基于审计博弈的安全协作频谱感知方案

王云涛[1]，苏洲[1]，许其超[2]，刘怡良[1]，彭海霞[1]，栾浩[1]
（1. 西安交通大学网络空间安全学院，陕西 西安 710049；2. 上海大学机电工程与自动化学院，上海 200444）

摘 要：针对群智协作频谱感知中恶意感知终端的投毒与搭便车攻击，结合事前威慑与事后惩罚机制提出了一种基于审计博弈的新型防御方案。首先，考虑审计预算约束，构建了一种不完全信息下的混合策略审计博弈模型，在协作感知前设置惩罚策略威慑恶意协作者，并在感知数据融合后进行审计进而实施惩罚。其次，设计了链上链下协同的轻量审计区块链模型，其中，审计证据存储在链下数据仓库，其元数据公开发布在审计链上。再次，设计了基于强化学习的分布式智能审计算法，以在动态环境下自适应地计算审计博弈的渐近混合策略均衡。仿真结果表明，相比传统方案，所提方案能快速获取稳定且渐近最优的审计策略，并积极抑制恶意协作者的投毒与搭便车行为。

关键词：协作频谱感知；审计博弈；安全；区块链；强化学习
中图分类号：TN92
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023238

Secure and collaborative spectrum sensing scheme based on audit game

WANG Yuntao[1], SU Zhou[1], XU Qichao[2], LIU Yiliang[1], PENG Haixia[1], LUAN Hao[1]
1. School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China
2. School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China

Abstract: To defend against poisoning attacks and free-riding attacks conducted by malicious sensing terminals in crowd sensing-based collaborative spectrum sensing (CCSS), a novel audit game-based defense scheme was proposed, which combined the pre-deterrence and post-punishment mechanisms. Firstly, considering the audit budget constraint, a mixed-strategy audit game model under incomplete information was designed, which set a penalty strategy to deter malicious collaborators before spectrum sensing, and audited and punished them after the fusion of sensing data. Then, a lightweight audit chain model with on-chain and off-chain collaboration was designed, in which audit evidence was stored in an off-chain data warehouse and its metadata was publicly published on the blockchain. Furthermore, a distributed and intelligent audit algorithm based on reinforcement learning was devised to adaptively seek the approximate mix-strategy equilibrium of the audit game. Simulation results demonstrate that the proposed scheme can quickly obtain the stable and approximately optimal audit strategies and actively suppress the poisoning and free-riding behaviors of malicious collaborators, in comparison with conventional schemes.

Keywords: collaborative spectrum sensing, audit game, security, blockchain, reinforcement learning

收稿日期：2023−07−04；**修回日期：**2023−09−27
通信作者：苏洲，zhousu@xjtu.edu.cn
基金项目：国家自然科学基金资助项目（No.62302387, No.62273223, No.U23A20276, No.62101429）；博士后创新人才支持计划基金资助项目（No.BX20230282）
Foundation Items: The National Natural Science Foundation of China (No.62302387, No.62273223, No.U23A20276, No.62101429), Postdoctoral Innovative Talent Support Program of China (No.BX20230282)

0 引言

随着智能物联网（IoT, Internet of things）设备及其数据规模的爆炸式增长，无线通信中的数据流量急剧攀升。根据Gartner报告，预计2025年全球物联网连接规模将超过270亿并将持续产生超过2 ZB的数据流量[1]。面向物联网应用规模的快速扩张以及自动驾驶、智慧医疗等各类新型物联网应用的蓬勃发展[2]，如何高效利用有限且稀缺的频谱资源来支撑高带宽、高动态、广连接、低时延的多样化物联网服务成为空天地一体化网络和6G系统面临的重要挑战[3-4]。

