参数不同。目标Q网络被用来稳定训练过程，即达到一个特定的更新步数时，才会被替换为当前Q网络的参数。与式(1)相同，给定观察到的下一状态，目标Q值为当前状态下的奖励r加上折扣因子乘以目标Q网络下的最大Q值。

⑤ 防御策略生成。子网的重构对防御未知网络威胁的扩散具有重要的作用，这可以通过阻断子网的连接来实现。防御策略的生成依赖当前Q网络和当前状态$\boldsymbol{s}$，根据式(2)得到当前状态对应的最大Q值的防御动作。防御动作能查找到从源节点出发的某个端口最具有潜在风险。随机选择Δt内从源节点的风险端口连接的目标节点即为要阻断的恶意连接(t,u,v,p)。根据式(1)计算状态−动作值函数评估不同阻断操作的长期收益。式(1)中表示当前状态下执行防御动作得到的奖励r计算方式如下。当正确阻断攻击事件时获得奖励$r = c\left(v_{\text{src}} + v_{\text{dst}}\right)$，错误阻断攻击事件则设置奖励$r = 0$。其中，$v_{\text{src}}$和$v_{\text{dst}}$表示源节点的基础价值，如对所有节点的比特数标准化后的值，c是一个常数。当前Q网络的损失函数与式(4)保持一致。该方法通过深度Q网络实时学习子网拓扑状态，动态生成与当前网络状态匹配的阻断策略，实现了不需要人工干预的自适应防御。智能决策模型根据当前异构子网状态产生对应的防御策略过程如算法1所示。

算法1　智能决策模型算法

定义　给定网络状态$\boldsymbol{s}$，训练轮次E，步数K，批次大小B，经验池大小D，目标Q网络更新次数N

1) for e = 1:1:E
2)　重置观察到的网络状态$\boldsymbol{s}$，初始化异构Q网络策略生成的前2次防御奖励$r = 0$，前一次防御奖励$r' = 0$，初始防御时间t_0，上一次防御时间t_l
3) for k = 1:1:K
4)　根据式(5)计算时间门控函数T
5)　if $T = 0$
6)　　采用完全随机的策略生成动作a_1；根据式(6)，采用基于密度的随机策略生成动作a_2；根据式(7)，采用基于条件概率的随机策略生成动作a_3
7)　　判断动作a_1,a_2,a_3的一致性，选择防御动作并记为a
8)　else
9)　　根据式(8)计算异构Q网络的得分c
10)　选择得分最高的Q网络，根据式(2)，得到防御动作a
11)　end if
12) 执行动作a，观察奖励r以及下一个状态$\boldsymbol{s}'$
13) 存储当前状态$\boldsymbol{s}$、动作a、奖励r和下一个状态$\boldsymbol{s}'$到经验池M中
14) if $|M| > B$
15)　从M中随机选择B个$(\boldsymbol{s},a,r,\boldsymbol{s}')$，根据式(1)、式(9)~式(14)计算目标Q网络的值函数
16)　根据式(4)计算损失
17)　更新下一个状态$\boldsymbol{s}'$给当前状态$\boldsymbol{s}$
18) end if
19) end for
20) if $e \,\%\, N == 0$
21)　赋值当前Q网络参数给目标Q网络
22) end if
23) end for

3　实验与分析

本节介绍实验数据、实验环境和对比实验，并从整体防御性能、各模块设计的优势以及模型开销的对比来分析实验结果。

3.1 实验数据

攻击者的策略和技术难以被防御者了解，直接模拟攻击会带来偏见。因此，本文利用3个公开的高级持续威胁攻击数据，用实验证明防御模型可以应对不同攻击策略的未知网络威胁。此外，为了观察模型能否有效阻断攻击者向关键系统渗透和扩散威胁，实验中选择所有包含从源主机访问或者连接到其他目标主机的行为事件。于是选择了Myneni等[36]建立的用于高级持续威胁的半合成数据集Unraveled，由悉尼新南威尔士大学提供的包含真实网络流量和合成攻击网络流量的UNSW-NB15[37]数据集以及由加拿大网络安全研究所等生成的CICIDS2017数据集[38]。详细说明如下。

1) Unraveled。该数据集是在真实的网络环境中模拟产生，用于捕获高级持续威胁。该数据集利用来自著名的高级持续威胁数据库之一（即MITRE的高级持续威胁组数据库）的攻击信息。根据预定义的业务功能模拟多个普通员工在6周（2021年5月26日—7月3日）内的流量和活动。由于防御的目的是阻断攻击者从受损主机连接到其他主机的活动。因此，本文选取第5周第5天横向移动阶段的网关流量数据集。当阻断事件的类型为非良性事件时，表明防御模型阻断正确，并给与一定的奖励。

2) UNSW-NB15。该数据集是专门针对网络中存在的低足迹攻击环境设计的，这种低足迹攻击常出现在高级持续威胁中。攻击者通过减少网络活动、使用合法工具、利用零日（0 day）漏洞等手段，降低攻击痕迹，实现长时间的隐蔽渗透和持续攻击。该数据集是包含现实正常流量和Fuzzers、Analysis、Exploits、Generic、Shellcode、后门攻击、拒绝服务攻击、侦察攻击和蠕虫攻击9种攻击流量。攻击标签从包含所有模拟攻击类型的真值表中标记，时间跨度为2015年1月22日19:50至2015年1月23日08:25和2015年2月18日11:45至2015年2月18日20:21。

