基于时滞FlipIt博弈与多智能体强化学习的欺骗防御时机选取方法

裴金川[1],胡宇翔[1,2,3*],于洪涛[1],田 乐[1,2,3],李梦龙[1,2,3]

(1. 信息工程大学,河南郑州 450002;2. 河南省网络空间内生安全重点实验室,河南郑州 450000;
3. 网络空间安全教育部重点实验室,河南郑州 450000)

摘 要: 云边协同网络在6G通信、工业互联网等关键领域中广泛应用,其开放分布式特性面临持续演进的网络攻击威胁。网络欺骗防御通过蜜罐、诱饵等资源干扰攻击者认知,但其有效性高度依赖于欺骗资源的轮换时机。然而,在实际攻防过程中,欺骗资产部署、攻击渗透扩散均存在不可忽略的状态转换时滞,且攻防双方存在持续的策略交互,现有方法或假设状态转移瞬时完成,或采用固定周期等静态机制,忽视了真实攻防过程中存在的状态转换时滞及攻防双方的动态交互,无法准确刻画攻防状态演化的时序特征,导致策略在真实高动态环境中适应性不足。针对上述问题,本文提出一种基于时滞FlipIt博弈与多智能体强化学习的欺骗防御时机选取方法。首先,分析网络安全状态演化特性,将正常、保护、感染、受损四种节点状态间的六条转移路径分别引入对应的时滞参数,构建融合状态转换时滞的时滞微分方程组,形成更贴近实际物理过程的网络状态演化模型,弥补了传统SIS、SIR等动力学模型忽略时滞因素的缺陷。其次,分别建立攻击者与防御者模型,将攻防对抗形式化为具有时序特征的博弈过程,在此基础上提出时滞FlipIt博弈模型,显式刻画策略下发到状态生效之间的延迟机制,并给出攻防双方在时滞约束下的效用函数。然后采用多智能体近端策略优化(Multi-Agent Proximal Policy Optimization,MAPPO)算法,构建中心化训练、分布式执行的强化学习框架,设计状态空间、动作空间与奖励函数,通过攻防智能体的持续交互与策略演化,求解最优欺骗防御时机轮换策略。实验在Mininet仿真环境与小规模真实测试床上进行,结果表明所提方法能够有效优化防御时机选择,保护状态节点比例稳定在83.2%,攻击者效用值较MFD-PPO方法降低44.85%,攻击拦截成功率较FP方法提升38.4%,同时,在同等保护状态节点比例下IP跳变频率显著低于对比方法,验证了所提方法在保障防御效能与降低资源开销方面的综合优势。

关键词: 云边协同;欺骗防御;时滞;FlipIt博弈;多智能体强化学习;防御时机

基金项目: 国家重点研发计划(No.2024YFB2906704, No.2023YFB2903902);先进通信网全国重点实验室基金课题(No.FFX24641X028)

中图分类号: TP393;TN915.08 **文献标识码:** A **文章编号:** 0372-2112(2026)04-1823-12

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20260221

A Deception Defense Timing Selection Method Based on FlipIt Game with Time Delay and Multi-Agent Reinforcement Learning

PEI Jinchuan[1], HU Yuxiang[1,2,3*], YU Hongtao[1], TIAN Le[1,2,3], LI Menglong[1,2,3]

(1. *Information Engineering University*, *Zhengzhou*, *Henan* 450002, *China*;
2. *Key Laboratory of Cyberspace Endogenous Safety & Security of Henan Province*, *Zhengzhou*, *Henan* 450000, *China*;
3. *Key Laboratory of Cyberspace Security Ministry of Education of China*, *Zhengzhou*, *Henan* 450000, *China*)

Abstract: The cloud-edge collaborative network is widely applied in key fields such as 6G communication and industrial internet. Its open and distributed characteristics are facing continuous evolving network attack threats. Network deception defense uses resources such as honeypots and decoys to interfere with the attacker's cognition, but its effectiveness highly depends on the timing of the rotation of deception resources. However, in actual offensive and defensive processes, there are unignorable state transition delays in the deployment of deception assets and the spread of attack penetration, and both the attacker and defender have continuous strategic interactions. Existing methods or assumptions that the state transition is instantaneous or adopt static mechanisms such as fixed cycles ignore the state transition delays and the dynamic interaction between the attacker and defender in the actual offensive and defensive process, making it impossible to accurately depict the temporal characteristics of the evolution of the offensive and defensive states, resulting in insufficient adaptability

收稿日期:2026-03-30;录用日期:2026-04-14;责任编委:崔来中
*通信作者:胡宇翔

of the strategies in the real high-dynamic environment. To address these issues, this paper proposes a deception defense timing selection method based on time-delay FlipIt game and multi-agent reinforcement learning. Firstly, it analyzes the evolution characteristics of network security states and introduces six transfer paths between four node states (normal, protected, infected, and damaged) into corresponding time-delay parameters to construct a time-delay differential equation system that integrates state transition time delays, forming a network state evolution model that is closer to the actual physical process and compensates for the shortcomings of traditional SIS, SIR, etc. dynamic models that ignore time-delay factors. Secondly, it establishes models for the attacker and defender, formalizes the offensive and defensive confrontation as a game process with temporal characteristics, and proposes a time-delay FlipIt game model to explicitly depict the delay mechanism between the strategy being issued and taking effect in the state, and gives the utility functions of both the attacker and defender under the time-delay constraints. Then, using the multi-agent proximal policy optimization (MAPPO) algorithm, a centralized training and distributed execution reinforcement learning framework is constructed, designing the state space, action space, and reward function, and solving the optimal deception defense timing rotation strategy through the continuous interaction and strategy evolution of the offensive and defensive agents. Experiments are conducted in the Mininet simulation environment and a small-scale real testbed. The results show that the proposed method can effectively optimize the selection of defense timing, maintaining a stable protection state node ratio of 83.2%, reducing the attacker's utility value by 44.85% compared to the MFD-PPO method, increasing the attack interception success rate by 38.4% compared to the FP method, and significantly reducing the IP hop frequency compared to the comparison method under the same protection state node ratio, verifying the comprehensive advantages of the proposed method in ensuring defense effectiveness and reducing resource overhead.

Keywords: cloud-edge collaboration; deception defense; time delay; FlipIt game; multi-agent reinforcement learning; defense timing

Foundation Item(s): National Key Research and Development Program of China (No.2024YFB2906704, No.2023YFB2903902); Advanced Communication Network National Key Laboratory Research Project Fund (No.FFX24641X028)

0 引言

随着6G通信与工业互联网的飞速发展,云边协同[1]通过云端强大算力与边缘实时响应能力实现了资源的动态协同调度,为6G[2]、工业互联网[3]等关键领域提供了核心支撑。然而,云边协同网络具有拓扑动态、节点异构、跨域耦合等特性[4],这种高度开放、复杂协同的环境使其面临持续演进的网络攻击威胁。攻击者作为具备个体理性的智能体,可利用分布式节点间的管理接口与数据流发起横向渗透,发起动态且持续演进的攻击,严重威胁系统稳定与数据安全[5]。

欺骗防御[6]作为网络主动防御技术,通过在网络环境中部署欺骗防御资源来构建虚假的攻击面,主动干扰攻击者的认知,增加其攻击难度与成本,可在云边协同网络中用以应对网络攻击。云边协同网络的安全防御本质上是一个多智能体在对抗环境下的协同决策问题,涉及攻防双方的动态博弈过程,博弈论[7]作为一种行为策略建模的数学工具,可为攻防决策提供理论支撑,博弈中的基本要素与云边协同网络的欺骗防御存在映射关系,如博弈的局中人、状态、信息和策略集可分别映射为攻防对抗双方、网络状态、攻防信息和攻防策略[8],可有效增强欺骗防御技术的策略性。