频谱感知[5]作为频谱共享系统的关键技术，被认为是缓解无线网络频谱资源匮乏问题的有效途径。其中，非授权的次用户（SU, secondary user）通过周期性频谱监测来准确感知并机会性地接入授权主用户（PU, primary user）空闲的频段，从而最大限度地提升有限频谱的利用率与系统容量。群智协作频谱感知[6-7]通过融合群智感知与协作频谱感知技术，提供了一种低成本、高弹性、可按需的高效频谱感知方案，受到了学术界的广泛关注。首先，群智协作频谱感知通过高效融合多个协作用户的频谱感知信息，可以有效克服无线网络的多径衰落、信号遮挡和阴影效应导致的深衰落和隐终端现象，从而提供更精确及时的PU状态监测；其次，群智协作频谱感知利用分布广泛、低成本、低功耗、高移动性的智能IoT设备作为候选频谱感知协作者，以保证频谱感知节点的数量充足，从而提高协作频谱感知性能并降低服务成本；再次，群智协作频谱感知可通过动态调整感知协作设备的数量、部署位置、感知数据精度以及感知周期等，以动态适应时变的通信链路与移动网络状态，从而保证较高的部署灵活性。因此，群智协作频谱感知技术研究具有重要的理论意义和应用前景。

在群智协作频谱感知中，由于网络开放化和攻击行为多样化，部分IoT设备作为分布式频谱感知协作者，可能发送虚假甚至精心构造的本地频谱感知数据，即频谱感知数据篡改（SSDF, spectrum sensing data falsification）[8]或投毒攻击[9-10]，从而影响甚至操纵融合中心（FC, fusion center）最终的频谱感知结果。例如，恶意感知终端可能通过发送专门构造的虚假频谱感知数据使最终频谱感知结果为占用，从而危害系统的正常工作。同时，恶意感知设备可能发动搭便车攻击[11]，通过贡献低质量甚至冗余的频谱感知数据以减小感知成本，在不做出贡献的情况下享受协作频谱感知服务。此外，无线网络多径衰落和阴影效应等因素以及5G小基站部署的超密集化与异构化导致的频谱分布多维化与高动态等特征，加大了恶意用户监测的难度。因此，如何防御群智协作频谱感知中恶意感知协作者的投毒与搭便车攻击，对于确保协作频谱检测结果的可靠性具有重要意义。

近年来，国内外学者们在抗SSDF攻击的安全协作频谱感知方面开展了大量的相关研究，目前的防御方案主要分为基于可信评估的防御[6,8,12-15]、基于鲁棒融合策略的防御[16-19]、基于博弈论的防御[20-23]三类。另外，当前也有部分文献基于密码学[24]以及面向用户隐私保护[25]进行研究。

基于可信评估的防御[6,8,12-15]一般通过挖掘历史交互信息进行动态信任或声誉评估，从而甄别并筛选出可信用户或终端。针对众包模式下协作频谱感知中的恶意数据注入攻击，联合考虑移动感知设备的即时可信度及其在数据融合过程中的信誉评分，文献[6]提出了基于即时信任评估的高效安全的协作频谱感知策略。针对协作频谱感知系统中的恶意SU，文献[14]提出了两级防御方案来抵御共谋伪造反馈攻击，在请求级防御中引入反馈信任来检测伪装成正常SU的共谋攻击者，在反馈级防御中分析历史感知数据和反馈数据进行频率相关性分析以防止攻击者提升信任值。

基于鲁棒融合策略的防御[16-19]一般通过基于规则、特征或者学习的方法来动态调整用户频谱感知数据的权重，或者通过聚类方法找出并剔除异常值，从而设计鲁棒的数据融合策略。文献[16]研究了存在错误频谱测量数据下的安全众包无线电环境图构造，联合考虑移动用户测量值的时空可信度并仅用最可信的测量值构建无线电环境图。文献[19]提出了基于图神经网络的空间插值方法来提升协作频谱感知安全性，在频谱传感器可信度评估中融合考虑频谱传感器的异质性，在恶意频谱传感器占多数的情况下仍可靠地检测频谱占用状态。