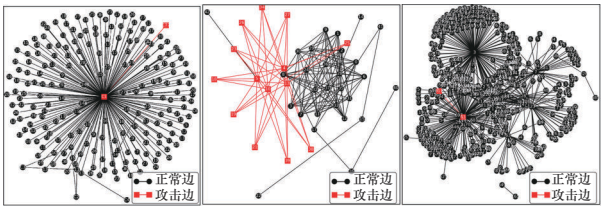
3) CICIDS2017。该数据集旨在模拟更加全面和多样的流量数据集用于评估入侵检测。数据集涵盖了常见的攻击，如拒绝服务攻击、分布式拒绝服务攻击、暴力破解、跨站脚本攻击、结构化查询语言（SQL, structured query language）注入、渗透攻击、端口扫描和僵尸网络。本文的重心在于通过动态隔离和重构子网，有效限制攻击者向网络要地前进。渗透攻击往往具有更高的隐蔽性，评估渗透攻击可以更好地测试子网重构在保护网络要地方面的表现。因此，实验选取第4天下午的渗透攻击场景流量数据集。

为了直观地给出数据的分布情况，本文给出了用于构建强化学习环境的数据分布（如表1所示）和状态分布（如图6所示）。表1中选择Unraveled数据集第5周第5天前1 h的数据、UNSW-NB15数据集前5 min数据以及CICIDS2017数据集第4天14:30之后的5 min数据进行分析。经过对数据的预处理，包含去除数据中重复值以及和目标主机一样的流量，最终得到实验数据。

图6给出了Unraveled、UNSW-NB15和CICIDS2017这3种数据集的状态分布，分别代表了不同类型的网络攻击场景，其中，方块节点及其节点对之间的连接表示受损节点和攻击边，圆圈节点及其节点对之间的连接表示正常节点和正常边。通过分析这些数据集，能够较为全面地评估本文模型在不同环境下的适应能力和防御效果。图中Unraveled数据集显示的是一个较为集中且呈现星状分布的网络结构。中心节点拥有大量正常边和少量攻击边。攻击边的数量较少，且攻击的目标是网络的核心节点或与核心节点相连的部分节点。UNSW-NB15数据集的网络结构更加分散，攻击边分布较为广泛，连接了多个不同的节点。相比于Unraveled数据集，UNSW-NB15数据集的攻击行为更加频繁且覆盖范围更大。CICIDS2017数据集的图结构相比前2个数据集更为复杂，且呈现出多团簇的特征，每个簇内部连接密集。

表1 用于构建强化学习环境的数据分布

数据集	节点数/个	流量数/条	受损节点数/个	攻击流/条	时间
Unraveled	218	3 843	2	122	6月25日16:15至6月25日17:15
UNSW-NB15	42	6 466	14	347	2月18日16:59至2月18日17:04
CICIDS2017	439	7 216	2	5	6月7日14:30至6月7日14:35



(a) Unraveled　(b) UNSW-NB15　(c) CICIDS2017

图6　3个数据集的状态分布

3.2　实验环境

实验运行在配备 Intel(R) Xeon(R) Gold 6242R 3.10 GHz CPU 和 Nvidia RTX 3090 GPU 的服务器上。本文根据3个数据集构建强化学习的网络环境，并设置防御执行体根据当前的网络环境进行防御，优化了所有参数以获得最佳性能。本文设置防御者的步数为2 500，迭代3次以获得平均值。简单线性Q网络包括4层，中间隐层维度每层的输出维度分别为1 024、512和128。序列感知Q网络的GRU输出维度为256，中间线性层输出维度为128。结构感知Q网络的第一层线性输出维度为128，第一个卷积层输入通道为1，输出通道为256，卷积核大小为9，卷积步数为1。胶囊层输入通道为256，输出通道为32，卷积核大小为9，卷积步数为2，包含8个胶囊，每个胶囊维度为16。动态路由层输出维度为16。所有Q网络的学习率为0.05，经验回放池大小为10 000。模型使用均方根传播（RMSprop, root mean square propagation）优化器进行优化，目的是防止自适应梯度（Adagrad, adaptive gradient）优化算法中学习率过小导致训练过早停滞问题。

3.3　对比实验

为全面评估自适应防御模型在防御未知网络威胁方面的性能，要求对比模型同样需要通过与环境进行实时交互给出防御策略。现有基于强化学习的防御模型大多是基于DQN、DoubleDQN和演员评论家算法的强化学习方法实现，而演员评论家算法更适合对连续防御动作进行优化。因此对比实验中选取以下3种典型防御策略作为对比模型：随机探索防御、利用简单线性Q网络的DQN防御策略，以及利用DoubleDQN的防御策略。

1) 随机防御策略。随机防御策略不依赖于历史经验或网络状态信息，随机选择网络防御动作。该方法简单易行，但由于缺乏对网络状态的理解，在一些特定攻击上不一定能给出有效的决策。

2) 利用简单线性Q网络的DQN防御策略。为了便于比较，称利用简单线性Q网络的防御策略为DQN。该模型采用简单线性Q网络来学习网络状态和防御动作之间的关系。Q网络通过对状态−动作对进行评估，不断调整决策以最大化长期奖励。简单线性Q网络的模型结构较为简单，所以应对复杂攻击模式的适应能力有限。

3) 利用DoubleDQN的防御策略。DoubleDQN分离动作选择和Q值估计过程，这种分离使DoubleDQN能更准确评估动作的实际价值，减少了Q值估计中的偏差，从而有助于提高防御策略的质量。

3.4　实验结果

3.4.1　防御性能分析

为了全面、系统地评估本文模型的整体性能。下面先分析防御模型累积奖励的趋势，用以评估模型策略的效果。之后给出模型在训练过程中损失函数的变化来分析模型学习过程中的收敛性和稳定性。