然而,在云边协同的高动态网络环境中,欺骗防御策略的有效性不仅取决于防御资源的部署,更取决于欺骗防御时机的有效轮换[9],且防御动作的生效往往伴随着物理时滞,若时机选择不当,不仅无法诱捕攻击,反而可能因资源无效调度导致系统震荡。Chowdhury等人[10]指出时间是网络安全行为背后的重要驱动因素之一,随后Zhang等人[11]建立了一个攻防框架来分析网络安全中的时间弹性问题。Merlevede等人[12]通过引入时间指数折扣机制以扩展攻防模型,详细描述了网络安全模型中的时机决策问题。Zhang等人[13]将微分博弈的连续时间特征与马尔可夫决策的状态转移机制相结合,以刻画策略激活过程中的时域随机特性,进而构建多阶段对抗模型。Farhang等人[14]针对时序化攻防对抗场景,在周期性策略假设基础上建立收益函数模型,并将保护、检测与响应三阶段时间参数纳入安全博弈框架,为防御者确定最优防御重置时机提供有效办法。然而,上述基于时间维度决策的方法多采用固定周期或随机触发机制[13-14],这种静态的防御触发机制难以在动态的攻防博弈中捕捉到最优的干预窗口。

为了提升欺骗防御的适应性,需对攻防双方的动态博弈关系进行建模,并确定欺骗防御资源的轮换时机。FlipIt博弈[15]的时序特征能够有效建模欺骗防御时机选取问题,其将攻防双方对隐蔽资源的控制权争夺抽象为抢占(Flip)动作,而抢占时机则决定了防御

收益[16]。何威振等人[17]提出基于深度强化学习的多阶段FlipIt博弈网络拓扑欺骗防御方法,通过构建欺骗防御模型并引入折扣因子与状态转移概率,实现时空维度上的最优防御策略求解。Tan等人[18]采用SIRM传染病模型构建攻击面状态转移模型,进而建立FlipIt博弈驱动的攻防时机决策框架FG-MTD,并设计最优时机选择算法验证不同攻防策略的时机规律。Zhu等人[19]将SIR传播模型与FlipIt博弈融合,用以求解最佳防御动作时机,并发现攻击周期与防御周期之间的相对大小关系直接影响防御效果。熊鑫立等人[20]将FlipIt博弈模型进行扩展,提出Cheat-FlipIt博弈模型,引入攻击者的欺骗能力,并通过深度强化学习求解Cheat-FlipIt博弈中的最优防御策略。

然而,上述现有研究模型简化了攻防交互的时间过程,假设攻防动作发生后能瞬时改变系统状态,且现有网络安全动力学模型如SIS、SIR等[21-22],普遍假设安全状态转移为瞬时过程,忽略了真实攻防操作中普遍存在的状态转换时滞(Time Delay, TD)[23]。例如,欺骗资产的部署、攻击的渗透扩散等环节均需一定时间完成,这些时滞因素不仅影响防御响应的及时性,也可能被攻击者利用,例如在防御重置或策略更新的窗口发起精准攻击。时滞的不确定性在攻防双方博弈中起到关键作用,可使防御方从被动响应转向主动预测,比如通过预判攻击者的行动窗口来调整欺骗策略的轮换节奏;忽略滞因将导致策略误判,难以准确反映状态转移的时序特征,从而导致防御策略在高动态环境中的适应性不足。强化学习是求解复杂的动态攻防博弈的有效方法,其中,何威振等人[17]的工作中采用的单智能体强化学习方法将攻击者视为被动环境,而非具备独立决策能力的理性主体,难以体现动态攻防场景中的复杂行为交互;多智能体强化学习(Multi-Agent Reinforcement Learning, MARL)则为动态策略求解提供了可行路径,多智能体强化学习不仅能显式建模攻防双方作为独立理性智能体的策略互动,其通过与环境持续交互学习的特性,这也使其天然具备从历史经验中学习并适应状态转换时滞规律的潜力,从而为求解此类复杂的时序博弈问题提供了有力工具。然而现有MARL应用于攻防博弈的工作[24]同样未能将状态转换时滞有效嵌入智能体的决策过程中,限制了策略在真实环境中的有效性与鲁棒性。

综上所述,防御方须在状态转换时滞不确定的条件下,通过权衡防御收益与防御轮换成本,寻求最优的欺骗防御时机轮换策略。基于此,本文提出基于时滞FlipIt博弈与多智能体强化学习的欺骗防御时机选取方法TD-FlipIt-MAPPO。

本文主要工作如下:

(1)分析网络安全状态演化过程,结合时滞因素改进网络安全传播动力学模型,构建耦合状态演化的时滞微分方程组,以解决现有动力学模型忽略时滞的问题。

(2)将欺骗攻防过程建模为具有时序特征的博弈模型,设计时滞FlipIt博弈模型,以形式化刻画攻防双方在考虑不确定时滞因素下的攻防博弈过程。

(3)采用多智能体近端策略优化算法(Multi-Agent Proximal Policy Optimization, MAPPO)求解欺骗防御时机轮换策略,从而弥补单智能体方法无法刻画攻防动态交互的不足,实验结果表明,该方法能够有效优化防御时机的选择,在保障防御效能的同时降低资源开销。

1　网络安全状态演化

网络欺骗防御为攻防双方的多阶段持续交互过程,本节构建了一种基于时滞微分方程的网络安全状态演化模型。设云边协同网络中的任意节点在任一时刻t处于且仅处于正常、保护、感染及受损状态四种离散状态之一,四种节点比例分别为$S_{NS}(t)$、$S_{PS}(t)$、$S_{IS}(t)$和$S_{DS}(t)$,其中正常状态节点未部署欺骗资源,但未受攻击影响,正常运行网络功能,保护状态节点已部署欺骗防御资源,可干扰攻击者,感染状态节点已被攻击者渗透,但尚未完全失控,仍可提供网络服务,攻击者可将其做跳板攻击其他节点,受损状态节点完全失控,数据被窃取且服务不可用。

1.1　状态演化路径

网络节点的安全状态并非瞬时切换,而是呈现出具有显著时滞特性的动态演化过程,例如防御方部署蜜罐、诱饵等欺骗资产后,需在一定时间内完成资源配置与服务启动,方可使节点进入保护状态;同样,攻击者对节点的攻击需经历侦察、提权等阶段,导致正常节点不会立即转变为感染或受损状态。

在云边协同网络的攻防对抗过程中,不同网络状态的节点之间可相互转化,网络安全状态转换关系如图1所示,存在六条状态转移路径,分别为正常状态到保护状态、保护状态到正常状态、保护状态到感染状态、正常状态到感染状态、感染状态到保护状态、感染状态到受损状态,每条转移路径存在对应的转移系数。

(1)正常状态→保护状态:该路径为防御方对正常节点部署欺骗资源的过程,当防御策略触发时,节点并非立即进入保护状态,而是经历一个部署时滞τ_{NP},即从欺骗资源部署指令下发到实际生效存在时间滞后。该时滞反映了从指令下发、资源配置到服务

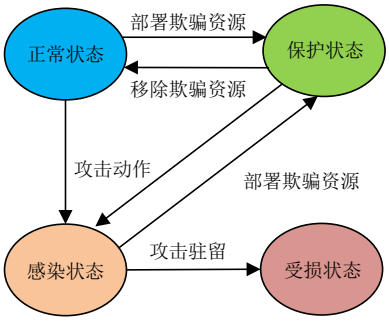


图1 网络安全状态演化示意图

Figure 1 Evolution diagram of cybersecurity status

激活的完整物理过程,对应路径转移系数为λ_{NP}。

(2)正常状态→感染状态:针对未受保护的正常节点,攻击者发起的漏洞利用或横向移动等渗透行为,需经历侦察与权限提升阶段。因此,节点从遭受攻击到实际转化为感染状态存在一个潜伏时滞τ_{NI},对应的转移系数为λ_{NI}。

(3)保护状态→正常状态:当保护状态下的欺骗策略失效(如被攻击者识破)或因资源轮换策略被主动移除时,节点将回归正常状态,这一过程同样存在响应时滞τ_{PN},对应的转移系数为λ_{PN}。

(4)保护状态→感染状态:若攻击者识别出防御者部署的欺骗资源并成功攻击保护状态节点,节点则从保护状态经过时滞τ_{PI}转化为感染状态,对应的转移系数为λ_{PI}。

(5)感染状态→保护状态:当节点被感染后,防御方通过系统重置并重新部署欺骗资源以夺回控制权,在启动防御恢复响应延迟τ_{IP}后将其转化为受保护状态,对应的转移系数为λ_{IP}。