综上，基于可信评估的防御方案一般依赖于历史交互经验，存在冷启动等问题，难以适用于车联网和无人机网络等高动态、低交互的移动物联网中。基于鲁棒融合策略的防御方案通常需要更大的

时间开销，难以满足自动驾驶等时变动态网络下的时延敏感性需求。此外，上述方案中攻击者与防御者的成本、收益、预算约束等往往被忽略，在成本管理下，何时、何种方式的攻击防御能满足给定攻防预算下的安全协作频谱感知需求成为挑战。

基于博弈论的防御方案[20-23]通过运用博弈理论分析攻击者/防御者的成本最小化或收益最大化的均衡策略，可实现成本高效的安全协作频谱感知。文献[21]提出了一种基于联盟−讨价还价的分层博弈模型，通过度量用户的贡献值并求取分层博弈的纳什均衡策略来抵御搭便车攻击。文献[22]设计了基于信号博弈的搭便车行为防御机制，通过恶意 SU 与正常 SU 间的多轮次信号博弈，来获取对用户诚实度的稳定信念评判，以甄别恶意协作者。然而，当前博弈论方案主要通过激励用户的协同性来抑制其恶意行为，很少考虑对恶意行为的追溯与惩罚，很少考虑用户与融合中心的双向信息不对称性以及参与者的混合策略决策，导致防御效果有限。

基于审计博弈的防御方案结合事前威慑与事后惩罚机制，为协作频谱感知提供了一种新的防御思路。一方面，在协作感知开始前设置惩罚策略对恶意用户进行威慑，并在感知数据融合后对参与的协作者进行审计进而实施惩罚；另一方面，它可与现有防御策略高效融合，从而提升整体防御效率。本文研究利用审计博弈机制来抵御群智协作频谱感知中的投毒及搭便车攻击。然而，基于审计博弈的群智协作频谱感知防御服务的实际部署仍面临以下主要挑战。1) 审计规模与效率：融合中心（FC, fusion center）通常具有有限的审计资源，难以对规模庞大的协作群体及其频谱感知数据进行全面审计，亟须高效权衡审计规模和审计效率。2) 隐私信息造成信息不对称：感知协作者通常具有差异化的部署位置和感知精度等隐私信息，造成审计双方的信息不对称，如何设计信息不对称下的最优审计策略成为挑战。3) 混合策略及高度动态环境：审计者（即 FC）和被审计者（即感知协作者）通常可采用混合策略而非纯策略以提升自身效用，同时网络状态与通信链路的动态时变使双方最优决策愈加困难。

本文主要的研究工作及贡献如下。

1) 提出了一种不完全信息下的动态重复序贯审计博弈模型来高效审计恶意用户，其中，FC 作为审计者基于审计预算选择一组最可疑的频谱感知终端进行审计，并决定其混合惩罚策略，然后感知终端作为被审计者根据惩罚策略序贯地决定其混合投毒策略，博弈过程重复进行直至达成 Stackelberg 均衡。设计了链上链下协同的轻量级审计区块链模型，其中审计证据数据存储在链下数据仓库，同时其元数据信息公开发布在审计链上。

2) 理论计算了完全信息下的纯策略 Stackelberg 均衡策略。针对实际环境中环境高度动态性、参与者隐私参数和混合策略的存在导致审计双方的信息不对称，提出了基于强化学习的分布式智能审计算法，以在动态环境下自适应地计算博弈的近似均衡策略，其中审计者和被审计者分别采用策略爬山（PHC, policy hill climbing）学习算法以序贯的方式动态决策并更新其最优策略。