图7对比了本文模型与其他防御模型在3个数据集上的累积奖励。累积奖励代表防御策略在面对攻击时的总体性能表现，累积奖励越高，表示模型防御性能越好。在更少的步数达到最高的奖励，模型防御性能越好。图7(a)展示了不同模型在Unraveled数据集上随迭代步数增长的累积奖励变化。本文模型的累积奖励增长明显优于其他对比模型，特别是在训练后期，表现出更强的防御策略学习能力。采用DQN防御的累积奖励到1 500步之后才能比随机探索防御和采用DoubleDQN防御效果好。这表明基础的线性Q网络模型虽然通过学习有一定的防御能力，但该模型未能充分捕捉到多维复杂状态特征，导致其防御动作的选择不够精确，累积奖励低于本文模型。随机探索防御和采用DoubleDQN防御表现出明显较低的累积奖励，且需要迭代更多的步数（即分析更多的网络节点和连接）才能完成一定的防御任务。这表明随机探索防御动作在应对复杂的网络攻击时表现较差，且DoubleDQN未能学习到防御模式。而本文模型在训练早期和后期均比其他防御模型能迅速地防御攻击，这表明本文模型能够更有效地找到防御模式，从而精准阻断受损主机到其他主机的恶意连接，达到自适应弹性防御。图7(b)在UNSW-NB15数据集上具有与Unraveled相似的防御性能。图7(c)展示的CI-

CIDS2017数据集具有多团簇的复杂网络结构，网络环境更为复杂，防御任务也更具挑战性。本文模型在该环境中仍能取得较高奖励，表明其具有适应复杂攻击模式的能力。

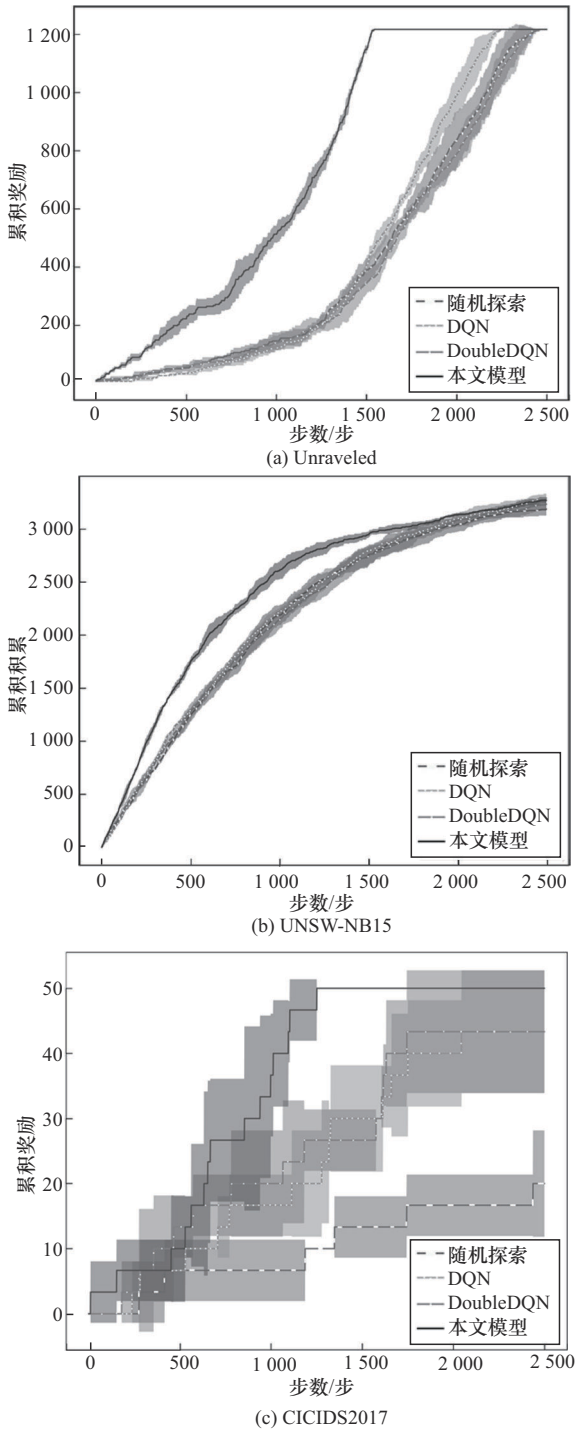


图7　不同模型在不同数据集上的防御性能对比

图8为模型训练的收敛情况对比。结果显示，本文模型（实线）在所有数据集上的损失波动最小且收敛速度最快，表现出更好的稳定性和学习能力。DoubleDQN（点划线）次之，损失波动较大但逐渐收敛。DQN（虚线）损失波动明显更大且收敛速度较慢。尤其在UNSW-NB15和CICIDS2017数据集中，本文模型的损失值在约500步以内达到较低值并逐渐趋于平稳，而其他模型的损失波动显著，难以稳定收敛。因此，在不同的网络环境下，本文模型均能迅速学会有效的防御策略，并且具有更高的稳定性。