(6)感染状态→受损状态:当节点处于感染状态时,若防御方未能及时响应,攻击者将执行渗透操作,逐步获取系统最高权限攻击者,导致该节点经过时滞τ_{ID}转换为受损状态,对应的转移系数为λ_{ID}。

1.2 基于时滞的攻防动力学模型

在云边协同网络中,攻防博弈的动态过程往往伴随着不可忽略的时间滞后效应,且这种时滞具有一定的随机性与不确定性。为了更精确地捕捉网络安全状态在不同节点间的演化规律,本节在经典传播动力学模型的基础上进行了扩展,引入时滞项并建立时滞微分方程来刻画这一复杂的动态过程。

基于上述状态转移路径,为系统刻画时滞对安全状态演化的影响,构建如下能反映攻防过程中状态转移时序特性的微分方程组:

$$\frac{\mathrm{d}S_{NS}(t)}{\mathrm{d}t}=-\lambda_{NP}S_{NS}(t-\tau_{NP})\mathrm{e}^{-\tau_{NP}}-\lambda_{NI}S_{NS}(t-\tau_{NI})\mathrm{e}^{-\tau_{NI}}+\lambda_{PN}S_{PS}(t-\tau_{PN})\mathrm{e}^{-\tau_{PN}} \tag{1}$$

$$\frac{\mathrm{d}S_{PS}(t)}{\mathrm{d}t}=\lambda_{NP}S_{NS}(t-\tau_{NP})\mathrm{e}^{-\tau_{NP}}+\lambda_{IP}S_{IS}(t-\tau_{IP})\mathrm{e}^{-\tau_{IP}}-\lambda_{PN}S_{PS}(t-\tau_{PN})\mathrm{e}^{-\tau_{PN}}-\lambda_{PI}S_{PS}(t-\tau_{PI})\mathrm{e}^{-\tau_{PI}} \tag{2}$$

$$\frac{\mathrm{d}S_{IS}(t)}{\mathrm{d}t}=\lambda_{PI}S_{PS}(t-\tau_{PI})\mathrm{e}^{-\tau_{PI}}+\lambda_{NI}S_{NS}(t-\tau_{NI})\mathrm{e}^{-\tau_{NI}}-\lambda_{IP}S_{IS}(t-\tau_{IP})\mathrm{e}^{-\tau_{IP}}-\lambda_{ID}S_{IS}(t-\tau_{ID})\mathrm{e}^{-\tau_{ID}} \tag{3}$$

$$\frac{\mathrm{d}S_{DS}(t)}{\mathrm{d}t}=\lambda_{ID}S_{IS}(t-\tau_{ID})\mathrm{e}^{-\tau_{ID}} \tag{4}$$

式(1)~(4)可描述时滞因素下网络节点状态的演化过程,其中引入的指数项$\mathrm{e}^{-\tau}$并非表示时滞本身的指数衰减,而是用于刻画在延迟期间系统状态可能发生变化的概率补偿机制,该形式借鉴了时滞系统建模中的有效转移率概念,即状态转移的发生不仅依赖于历史状态,还受到延迟区间内系统动态的影响,因此引入指数衰减因子以描述这一过程。状态之间的转换不仅依赖于当前攻防强度,更受历史状态与操作延迟的共同影响,有效的欺骗防御需考虑时滞因素,当前t时刻的网络状态转移并非取决于即时攻防行为,而是由$t-\tau$前的历史动作延迟驱动,这种滞后性导致防御者观测到的安全状态实际是历史攻击行为的延迟效应,因此传统的基于事件的实时响应策略已失效,防御者须对时滞窗口内的攻击预测进行前置防御,故时机选择成为防御效能的核心,即通过预测攻击窗口抢占节点状态演化的主动权,在攻击生效前完成欺骗资源的部署与轮换。此外,在网络状态转移的全过程中,四种节点状态比例之和恒为1,如式(5)所示:

$$S_{NS}(t)+S_{PS}(t)+S_{IS}(t)+S_{DS}(t)=1 \tag{5}$$

在此框架下,防御者的核心目标转化为在有限资源约束下,结合状态转移的延迟特性,动态规划欺骗行动的轮换时机,以抑制攻击者的控制窗口并最大化长期防御效用。引入时滞因素的状态演化模型不仅更贴近云边协同网络的实际运行机制,也为后续攻防博弈建模提供了动态且带时滞约束的状态空间。

2 攻防模型

2.1 攻击者模型

攻击者建模为具备个体理性的智能体,作为外部威胁实体由边缘层发起攻击,如图2所示为攻击者渗透的典型攻击链,首先扫描并入侵边缘节点,利用其暴露的攻击面及潜在漏洞建立初始立足点,继而实施横向移动,通过解析跨域连接关系逐步渗透至云层关键节点,最终达成数据窃取或持久化控制云边协同系统。在此架构中,攻击者行为构成驱动防御系统进行动态策略调整与时机决策的外部博弈输入。

为形式化攻击者的动作空间,假设攻击者对目标

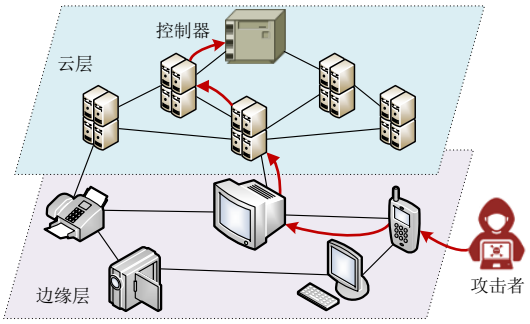


图2　攻击者渗透示意图

Figure 2　Schematic diagram of attacker penetration

的控制权的时间间隔为服从参数为 $\lambda>0$ 的指数分布,则攻击者获得目标控制权的时间间隔的概率分布为 $p(t)=\lambda e^{-\lambda t}(t>0)$,其中,参数 λ 为攻击者的攻击强度因子,其取值直接决定攻击频率与渗透速度,λ 越大,攻击者的攻击频率越高、渗透速度越快,但伴随更高的攻击成本;反之则攻击频率低、成本低。攻击者的核心动作是调整攻击强度因子 λ,通过改变 λ 调控节点状态向感染、受损状态的转移系数,实现攻击策略的动态调整,转移系数的调控关系分别为

$$\lambda_{\mathrm{NI}}=\frac{\lambda S_{\mathrm{NS}}(t)}{1+S_{\mathrm{IS}}(t)} \tag{6}$$

$$\lambda_{\mathrm{PI}}=\frac{\lambda S_{\mathrm{PS}}(t)}{1+S_{\mathrm{IS}}(t)} \tag{7}$$

$$\lambda_{\mathrm{ID}}=\frac{\lambda S_{\mathrm{IS}}(t)}{1+S_{\mathrm{DS}}(t)} \tag{8}$$

2.2　防御者模型

防御者同样为具备个体理性的智能体,云边协同防御架构中,为应对攻击时序不对称、行为延迟及资源受限所引发的动态防御问题,云层负责防御轮换时机策略生成与时序规划,策略的下发和执行生效之间存在不可避免的时滞 τ,包括指令下发、资源配置以及状态生效等延迟。由此可知,网络攻防过程中云边协同网络作为防御者,响应攻击者动作,下发欺骗时机轮换策略后存在相应时延,具有不确定性,而云层中的控制器基于历史攻防轨迹与当前网络状态,采用时滞 FlipIt 博弈模型与多智能体强化学习算法求解博弈策略,动态计算最优欺骗资源轮换时机,并将策略分发到边缘层设备,同时接收来自边缘层的环境状态反馈,以此迭代更新博弈模型参数,在攻防双方的持续对抗中不断优化欺骗防御时机决策,实现防御策略与攻击行为的动态适配。

防御者的核心动作是欺骗资源轮换动作,在时刻 t 的策略可映射为二元动作序列 $\pi_{\mathrm{D}}(t)\in\{0,1\}$,其中 $\pi_{\mathrm{D}}(t)=1$ 表示防御者在时刻 t 发起欺骗资源轮换动作,重置攻击面;$\pi_{\mathrm{D}}(t)=0$ 表示防御者在时刻 t 不执行任何轮换动作,维持当前欺骗资源部署状态。