3) 仿真实验验证了所提方案的可行性和有效性。仿真结果表明，与传统方案相比，所提方案能快速获取稳定且近似最优的审计策略，积极抑制恶意协作者的投毒与搭便车行为。

1　系统模型

1.1　网络模型

群智协作频谱感知网络模型如图 1 所示。本文考虑蜂窝网基站或电视塔等大范围且固定式 PU 下的群智协作频谱感知场景[6-7]，主要包括一个 FC、一个 PU、M 个随机分布的 SU、K 个频谱感知协作者以及审计链。频谱感知协作者主要由手机、平板电脑等各类低成本、低功耗的智能 IoT 设备组成，通过内置的频谱传感器参与频谱感知[6]，以保证足够数量的频谱感知节点参与，从而提升协作频谱感知性能。频谱感知终端包括两部分：M 个 SU 和 K 个感知协作者，分别用集合 $\mathcal{M}=\{1,\cdots,M\}$ 和 $\mathcal{K}=\{1,\cdots,K\}$ 表示。令 $N=M+K$，频谱感知终端的集合可表示为 $\mathcal{N}=\{1,\cdots,N\}$。每个长度为 T 的时间帧分为频谱感知和数据传输两部分[10]。N 个频谱感知终端在感知阶段独立地感知 PU 的活跃状态，并在传输阶段发送频谱感知报告至 FC 进行感知数据的融合计算。当全局检测结果为 PU 信号不存在时，SU 即可机会式地占用 PU 的授权频段进行信息传输。

FC 周期性地审计用户的频谱感知报告。为提升审计效率，FC 与频谱感知终端之间进行审计博弈，其中，频谱感知终端可发送真实/伪造的感知报告，审计者需要在审计规模与审计效率之间权衡，并根据被审计者的作恶程度设置相应惩罚策略，最后将审计结果上传至审计链。

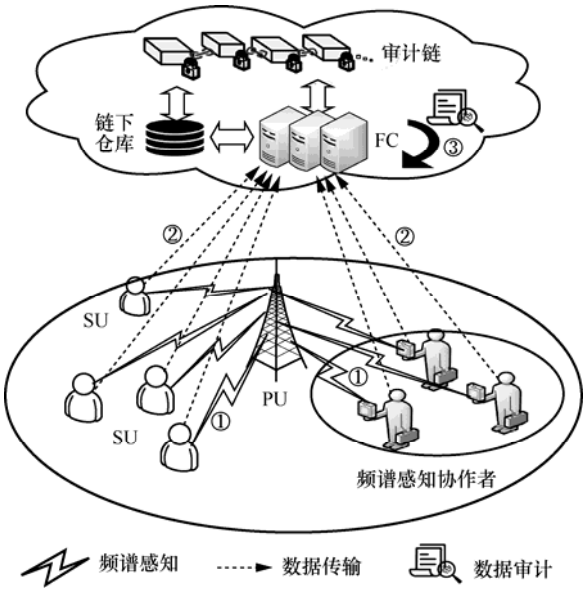


图 1 群智协作频谱感知网络模型

1.2 频谱感知与融合模型

根据信号传播模型[26]，对于频谱感知终端 $n\in\mathcal{N}$，其接收的 PU 信号强度可表示为

$$P_n=\left(\frac{d_0}{d_n}\right)^{\iota}P_0e^{A_n}e^{B_n} \tag{1}$$

其中，d_0 表示基准距离，d_n 表示频谱感知终端 n 与 PU 间的距离，P_0 表示在 d_0 处接收到的 PU 的信号强度，ι 表示路径损耗指数，e^{A_n} 和 e^{B_n} 分别表示阴影衰落和多径衰落的影响，并有 $A_n\sim\mathbb{N}(0,\sigma_n^2)$，$\sigma_n$ 表示背景噪声的标准差。假设信道带宽远大于相干带宽，则多径衰落的影响可以忽略不计[6]，即 $B_n=0,\forall n\in\mathcal{N}$。此外，假设 A_n 和 A_m 对于所有 $n\neq m$ 都是独立的，即所有频谱感知终端经历独立同分布（IID, independent and identically distributed）的高斯阴影和衰落（相同的平均信噪比 ψ_0）[27]。