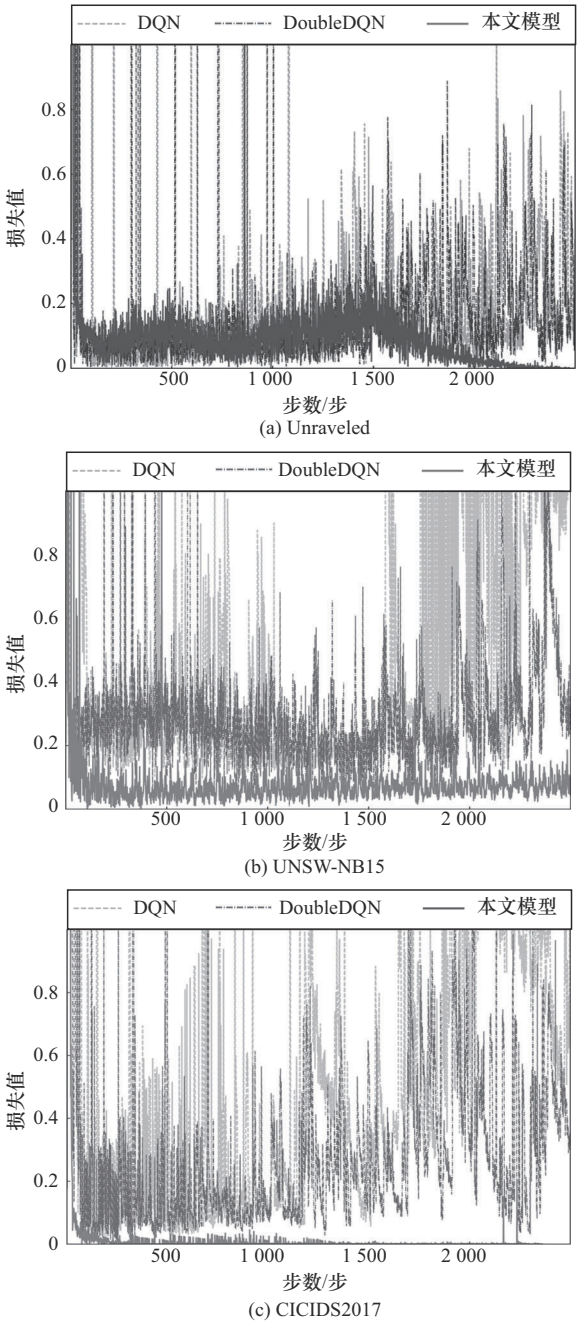


图8　模型训练的收敛情况对比

3.4.2　有效性分析

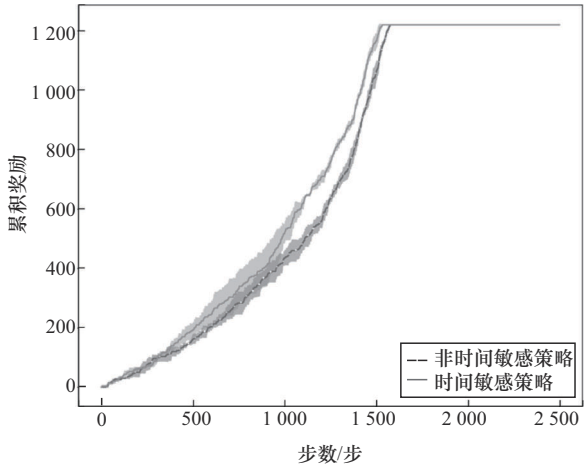
1) 时间敏感的策略选择机制有效性

时间敏感的策略选择机制在智能决策模型中起到了显著作用。通过在不同时间段内对防御策略进行动态调整，使模型能够更好地适应攻击模式的变化。在训练过程中，通过对比使用时间敏感的策略选择机制和使用非时间敏感的策略选择机制（ε-greedy，以ε的概率选择一个完全随机的动作）来说明使用时间门控机制的有效性。图 9 的实验结果表明，加入时间门控机制后，智能决策模型能够在一定的时间段内有效阻止攻击行为。对于 3 个数据集，在 500~1 500 步内，考虑加入时间门控机制累积奖励相比无时间门控机制的模型有明显提升。对于 Unraveled 数据集，加入时间门控机制累积奖励相比无时间门控机制的模型平均提升了约 10.5%。对于 UNSW-NB15 数据集，加入时间门控机制累积奖励相比无时间门控机制的模型平均提升约 4.4%，对于 CICIDS2017 数据集，加入时间门控机制累积奖励相比无时间门控机制的模型平均提升约 20%。