3　基于时滞 FlipIt 博弈的欺骗防御模型

为刻画云边协同网络中攻防双方在控制权争夺过程中的时序依赖和响应延迟特性,本节在经典 FlipIt 博弈框架基础上,融入状态转换时滞因素,构建时滞 FlipIt 博弈(Time-Delayed FlipIt, TD-FlipIt)模型,通过形式化定义博弈要素与优化目标,实现对欺骗防御时机的精准建模。

FlipIt 博弈由攻击方局中人、防御方局中人和系统资源 3 部分组成,是描述攻防控制权争夺的连续时间博弈模型,博弈双方通过离散的“翻转”动作竞争资源控制权,控制持续时间越长收益越高,但经典 FlipIt 模型假设动作立即生效,忽略了实际攻防中策略下发与状态响应间普遍存在的时间延迟 τ。为此,本节构建的 TD-FlipIt 博弈模型显式引入时滞参数以更贴近实际攻防场景。

图 3 为 TD-FlipIt 博弈的基本时序关系,攻防双方交替争夺系统资源控制权,图中带圆圈的箭头代表攻防动作的执行时刻,攻防双方做出响应动作后均需经过时滞 τ 才能实际获取资源控制权。因此,防御策略从下发到网络节点状态的改变,需要完整经历状态转移延迟,这种延迟机制从根本上重构了防御时机的决策逻辑,防御者不再是对攻击事件做出被动响应式抢夺,而是要在状态演化的时滞约束下,提前预判攻击者的攻击窗口,通过预测性抢占,在攻击生效前完成欺骗资源轮换,从而掌握对抗主动权。

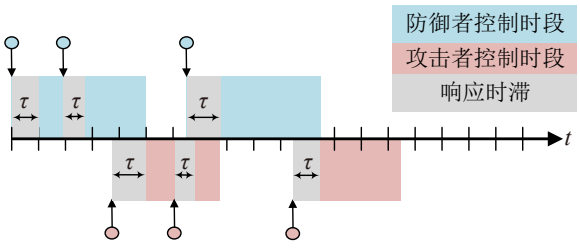


图3　TD-FlipIt 博弈示意图

Figure 3　TD-FlipIt game diagram

TD-FlipIt 博弈模型遵循以下假设。其一,初始时刻 $t=0$ 时,防御者拥有资源控制权。其二,若一方已控制资源,再次发起行动不会改变控制权,但仍会消耗相应成本,直至另一方采取行动夺取控制权。因此,为确保持续掌控资源,攻防双方均可能在已控制资源期间执行无效行动。其三,若防御者和攻击者同时执行翻转控制动作,则遵循“防御优先”原则,由防御者获得控制权。

TD-FlipIt 博弈模型可表示为元组 $\{N,S,\tau,\Pi,U,G\}$。

(1)博弈参与者 $N=(N_{\mathrm{D}}(t),N_{\mathrm{A}}(t))$ 表示防御者与攻

击者两个理性智能体，当$N_{\mathrm{A}}(t)=1$且$N_{\mathrm{D}}(t)=0$时表示当前时刻系统资源被攻击者控制，反之则被防御者控制。

（2）系统状态$S=\left(S_{\mathrm{NS}}(t),S_{\mathrm{PS}}(t),S_{\mathrm{IS}}(t),S_{\mathrm{DS}}(t)\right)$表示时刻$t$下正常、保护、感染、受损四类状态节点的比例，其动态演化由时滞微分方程组驱动，受攻防双方策略交互影响。

（3）时滞集合$\tau=(\tau_1,\tau_2,\cdots,\tau_T)$刻画攻防双方博弈全过程中发起动作从执行到生效的响应延迟，其中τ_i包含四类状态之间的转换时延集合

$$\left\{\tau_{\mathrm{NP}},\tau_{\mathrm{PN}},\tau_{\mathrm{PI}},\tau_{\mathrm{NI}},\tau_{\mathrm{IP}},\tau_{\mathrm{ID}}\right\}。$$

（4）策略集合$\Pi=(\Pi_{\mathrm{D}},\Pi_{\mathrm{A}})$为攻防双方的决策空间，防御者策略$\Pi_D$是动态选择欺骗资源（如蜜罐、虚假服务）的轮换时机，即决策在时刻t是否发起轮换动作以重置攻击面；攻击者策略Π_{A}是调节控制权争夺的时间间隔分布，通过调整攻击强度参数λ优化其控制收益。

（5）效用函数$U=(U_{\mathrm{D}},U_{\mathrm{A}})$考虑时滞不确定性与攻防异步性，时滞具有动态不确定性，且FlipIt博弈中攻防双方的操作可能不在同一时间点发生，这种异步博弈会影响双方的收益，因此防御者与攻击者的效用函数基于长期累积收益构建，考虑资源控制权持有时间与行动成本，U_{D}为防御者效用函数。

$$U_{\mathrm{D}}=\int_0^T\left(r_{\mathrm{NS}}\cdot S_{\mathrm{NS}}(t)+r_{\mathrm{PS}}\cdot S_{\mathrm{PS}}(t)\right)\mathrm{d}t-c_{\mathrm{d}}\cdot K_{\Pi_{\mathrm{D}}}\quad(9)$$

其中，r_{NS}和r_{PS}分别为单位时间内防御者控制正常或保护状态节点的收益，c_{d}为单次欺骗防御轮换的资源成本，$K_{\Pi_{\mathrm{D}}}$为防御者策略Π_{D}在$[0,T]$内执行轮换动作的次数。U_{A}为攻击者效用函数，同样可表示为

$$U_{\mathrm{A}}=\int_0^T\left(r_{\mathrm{IS}}\cdot S_{\mathrm{IS}}(t)+r_{\mathrm{DS}}\cdot S_{\mathrm{DS}}(t)\right)\mathrm{d}t-c_{\mathrm{a}}\cdot\lambda T\quad(10)$$

其中，r_{IS}和r_{DS}分别为单位时间内攻击者控制感染或受损状态节点的收益，c_{a}为攻击者单次攻击动作控制目标资源的成本，λT为攻击者在$[0,T]$时间内控制目标资源的次数。

（6）G为博弈约束条件，体现真实攻防场景的物理限制与行为规则，包括状态归一化约束和时滞非负约束，状态归一化约束即$S_{\mathrm{NS}}(t)$、$S_{\mathrm{PS}}(t)$、$S_{\mathrm{IS}}(t)$和$S_{\mathrm{DS}}(t)$四种状态节点和为1，时滞非负约束即$\tau>0$。

TD-FlipIt模型为非合作零和博弈，攻防双方的策略空间为紧空间，且其效用函数均为有界且上半连续，网络状态对策略扰动具有连续依赖性，根据Glicksberg定理[25]可知，此类非合作博弈存在混合策略纳什均衡，在此均衡策略中攻防双方无法单方面调整自己的策略以获得更优效用值，因此后续可通过多智能体强化学习方法逼近该均衡策略。在所构建的时滞FlipIt博弈中，攻防双方均为理性参与者，各自追求长期效用的最大化，故其优化目标均为最大化各自的效用函数U_{A}和U_{D}。

综上，TD-FlipIt模型融合时滞机制和状态感知的控制权，使得攻防双方通过策略博弈推动网络状态的演化，为云边协同网络中欺骗防御时机的优化提供了形式化的博弈基础。

4 欺骗防御策略求解

4.1 TD-FlipIt-MAPPO方法

TD-FlipIt博弈的混合策略纳什均衡无法通过传统静态规则或启发式方法有效求解，且攻防双方均为具备个体理性的智能体，攻防策略相互依赖且环境状态持续演化，且攻防双方的行动具有不确定性和动态性，在博弈过程中其策略不断迭代优化。为此，本节采用基于MAPPO的MARL方法，并与TD-FlipIt博弈模型深度融合以构建TD-FlipIt-MAPPO方法，实现攻防策略的协同演化与自适应优化。

云边协同网络中，TD-FlipIt-MAPPO方法中存在防御者与攻击者两个智能体，系统在当前时间步t的状态表示为$s_t\in S$，每个智能体的观测为$o_t\in O$，动作为$a_t\in A$，智能体通过执行动作更新状态并获得奖励，形成与环境的交互机制。每个智能体的目标是最大化期望累积奖励$\hat{R}$，如式（6）所示：

$$\hat{R}=\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t r(s_t,a_t)\right]\quad(11)$$