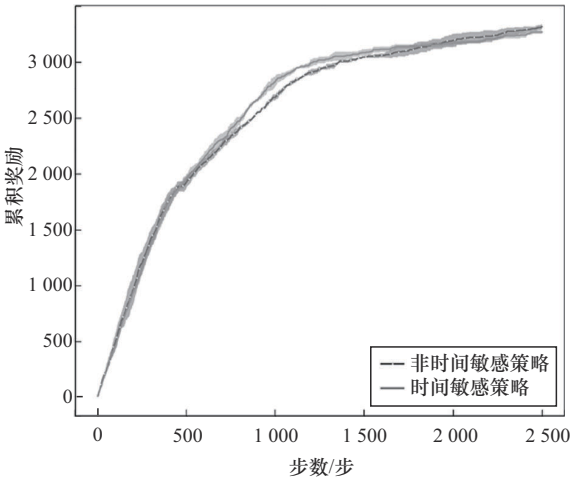
2) 异构随机探索策略生成有效性

异构随机探索策略生成机制考虑 3 种方式引入随机探索，增强了模型的优先探索能力。实验分别测试了单独采用每种随机探索方式和多种探索组合方式的防御性能。

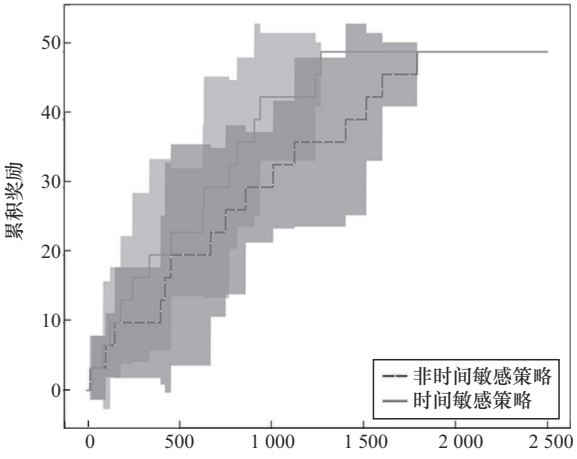
图 10 的实验结果表明，在 Unraveled 数据集上，异构随机探索机制相比单独使用的随机探索方式展现了最优的防御性能，累积奖励增长迅速且稳定。采用完全随机探索策略表现最差，累积奖励增长缓慢，且始终低于其他策略。基于密度和条件概率的随机探索策略均显著提高了累积奖励的增长速度。在 UNSW-NB15 数据集上，所有随机探索策略的累积奖励在初期阶段增长较为相似，完全随机探索策略在中期表现好一些，基于密度的随机探索策略则在后期表现较好一些。在 CICIDS2017 数据集上，基于密度和条件概率的随机探索策略在 750~1 500 步范围内的累积奖励提升较为明显，完全随机探索策略在后期的累积奖励增长较为缓慢。本文设计的多种探索方式融合的异构随机探索机制综合了 3 种随机探索策略，在整个防御步数内维持较优越的效果。由此可见，异构随机探索机制能够从总体上提高防御模型在不确定状态下的探索广度，使模型可以尝试新的防御策略，增强模型的防御动作多样性，从而更全面地防御各种攻击。



(a) Unraveled

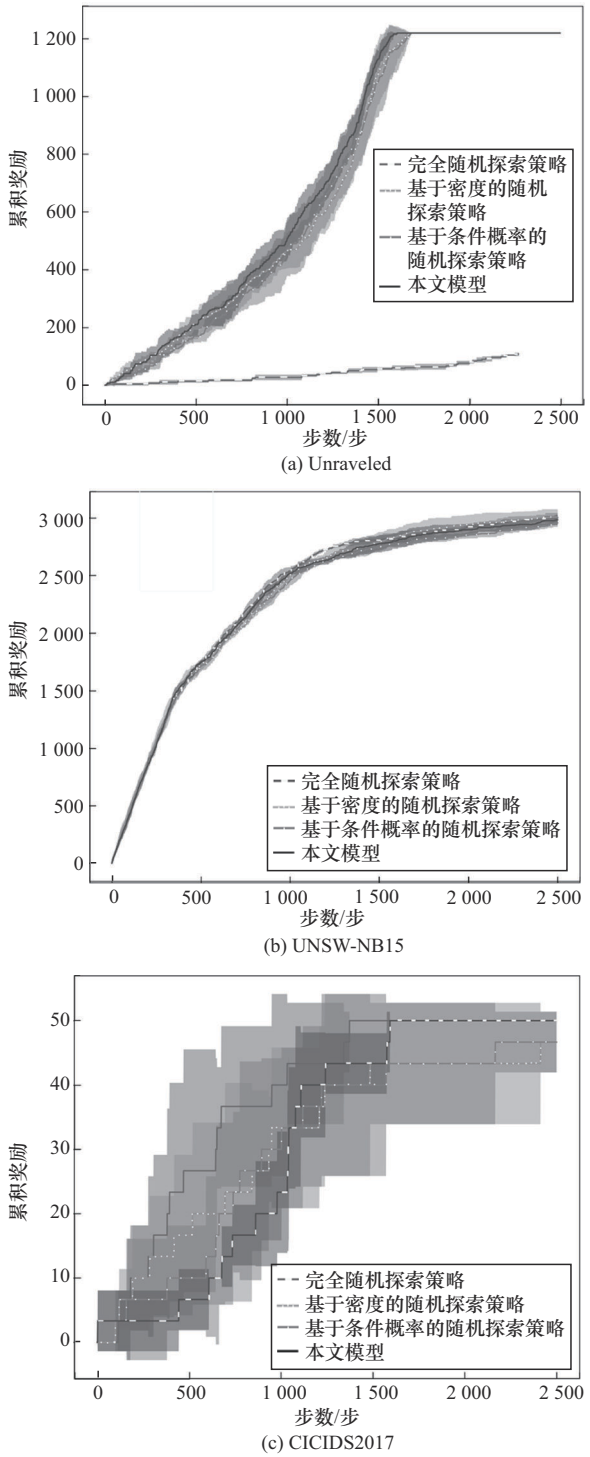


(b) UNSW-NB15



(c) CICIDS2017

图 9　时间敏感的策略选择机制有效性分析



(a) Unraveled

(b) UNSW-NB15

(c) CICIDS2017

图10　异构随机探索策略生成有效性分析

3) 异构Q网络防御策略生成有效性

除了探索机制外，异构Q网络防御策略生成在实验中也展现出卓越的有效性。相较于传统Q网络，异构Q网络能够更精准地评估当前状态和动作的价值，尤其在处理高维度和多变化的状态时表现更加出色。图11可以观察到每种Q网络随着防御步数增加获得奖励频次的情况。实验结果表明，针对不同的子网环境，模型选择的防御Q网络有着显著区别。同时，随着步数的变化，模型进行防御时前后选择Q网络准确防御的结果也不同。例如，选用序列感知Q网络策略在Unraveled数据集上在后期防御能够获得更多的奖励，而在UNSW-NB15数据集上则在前期获得的奖励较多。使用结构感知Q网络策略与使用序列感知Q网络策略有着类似的趋势。简单线性Q网络策略在Unraveled数据集上获得较少的奖励，而在UNSW-NB15数据集上则能在长期步数内能获得有效奖励。对于CICIDS2017数据集，在整个防御期使用结构感知Q网络策略能获得较多的奖励。因此，实验结果验证了采用异构Q网络能够自适应面对不同风险环境，并实现有效的防御。

3.4.3　模型效率

除了防御有效性外，模型效率也是一个重要的考量因素。在实验中，本文首先分析了智能决策模型的时间复杂度，随后对比了智能决策模型的训练时间。

智能决策模型的主要开销在于强化学习模型与环境的交互和深度Q网络的学习部分。环境交互步骤包括从环境中采样状态 $\boldsymbol{s}$ 和执行动作 α，并获得新状态和奖励。这部分时间复杂度可简化为 $TO(k)$，其中 T 是时间步，k 是常数。深度Q网络的学习部分输入动作空间大小 N 输出预测Q值，时间复杂度为 $NO(M_W)$，其中 M_W 是神经网络的规模。简单线性Q网络的时间复杂度为 $O(NW+W^2)$，其中 W 是隐藏层维度。序列感知Q网络的时间复杂度为 $O(lNW)$，其中 l 是序列长度，W 是隐藏层维度。结构感知Q网络中采用胶囊网络实现，其中卷积核提取初始特征的时间复杂度简化为 $O(W^2)$，动态路由部分的时间复杂度简化为 $O(RCW^2)$。其中 R 是迭代次数，C 是胶囊数量，W 是隐藏层维度，因此，结构感知Q网络的时间复杂度可简化为 $O(RCW^2)$，整体来看，智能决策模型融合了上述3种深度Q网络，时间复杂度小于 $O(n^3)$。虽然采用结构感知Q网络增加了计算成本，但通过减少迭代次数、并行计算等，可以在保证性能的同时降低计算复杂度。

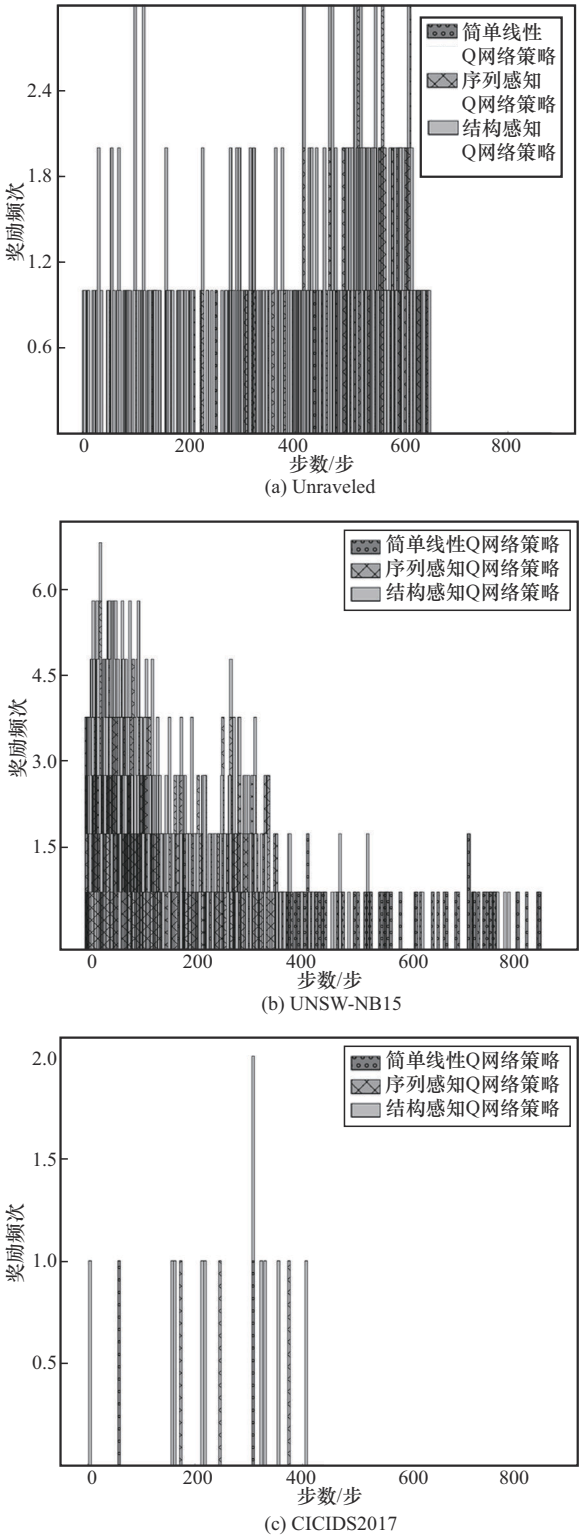


图11　异构Q网络防御策略生成有效性分析

图12为模型平均每轮训练时间对比。随机防御模型具有较快的决策速度，但由于不涉及训练，因此未进行对比。DQN和DoubleDQN虽然在防御决策速度上较快，但针对较大的数据在效率方面仍然受限。UNSW-NB15数据集包含更多的攻击边，模型训练的时间明显高于在Unraveled和CICIDS2017数据集上的训练时间。而CICIDS2017数据集包含更多的正常边，模型的训练时间并没有高于UNSW-NB15数据集，这表明模型能够快速过滤掉正常边的影响，将重点放在学习攻击相关的特征上，从而避免了因正常边数量增多而显著增加训练时间。本文模型虽然相较于DQN和DoubleDQN运行时间稍高，但对于过大的数据集UNSW-NB15和CICIDS2017而言，本文模型没有增加过多的运行时间，保持在可接受范围内，在防御精度和效率之间达到了良好的平衡。