其中，r_t是智能体在状态s_t下采取动作a_t所获得的即时奖励，$\gamma\in[0,1)$为折扣因子。每个智能体根据其策略$\pi_i(o_i)$选择动作a_i，获得即时奖励r_i和下一状态s'，并得到经验组(s,a_i,r_i,s')存入经验回放缓冲区$\mathcal{D}$中。

如图4所示，欺骗防御轮换时机求解算法基于actor-critic架构，防御者和攻击者智能体之间存在动态的攻防交互，防御者actor网络根据其局部观测执行策略动作a_t^{def}，调整欺骗防御轮换时机，攻击者actor网络则同样执行其策略动作a_t^{att}以调整抢占时机参数，在每一轮攻防交互中，网络节点状态根据防御者和攻击者的动作a_t发生演化。而critic网络则使用全局状态s_t评估策略动作，从而对当前策略下每个智能体的价值函数V_ϕ进行准确评估。此外，通过引入广义优势估计（Generalized Advantage Estimation，GAE）和clip裁剪函数，可在策略更新过程中有效约束策略变化幅度，避免因剧烈更新导致训练不稳定。

攻击者和防御者智能体中的actor网络和critic网络训练的目标均为最小化其对应的损失函数，actor网络π_θ和critic网络V_ϕ的网络参数分别为θ和ϕ。智能体i的actor网络以局部观察o_t^i作为输入，输出动作

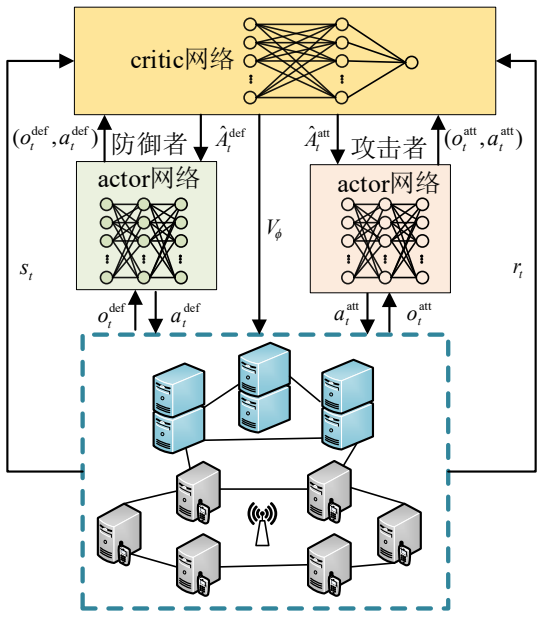


图4　欺骗防御策略求解框架

Figure 4　Deception defense strategy solving framework

概率分布,actor网络的损失函数为

$$L(\theta)=\frac{1}{Bn}\sum_{t=1}^{B}\sum_{i=1}^{n}\min\left(\lambda_{\theta,t}^{i}\hat{A}_t^i,\mathrm{clip}(\lambda_{\theta,t}^{i},1-\varepsilon,1+\varepsilon)\hat{A}_t^i\right)+\sigma\frac{1}{Bn}\sum_{t=1}^{B}\sum_{i=1}^{n}H\left(\pi_\theta(o_t^i)\right)\tag{12}$$

其中:n为智能体个数;B为训练的批量大小;$\hat{A}_t^i$为采用GAE方法计算得到的优势函数;$H(\pi_\theta(o_t^i))$是策略π_θ在观测o_t^i下的熵,用于衡量策略的不确定性;σ是熵系数,用于控制熵正则化的强度;$\lambda_{\theta,t}^{i}$是新策略与旧策略在相同观测和动作下的概率比值,如下式所示:

$$\lambda_{\theta,t}^{i}=\frac{\pi_\theta(a_t^i|o_t^i)}{\pi_{\theta_{\mathrm{old}}}(a_t^i|o_t^i)}\tag{13}$$

同时为确保训练的稳定性,防止策略更新过大,利用clip裁剪函数将策略概率比值约束在$(1-\varepsilon,1+\varepsilon)$范围内。

critic网络以全局状态s_t作为输入,输出状态值函数$V_\phi(s_t^i)$,其损失函数为

$$L(\phi)=\frac{1}{Bn}\sum_{t=1}^{B}\sum_{i=1}^{n}\max\left[\left(V_\phi(s_t^i)-\hat{R}_t\right)^2,\left(\mathrm{clip}\left(V_\phi(s_t^i),V_{\phi_{\mathrm{old}}}(s_t^i)-\varepsilon,V_{\phi_{\mathrm{old}}}(s_t^i)+\varepsilon\right)-\hat{R}_t\right)^2\right]\tag{14}$$

其中,$\hat{R}_t$是累积折扣奖励,$V_{\phi_{\mathrm{old}}}(s_t^i)$表示旧的critic网络对状态$s_t^i$的价值估计,同理此处也应用clip裁剪函数限制新旧价值函数的范围,以确保价值函数的更新不会过大。随后将损失函数反向传播到actor网络和critic网络以更新其各自对应的网络参数θ和ϕ,逐步优化欺骗防御与攻击响应策略。

4.2　算法及接口设计

基于TD-FlipIt-MAPPO的欺骗防御时机选择算法通过中心化训练、分布式执行实现,算法过程如算法1所示。所提TD-FlipIt-MAPPO方法基于actor-critic架构,其复杂度主要来源于神经网络的前向传播与反向传播,智能体数量为n,状态空间维度为$|S|$,动作空间维度为$|A|$,网络隐藏层维度为d,则在训练阶段的每次迭代中,actor网络和critic网络的前向传播与反向传播复杂度均为$O(B\cdot d^2)$,其中B为批量大小,整体训练复杂度与n呈线性关系,即$O(n\cdot B\cdot d^2)$;在推理阶段,防御者在单步决策时仅需actor网络的前向传播以生成动作,其复杂度为$O(d^2)$。

算法1　基于MAPPO的欺骗防御时机选择算法

输入:MAPPO的超参数,状态和奖励

输出:每个智能体的动作和参数

1. 创建云边协同网络中的防御者和攻击者智能体
2. 初始化actor网络和critic网络的参数θ,ϕ
3. 初始化经验回放缓冲区$\mathcal{D}$
4. **for** episode = 1, 2, …, N_e **do**
5. 　重置云边协同网络环境并获取状态s
6. 　**for** time step t = 1, 2, …, T **do**
7. 　　防御者智能体actor网络根据观测o_t^{def}选择并执行动作a_t^{def}
8. 　　攻击者智能体actor网络根据观测o_t^{att}选择并执行动作a_t^{att}
9. 　　获得奖励r_t和下一状态s'
10. 　　将经验组(s,a,r,s')存入缓冲区$\mathcal{D}$
11. 　**end for**
12. 　使用广义优势估计计算优势评估$\hat{A}$
13. 　计算累积折扣奖励$\hat{R}$
14. 　**for** mini-batch k = 1,2, …, K **do**
15. 　　从缓冲区$\mathcal{D}$中随机抽取小批量数据$\mathcal{B}$
16. 　　使用数据$\mathcal{B}$更新策略actor网络参数
17. 　　使用数据$\mathcal{B}$更新价值critic网络参数
18. 　**end for**
19. **end for**

然而,在时滞环境中,智能体在时刻t执行的动作a_t实际影响的是$t+\tau$时刻的系统状态。为此,在经验回放机制中对样本进行了延迟对齐处理,在计算优势函数时,将动作与延迟后的下一状态进行配对,确保价值函数能够正确反映动作的长期影响,并提升时滞环境下多智能体学习的稳定性。

通过与环境的持续交互学习各自最优策略。智能体对环境给出的每个状态进行价值评估,并据此选择动作;攻防博弈的状态信息更新后同样影响着攻击者和防御者的动作选择,经过多次迭代更新后,最终收敛至一个稳定的交互式策略。由此防御者可根据攻击者的行为模式动态调整欺骗轮换的触发时间,智

能体通过与环境交互过程中学习的“状态-动作”的价值映射，可适应多维度的欺骗策略编排，从而实现更高效的主动防御。

TD-FlipIt-MAPPO的状态、动作及奖励的MARL接口设计如下：状态空间包括当前网络中四类节点的状态比例$S_{\mathrm{NS}}(t)$、$S_{\mathrm{PS}}(t)$、$S_{\mathrm{IS}}(t)$和$S_{\mathrm{DS}}(t)$，攻防双方能观测到当前网络状态及自身执行翻转动作的时间；动作空间考虑攻防双方的策略，防御者动作a^{def}为调整欺骗防御轮换时机t，攻击者动作a^{att}同样为调整控制目标系统的时机，表示为参数λ的调节幅度；奖励函数则为博弈模型中的参与者效用函数，防御者奖励为防御者效用函数U_{D}，攻击者奖励为攻击者效用函数U_{A}。

5 实验与分析

5.1 实验设置

为评估所提基于时滞FlipIt博弈的欺骗防御时机选取方法TD-FlipIt-MAPPO的性能，实验选取的硬件配置环境为Intel（R）Xeon（R）Gold 5218 CPU与128 GB RAM，操作系统为Ubuntu 18.04，并基于Python 3.8.20和PyTorch 1.13.0深度学习框架进行实验环境的编程搭建。防御者与攻击者智能体的actor网络和critic网络均采用两层全连接网络结构，隐藏层维度均为128，并采用ReLU激活函数。actor网络输出层使用Softmax函数输出动作概率分布，critic网络输出状态价值函数值。此外，采用Mininet网络模拟器创建云边协同的攻击场景，通过Ryu控制器管理20个OpenFlow交换机，并部署500个终端主机，攻击者执行端口扫描以探测网络服务，防御方采用IP地址动态变换进行实例验证模拟欺骗资产的轮换，从而构建贴近实际的对抗场景。

参照文献[26]，实验设置单位时间内攻防双方控制节点的收益在0.5~2.5之间，且攻防博弈中网络状态转化时滞参数作为网络环境的物理特性，具有不确定性，并设置网络状态转移时滞和转移速率系数在相应区间，实验均在10个不同的随机种子下独立重复运行，并取平均值作为最终结果，具体参数设置如表1所示。

为了证明本文所提的TD-FlipIt-MAPPO方法在选取欺骗防御轮换时机方面的表现，对比方法选取如下：（1）FP，作为移动目标防御的经典静态触发机制，该方法以固定周期执行欺骗防御的轮换动作，不随攻防状态动态调整策略；（2）MFD-PPO[26]，该方法将攻防过程建模为多阶段FlipIt博弈，通过引入折扣因子和阶段转移概率来刻画系统的动态演变，该方法采用单智能体近端策略优化算法求解防御者策略，将攻击者视为被动环境因素而非独立决策主体；（3）SF-MAWP[24]，该方法构建Stackelberg-FlipIt博弈模型，采用多智能体WoLF-PHC算法求解均衡策略，考虑了攻防双方的信息不对称与序贯决策特性，但未显式建模状态转换时滞。

表1 仿真参数设置

Table 1 Simulation parameter settings

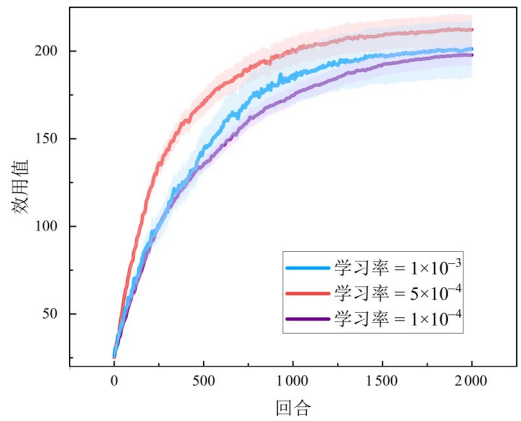
仿真参数	取值
单位时间内防御者控制节点收益	0.5~2.0
单位时间内攻击者控制节点收益	1.0~2.5
网络状态转移时滞τ	0.5~1.2
转移速率系数λ	0.1~1.0
攻击者单次控制目标资源的成本c_a	0.4
单次欺骗防御轮换的资源成本c_d	0.2
经验回放池$\mathcal{D}$	100 000
折扣因子γ	0.95
批量大小	128
回合	2 000
时间步	512
裁剪参数	0.2
熵系数	0.01

5.2 实验分析

5.2.1 收敛性分析

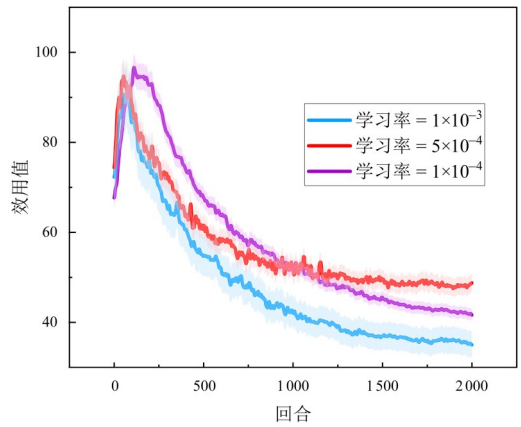
为评估TD-FlipIt-MAPPO算法中actor网络学习率对算法收敛性影响，本节设置学习率分别为1×10^{-4}、5×10^{-4}和1×10^{-3}进行训练。图5展示了在不同学习率下防御者智能体与攻击者智能体的效用函数随训练回合的变化曲线。

从图5（a）中可以看出，三种学习率均能使算法实现收敛，且防御者效用整体呈现上升趋势。其中，学习率为1×10^{-4}时，训练过程较为平缓，初期收敛速度较慢；学习率为1×10^{-3}时，虽然初期上升迅速，但在训练中期出现震荡，且最终收敛值明显低于学习率为5×10^{-4}时的收敛值；学习率为5×10^{-4}时，不仅具备较快的初始收敛速度，且在整个训练过程中保持良好的平稳性，最终达到最高的防御者效用水平。同时，图5（b）攻击者智能体的效用曲线也表现出相应趋势，在学习率过高时，攻击者策略响应波动后趋于较低水平；而在学习率为5×10^{-4}时，攻击者效用收敛到相对较高值。综上所述，学习率为5×10^{-4}在保证训练效率的同时兼顾了模型稳定性与收敛性能，能够有效促进攻防双方策略的协同演化。因此，本文最终选择5×10^{-4}作为TD-FlipIt-MAPPO算法的学习率，以实现最优的训练效果与策略鲁棒性。



(a) 防御者效用函数收敛曲线

(a) Convergence curve of the defender's utility function



(b) 攻击者效用函数收敛曲线

(b) Convergence curve of the attacker's utility function

图5　效用函数收敛曲线

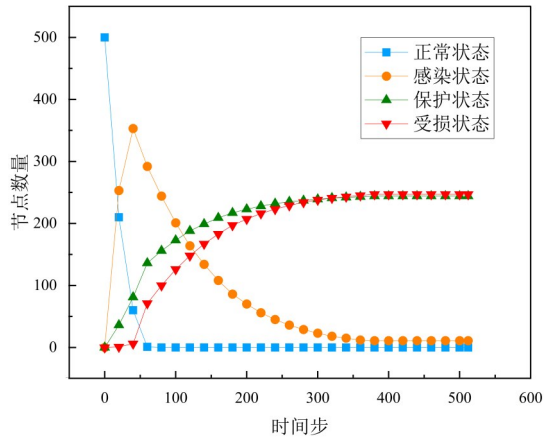
Figure 5　Convergence curves of utility functions

5.2.2　有效性分析

为验证状态转换时滞显式建模对欺骗防御时机决策的关键作用,本节设计消融实验,对比两种差异化配置的模型有效性:一是本文所提 TD-FlipIt-MAPPO 方法,显式融入状态转换时滞机制;二是 No-Delay MAPPO 方法,剔除时滞因素,假设所有节点状态转移瞬时完成,即 $\tau=0$,两种方法的节点状态变化曲线如图6所示。