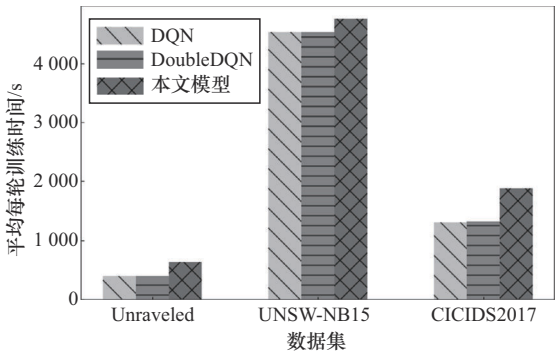


图12　模型平均每轮训练时间对比

4　结束语

针对以网络要地为目标的未知网络威胁，考虑到攻击者采用隐蔽的渗透手段导致一些现有防御机制难以有效防御未知威胁，本文提出了基于拟态防御的未知网络威胁自适应防御模型。借鉴拟态防御机制的思想，自适应防御模型基于感知的动态网络环境，构建基于强化学习的智能决策模型，动态生成子网重构防御策略，由调度控制层完成子网重构，实现拟态伪装。基于强化学习的智能决策模型同时具备时间和空间上的自适应性。在时间维度上，智能决策模型能够捕捉长期变化的环境特征，自适应选择防御策略以适应动态网络环境。在空间维度上，智能决策模型可以针对攻击模式的变化，探索并学习新的防御动作，从而有效应对多种攻击模式，提高模型整体防御效果。

实验结果显示，本文模型具有较好的防御性能。在3个不同类型的网络环境中，本文模型不仅可以快速收敛，还能最先达到最优的防御性能。此外，本文模型通过与动态变化的网络环境实时交互，

自适应调整子网重构策略，以有效应对快速演化的网络威胁和复杂的网络环境。本文模型的防御效率还存在提升空间，未来将继续针对该部分问题进行优化和提升，从整体上提高本文模型的防御效率。

参考文献：

[1] UDDIN M A, ARYAL S, BOUADJENEK M R, et al. usfAD based effective unknown attack detection focused IDS framework[J]. Scientific Reports, 2024, 14(1): 29103.

[2] 王奕, 滕克难, 程业, 等. 基于图论与PageRank的要地反导己方目标重要性排序[J]. 系统工程与电子技术, 2021, 43(3): 709-715.
WANG Y, TENG K N, CHENG Y, et al. Importance ranking of anti-missile targets in important places based on graph theory and PageRank[J]. Systems Engineering and Electronics, 2021, 43(3): 709-715.

[3] 赵文飞, 陈健, 王奕, 等. 基于强化学习的海上要地群协同防空动态火力分配[J]. 兵工学报, 2023, 44(11): 3516-3528.
ZHAO W F, CHEN J, WANG Y, et al. Dynamic firepower allocation for cooperative air defense of strategic locations on the sea based on reinforcement learning[J]. Acta Armamentarii, 2023, 44(11): 3516-3528.

[4] 中国网络安全产业联盟. 中国网络安全产业分析报告[R]. 2024.
China Cybersecurity Industry Alliance. China cybersecurity industry analysis report[R]. 2024.

[5] SHI Y, CHEN G, LI J T. Malicious domain name detection based on extreme machine learning[J]. Neural Processing Letters, 2018, 48(3): 1347-1357.

[6] YAN G H, LI Q, GUO D, et al. Discovering suspicious APT behaviors by analyzing DNS activities[J]. Sensors, 2020, 20(3): 731.

[7] WANG X, ZHENG K F, NIU X X, et al. Detection of command and control in advanced persistent threat based on independent access[C]// Proceedings of the 2016 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2016: 1-6.

[8] DEBATTY T, MEES W, GILON T. Graph-based APT detection[C]//Proceedings of the 2018 International Conference on Military Communications and Information Systems (ICMCIS). Piscataway: IEEE Press, 2018: 1-8.

[9] SHU L H, DONG S, SU H D, et al. Android malware detection methods based on convolutional neural network: a survey[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2023, 7(5): 1330-1350.

[10] DONG S, SHU L H, NIE S. Android malware detection method based on CNN and DNN bybrid mechanism[J]. IEEE Transactions on Industrial Informatics, 2024, 20(5): 7744-7753.

[11] PURVINE E, JOHNSON J R, LO C. A graph-based impact metric for mitigating lateral movement cyber attacks[C]//Proceedings of the 2016 ACM Workshop on Automated Decision Making for Active Cyber Defense. New York: ACM Press, 2016: 45-52.

[12] BOHARA A, NOUREDDINE M A, FAWAZ A, et al. An unsupervised multi-detector approach for identifying malicious lateral movement[C]// Proceedings of the 2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS). Piscataway: IEEE Press, 2017: 224-233.

[13] POWELL B A. The epidemiology of lateral movement: exposures and countermeasures with network contagion models[J]. Journal of Cyber Security Technology, 2020, 4(2): 67-105.

[14] DONG S, XIA Y J, WANG T. Network abnormal traffic detection framework based on deep reinforcement learning[J]. IEEE Wireless Communications, 2024, 31(3): 185-193.

[15] KHALID M N A, AL-KADHIMI A A, SINGH M M. Recent developments in game-theory approaches for the detection and defense against advanced persistent threats (APTs): a systematic review[J]. Mathematics, 2023, 11(6): 1353.

[16] NOUREDDINE M A, FAWAZ A, SANDERS W H, et al. A game-theoretic approach to respond to attacker lateral movement[C]//Decision and Game Theory for Security. Berlin: Springer International Publishing, 2016: 294-313.

[17] RASS S, ZHU Q Y. GADAPT: a sequential game-theoretic framework for designing defense-in-depth strategies against advanced persistent threats[C]//International Conference on Decision and Game Theory for Security. Berlin: Springer International Publishing, 2016: 314-326.

[18] FOLEY M, HICKS C, HIGHNAM K, et al. Autonomous network defence using reinforcement learning[C]//Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security. New York: ACM Press, 2022: 1252-1254.