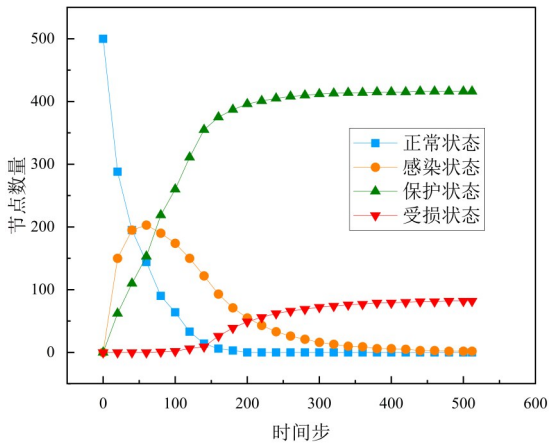
由图6(a)可见,No-Delay MAPPO 方法的节点状态变化呈现明显的瞬时响应特征,博弈初期正常状态节点数量急剧下降,感染状态节点占比快速上升,这一现象源于其忽略了真实网络中防御资源部署、攻击渗透等环节的固有物理延迟,导致防御策略过度激进。在攻防博弈后期,该方法难以维持稳定的保护状态节点比例,最终仅收敛至48.8%,防御效能受限。

对比而言,图6(b)中 TD-FlipIt-MAPPO 方法的节点状态演化更贴合实际场景,得益于状态转换时滞的显式建模,正常状态节点数量下降速率更为平缓,防御策略能够精准捕捉节点状态的动态演化规律,有效规避了无效的提前轮换动作,训练后期其保护状态节点比例稳定在83.2%,相较于 No-Delay MAPPO 方法提升了70.5%。



(a) 无时滞的节点状态变化曲线

(a) Curve of node state change without time delay



(b) 有时滞的节点状态变化曲线

(b) Curve of node state change with time delay

图6　网络节点状态变化曲线

Figure 6　Network node status change curve

综上,消融实验结果证明,忽略状态转换时滞会导致模型学习到偏误的次优策略,而显式建模该时滞因素可显著提升欺骗防御时机决策的合理性与有效性,为云边协同网络提供更可靠的主动防御支撑。

5.2.3　效用值分析

为评估所提方法在攻防博弈中的性能表现,本节通过比较不同方法在训练收敛后防御者与攻击者的最终效用值进行定量分析,结果如图7所示。

从防御者效用维度来看,FP 方法表现最差,其防

御者效用值仅为127.1。这一结果表明，静态固定周期的欺骗资源轮换机制缺乏对动态攻击行为的自适应能力，难以有效应对攻击者的策略调整，防御效能受限。与之形成鲜明对比的是，MFD-PPO、SF-MAWP和TD-FlipIt-MAPPO这三种基于强化学习的方法的防御效用均实现显著提升，验证了动态策略优化在欺骗防御时机决策中的核心优势。SF-MAWP与本文方法的防御效用略低于MFD-PPO方法，是因为MFD-PPO采用单智能体框架，未充分刻画攻防双方作为独立决策者的对抗行为，导致其在特定指标上表现偏优但模型真实性不足。

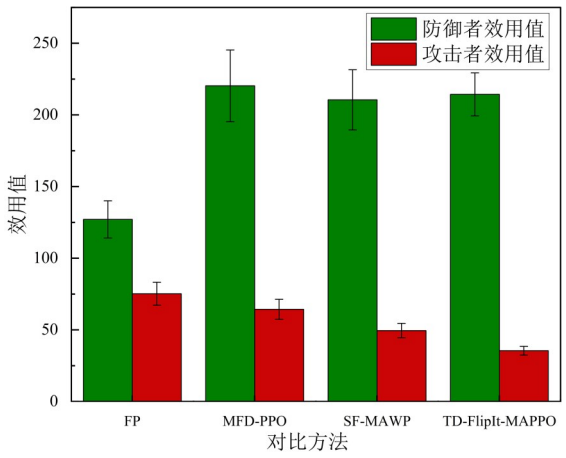


图7 不同方法的效用值对比

Figure 7 Comparison of the effectiveness values of different methods

从攻击者效用维度分析，本文所提方法得益于对状态转换时滞的显式建模，其攻击者效用值最低，为35.46，相较于MFD-PPO方法降低了44.85%，且防御者效用与MFD-PPO方法处于同一水平。

综上所述，尽管TD-FlipIt-MAPPO方法的防御者效用在绝对数值上略低于MFD-PPO方法，但该方法在抑制攻击者收益、贴合真实攻防场景等方面具备显著优势，综合性能与模型合理性更优，进一步验证了其作为欺骗防御轮换时机决策策略的有效性与可行性。

5.2.4 防御成本分析

为评估所提方法的防御资源开销控制能力，本节设计对比实验，在固定保护状态节点比例的约束下，以IP跳变频率为核心指标，量化不同方法的防御成本差异。实验中将保护状态节点定义为能规避攻击者扫描探测、未被纳入攻击列表的节点，将保护状态节点比例依次设置为0.1、0.2、0.3、0.4、0.5，该区间覆盖了普通数据传输（低安全需求）到关键业务承载（中等安全需求）的典型应用场景。为保障对比公平性，通过梯度调整各方法的防御策略参数，确保所有方法在训练过程中均能稳定达到预设的保护状态节点比例。IP跳变频率作为防御成本的直接表征，定义为单位时间内网络中IP地址发生变化的节点比例平均值，在相同保护状态节点比例下，IP跳变频率越低，则欺骗资源的轮换频率越低，资源消耗越少，即防御成本越低。

实验结果如表2所示，本文所提TD-FlipIt-MAPPO方法在所有预设的保护状态节点比例下，均实现了最低的IP跳变频率，表明该方法通过显式建模状态转换时滞，能够精准预判攻击生效窗口，避免了无意义的高频次防御轮换；同时，面对节点入侵时，防御者并非立即发起反击，而是结合时滞特性选择最优响应时机，进一步降低了不必要的资源消耗。

表2 不同方法的IP跳变频率

Table 2 The frequency of IP hopping for different methods

保护状态节点比例	FP	MFD-PPO	SF-MAWP	TD-FlipIt-MAPPO
0.1	0.033	0.025	0.028	0.023
0.2	0.075	0.048	0.052	0.039
0.3	0.098	0.072	0.087	0.068
0.4	0.135	0.096	0.104	0.095
0.5	0.182	0.120	0.161	0.121

综上，TD-FlipIt-MAPPO方法通过时滞感知的时机优化策略，在保障保护状态节点比例的前提下，有效降低了IP跳变频率，实现了安全性与资源开销的动态平衡，证实了其在防御成本控制方面的显著优势。

5.2.5 真实测试床验证实验

为进一步验证所提方法在接近真实网络环境中的可行性与实用性，本节开展小规模实测验证，重点对比TD-FlipIt-MAPPO与主流对比方法的核心性能差异。云层仍部署1台Intel（R）Xeon（R）Gold 5218 CPU，边缘层部署4台Raspberry Pi 4B作为边缘节点，接入20台虚拟终端，并配置端口扫描工具模拟攻击者从边缘层发起的多阶段渗透攻击，防御方通过动态IP配置工具模拟欺骗资源轮换，TD-FlipIt-MAPPO方法通过Python脚本部署于云层服务器，实时下发轮换指令，测试时长为1 h，每种方法独立运行5次取平均值。如表3结果所示，TD-FlipIt-MAPPO方法的攻击拦截成功率达到87.9%，较FP方法提升38.4%，较MFD-PPO方法提升14.3%，验证了时滞建模与多智能体交互机制在真实环境中的有效性，能够更精准地捕捉攻击节奏并阻断攻击渗透。

表3　真实测试床实验结果对比

Table 3　Comparison of actual test bed experimental results

方法	攻击拦截成功率/%
FP	63.5
MFD-PPO	76.9
SF-MAWP	80.1
TD-FlipIt-MAPPO	87.9

6　结束语

本文针对云边协同网络中欺骗防御轮换时机忽略状态转换时滞及攻防双方动态交互的问题,提出了一种基于时滞FlipIt博弈和多智能体强化学习的欺骗防御时机选取方法。通过分析网络安全演化特性的基础上,显式建模了节点状态转换过程中的物理时滞,构建了融合时滞微分方程的网络状态演化模型;随后给出了网络攻防模型,并在此基础上建立了时滞FlipIt博弈模型,最终采用多智能体近端策略优化算法求解最优欺骗防御轮换时机策略。实验结果表明,所提方法在保护状态节点比例和防御成本控制等关键指标上均优于FP、MFD-PPO、SF-MAWP等基线方法,有效平衡了安全性与资源开销,为云边协同环境下的主动、高效欺骗防御时机的轮换提供了可行的技术路径。然而现有攻防模型和时滞设定在一定程度上简化了实际攻防环境的复杂性,以便于分析时滞对防御时机决策的直接影响,且实例验证主要聚焦于网络IP跳变,尚未涵盖应用层蜜罐或数据层诱饵的联合编排,未来将通过构建多维异质欺骗资源池及协同编排跨层欺骗防御手段,并引入细粒度的攻防成本参数,针对自适应学习能力更强的攻击者模型和更灵活的时滞分布开展进一步研究,以验证所提方法在强对抗环境下的有效性与鲁棒性。

参考文献

[1] Wang Yingchao, Yang Chen, Lan Shulin, et al. End-edge-cloud collaborative computing for deep learning: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2024, 26(4): 2647-2683.

[2] Chen Xiang, Guo Zhiheng, Wang Xijun, et al. Toward 6G native-AI network: Foundation model-based cloud-edge-end collaboration framework[J]. IEEE Communications Magazine, 2025, 63(8): 23-30.

[3] Jamil M N, Schelén O, Monrat A A, et al. Enabling industrial internet of things by leveraging distributed edge-to-cloud computing: Challenges and opportunities[J]. IEEE Access, 2024, 12: 127308.

[4] 马博, 余应洁, 吴莎尘, 等. 云边端异构算力网络计算任务分割与路径优化方法研究[J]. 电子学报, 2025, 53(6): 1847-1864.
Ma Bo, Yu Yingjie, Wu Shachen, et al. Task segmentation and path optimization in heterogeneous cloud-edge-end computing power network[J]. Acta Electronica Sinica, 2025, 53(6): 1847-1864. (in Chinese)

[5] Stephanie V, Khalil I, Rahman M S, et al. Privacy-preserving ensemble infused enhanced deep neural network framework for edge cloud convergence[J]. IEEE Internet of Things Journal, 2023, 10(5): 3763-3773.

[6] Javadpour A, Ja'fari F, Taleb T, et al. A comprehensive survey on cyber deception techniques to improve honeypot performance[J]. Computers & Security, 2024, 140: 103792.

[7] Gill K S, Sharma A, Saxena S. A systematic review on game-theoretic models and different types of security requirements in cloud environment: Challenges and opportunities[J]. Archives of Computational Methods in Engineering, 2024, 31(7): 3857-3890.

[8] 蒋侣, 张恒巍, 王晋东. 基于多阶段Markov信号博弈的移动目标防御最优决策方法[J]. 电子学报, 2021, 49(3): 527-535.
JIANG Lü, ZHANG Hengwei, WANG Jindong. A Markov signaling game-theoretic approach to moving target defense strategy selection[J]. Acta Electronica Sinica, 2021, 49(3): 527-535. (in Chinese)

[9] Cai Feiyang, Koutsoukos X. Real-time detection of deception attacks in cyber-physical systems[J]. International Journal of Information Security, 2023, 22(5): 1099-1114.

[10] Chowdhury N H, Adam M T P, Teubner T. Time pressure in human cybersecurity behavior: Theoretical framework and countermeasures[J]. Computers & Security, 2020, 97: 101963.

[11] Zhang Yunxiao, Malacaria P. Optimization-time analysis for cybersecurity[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(4): 2365-2383.

[12] Merlevede J, Johnson B, Grossklags J, et al. Time-dependent strategies in games of timing[C]//Proceedings of the 10th International Conference on Decision and Game Theory for Security. Heidelberg: Springer, 2019: 310-330.

[13] Zhang Hengwei, Tan Jinglei, Liu Xiaohu, et al. Moving target defense decision-making method: A dynamic Markov differential game model[C]//Proceedings of the 7th ACM Workshop on Moving Target Defense. New York: ACM, 2020: 21-29.

[14] Farhang S, Grossklags J. When to invest in security? Empirical evidence and a game-theoretic approach for time-based security[PP/OL]. V2. arXiv (2017-06-06)[2026-03-10]. https://arxiv.org/abs/1706.00302.

[15] Van Dijk M, Juels A, Oprea A, et al. FLIPIT: The game of "stealthy takeover"[J]. Journal of Cryptology, 2013, 26(4):

655-713.
[16] Chen Xiaoguang, Cao Wenyuan, Chen Lili, et al. iCyber-Guard: A FlipIt game for enhanced cybersecurity in IIoT[J]. IEEE Transactions on Computational Social Systems, 2024, 11(6): 8005-8014.
[17] 何威振, 谭晶磊, 张帅, 等. 利用深度强化学习的多阶段博弈网络拓扑欺骗防御方法[J]. 电子与信息学报, 2024, 46(12): 4422-4431.
He Weizhen, Tan Jinglei, Zhang Shuai, et al. Multi-stage game-based topology deception method using deep reinforcement learning[J]. Journal of Electronics & Information Technology, 2024, 46(12): 4422-4431. (in Chinese)
[18] Tan Jinglei, Zhang Hengwei, Zhang Hongqi, et al. Optimal timing selection approach to moving target defense: A FlipIt attack-defense game model[J]. Security and Communication Networks, 2020, 2020(1): 3151495.
[19] Zhu Zinan, Zhou Lin. Application of complex network attack and defense time game model in network security defense decision[J]. Journal of Cyber Security and Mobility, 2025, 14(2): 311-338.
[20] 熊鑫立, 黄郡, 姚倩. 基于Cheat-FlipIt博弈的网络安全对抗建模与分析[J]. 信息对抗技术, 2024, 3(4): 63-80.
Xiong Xinli, Huang Jun, Yao Qian. Modeling and analysis of network security adversarial behavior based on Cheat-FlipIt game[J]. Information Countermeasure Technology, 2024, 3(4): 63-80. (in Chinese)
[21] Zheng Yan, Na Zhenyu, Ji Weidong, et al. An adaptive fuzzy SIR model for real-time malware spread prediction in industrial internet of things networks[J]. IEEE Internet of Things Journal, 2025, 12(13): 22875-22888.
[22] Qi Ju. Loss and premium calculation of network nodes under the spread of SIS virus[J]. Journal of Intelligent & Fuzzy Systems, 2023, 44(5): 7919-7933.
[23] Qiu Li, Xiang Caixia, Wen Yidan, et al. Predictive output feedback control of networked control system with Markov DoS attack and time delay[J]. International Journal of Robust and Nonlinear Control, 2023, 33(5): 3376-3395.
[24] Sun Rongbo, Fei Jinlong, Zhu Yuefei, et al. Multi-agent reinforcement learning for moving target defense temporal decision-making approach based on Stackelberg-FlipIt games[J]. Computers, Materials & Continua, 2025, 84(2): 3765-3786.
[25] He Weizhen, Tan Jinglei, Guo Yunfei, et al. Flipit game deception strategy selection method based on deep reinforcement learning[J]. International Journal of Intelligent Systems, 2023, 2023(1): 5560416.
[26] Glicksberg I L. A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points[J]. Proceedings of the American Mathematical Society, 1952, 3(1): 170-174.

作者简介

裴金川 男,1998年6月出生于河北省唐山市。现为信息工程大学博士研究生。主要研究方向为网络主动防御与新型网络体系结构。
E-mail: jinchuan_pei@163.com

田 乐 男,1985年出生于陕西省咸阳市。现为信息工程大学副研究员,硕士生导师。主要研究方向为新型网络体系结构。
E-mail: le.tian2019@outlook.com

胡宇翔 男,1982年出生于河南省周口市。现为信息工程大学教授,博士生导师。主要研究方向为网络空间安全与新型网络体系结构。
E-mail: huyuxiangchn@163.com

李梦龙 男,1992年出生于河南省新乡市。现为信息工程大学助理工程师。主要研究方向为云网融合与强化学习。
E-mail: growinglml@163.com

于洪涛 男,1970年出生于辽宁省丹东市。现为信息工程大学研究员,博士生导师。主要研究方向为大数据与人工智能。
E-mail: hgnn2kan@163.com