[19] WAQAS M, TU S S, WAN J L, et al. Defense scheme against advanced persistent threats in mobile fog computing security[J]. Computer Networks, 2023, 221: 109519.

[20] ZHU T Q, YE D Y, CHENG Z S, et al. Learning games for defending advanced persistent threats in cyber systems[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2023, 53(4): 2410-2422.

[21] LI H L, WU J, XU H S, et al. Explainable intelligence-driven defense mechanism against advanced persistent threats: a joint edge game and AI approach[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(2): 757-775.

[22] WU J X. Cyberspace mimic defense: generalized robust control and endogenous security[M]. Berlin: Springer International Publishing, 2020.

[23] 邬江兴. 网络空间内生安全: 拟态防御与广义鲁棒控制(下册)[M]. 北京: 科学出版社, 2020.
WU J X. Cyberspace endogenous security: mimic defense and generalized robust control[M]. Beijing: Science Publishing Company, 2020.

[24] 陈双喜, 吴安邦, 岐舒骏, 等. 新型主动防御框架的资源对抗模型分析[J]. 电子学报, 2019, 47(7): 1557-1565.
CHEN S X, WU A B, QI S J, et al. Analysis of resource defense model for novel active defense modeling[J]. Acta Electronica Sinica, 2019, 47(7): 1557-1565.

[25] 朱正彬, 刘勤让, 刘冬培, 等. 拟态多执行体调度算法研究进展[J]. 通信学报, 2021, 42(5): 179-190.
ZHU Z B, LIU Q R, LIU D P, et al. Research progress of mimic multi-execution scheduling algorithm[J]. Journal on Communications, 2021, 42(5): 179-190.

[26] 沈丛麒, 陈双喜, 吴春明, 等. 基于信誉度与相异度的自适应拟态控制器研究[J]. 通信学报, 2018, 39(S2): 173-180.
SHEN C Q, CHEN S X, WU C M, et al. Research on adaptive mimicry controller based on credibility and dissimilarity[J]. Journal on Communications, 2018, 39(S2): 173-180.

[27] 王禛鹏, 扈红超, 程国振. MNOS: 拟态网络操作系统设计与实现[J]. 计算机研究与发展, 2017, 54(10): 2321-2333.
WANG Z P, HU H C, CHENG G Z. Design and implementation of mimic network operating system[J]. Journal of Computer Research

and Development, 2017, 54(10): 2321-2333.

[28] 徐蜜雪, 苑超, 王永娟, 等. 拟态区块链: 区块链安全解决方案[J]. 软件学报, 2019, 30(6): 1681-1691.
XU M X, YUAN C, WANG Y J, et al. Mimic blockchain: solution to the security of blockchain[J]. Journal of Software, 2019, 30(6): 1681-1691.

[29] 朱勇刚, 孙艺夫, 姚富强, 等. 基于多智能超表面的信道空间内生抗干扰方法[J]. 通信学报, 2023, 44(10): 13-22.
ZHU Y G, SUN Y F, YAO F Q, et al. Channel-space endogenous anti-jamming method based on multi-reconfigurable intelligent surface[J]. Journal on Communications, 2023, 44(10): 13-22.

[30] GILL K S, SAXENA S, SHARMA A. GTM-CSec: game theoretic model for cloud security based on IDS and honeypot[J]. Computers & Security, 2020, 92: 101732.

[31] XIAO L, XU D J, MANDAYAM N B, et al. Attacker-centric view of a detection game against advanced persistent threats[J]. IEEE Transactions on Mobile Computing, 2018, 17(11): 2512-2523.

[32] HUANG L N, ZHU Q Y. A dynamic games approach to proactive defense strategies against advanced persistent threats in cyber-physical systems[J]. Computers & Security, 2020, 89: 101660.

[33] DEHGHAN M, SADEGHIYAN B, KHOSRAVIAN E, et al. ProAPT: projection of APT threats with deep reinforcement learning[J]. arXiv Preprint, arXiv: 2209.07215, 2022.

[34] NING B F, XIAO L. Defense against advanced persistent threats in smart grids: a reinforcement learning approach[C]//Proceedings of the 2021 40th Chinese Control Conference (CCC). Piscataway: IEEE Press, 2021: 8598-8603.

[35] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[36] MYNENI S, JHA K, SABUR A, et al. Unraveled: a semi-synthetic dataset for advanced persistent threats[J]. Computer Networks, 2023, 227: 109688.

[37] MOUSTAFA N, SLAY J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)[C]//Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS). Piscataway: IEEE Press, 2015: 1-6.

[38] SHARAFALDIN I, LASHKARI A H, GHORBANI A A. Toward generating a new intrusion detection dataset and intrusion traffic characterization[C]//Proceedings of the 4th International Conference on Information Systems Security and Privacy. SCITEPRESS-Science and Technology Publications, 2018: 108-116.

[作者简介]

郝宵荣（1995– ），女，山西晋中人，东南大学博士生，主要研究方向为网络行为分析与安全防御、复杂网络、动态图表示学习、图异常检测等。

刘波（1975– ），女，河南南阳人，博士，东南大学教授、博士生导师，主要研究方向为网络异常检测、网络舆情分析、内容安全研究、网络大数据行为分析等。

周鼎（1991– ），男，河南开封人，博士，紫金山实验室系统架构研究员，主要研究方向为网络信息物理系统安全和防御。

曹玖新（1967– ），男，河南洛阳人，博士，东南大学教授、博士生导师，主要研究方向为大数据智能处理与内容安全、大数据安全与隐私保护。

张进（1979– ），男，江苏镇江人，博士，紫金山实验室高级工程师，主要研究方向为宽带信息网络、网络安全、软硬件协同设计。