由于能量检测法[6,10]简单易用、计算复杂度低且不需要 PU 信号的相关信息，频谱感知终端采用能量检测法进行协作频谱感知。在感知阶段，每个频谱感知终端 n 在本地收集 h 个接收信号强度（RSS, received signal strength）样本，即 $\boldsymbol{s}_n=(s_{n,1},\cdots,s_{n,h})$，并在传输阶段将统计量（即平均 RSS）$S_n=\frac{1}{h}\sum_{z=1}^{h}s_{n,z}$ 作为频谱感知报告发送至 FC。FC 采用等增益合并（EGC, equal gain combination）规则[28]对所有协作终端的感知数据进行软融合来最终决策，其具有易于部署、较少依赖 PU 相关先验知识等优势。具体地，FC 融合各感知终端发送的频谱感知报告，计算切尾均值（TM, trimmed mean）得到 S_{FC} 并判定 PU 状态，即

$$S_{\mathrm{FC}}=\mathrm{TM}\left(S_n\right)\begin{cases}<\lambda_{\mathrm{FC}}, & \mathcal{H}_0\\ \geqslant\lambda_{\mathrm{FC}}, & \mathcal{H}_1\end{cases} \tag{2}$$

其中，TM(·) 表示去除最大和最小值后的算术均值，λ_{FC} 表示 FC 预设的门限值，$\mathcal{H}_0$ 和 $\mathcal{H}_1$ 分别表示 PU 处于空闲与活跃（即信道占用）状态。由于 $\{S_n,n\in\mathcal{N}\}$ 为相互独立的高斯变量，S_{FC} 近似满足

$$S_{\mathrm{FC}}\sim\begin{cases}\mathbb{N}\left(N\sigma_n^2,\dfrac{N\sigma_n^4}{h}\right),\mathcal{H}_0\\ \mathbb{N}\left(\left(N+\mu_{\Sigma}\right)\sigma_n^2,\dfrac{\left(N+2\mu_{\Sigma}\right)\sigma_n^4}{h}\right),\mathcal{H}_1\end{cases} \tag{3}$$

其中，$\mu_{\Sigma}=\sum_{n=1}^{N}|h_n|^2\psi_0$ 表示联合信噪比，h_n 表示感知信道衰落因子。因此，协作频谱感知下的全局检测概率和全局虚警概率分别为

$$P_{\mathrm{d}}=\Pr\left\{S_{\mathrm{FC}}\geqslant\lambda_{\mathrm{FC}}\mid\mathcal{H}_1\right\}=\mathcal{F}\left(\frac{\lambda_{\mathrm{FC}}-\left(N+\mu_{\Sigma}\right)\sigma_n^2}{\sigma_n^2\sqrt{\dfrac{\left(N+2\mu_{\Sigma}\right)}{h}}}\right) \tag{4}$$

$$P_{\mathrm{f}}=\Pr\left\{S_{\mathrm{FC}}\geqslant\lambda_{\mathrm{FC}}\mid\mathcal{H}_0\right\}=\mathcal{F}\left(\frac{\lambda_{\mathrm{FC}}-N\sigma_n^2}{\sigma_n^2\sqrt{\dfrac{N}{h}}}\right) \tag{5}$$

其中，函数 $\mathcal{F}(x)=\frac{1}{\sqrt{2\pi}}\int_x^{+\infty}\mathrm{e}^{\frac{-z^2}{2}}\mathrm{d}z$。全局漏检概率为 $P_{\mathrm{m}}=1-P_{\mathrm{d}}$。若检测结果为 PU 空闲，当仅有单个 SU 时，其接入全部空闲频谱；当存在多个 SU 时，同时隙内多个传输者的数据传输势必造成同频干扰，导致全局吞吐量的下降。多 SU 下空闲信道的优化利用可参考文献[29]。

1.3 威胁模型

假设 FC 是诚实的，即审计者诚实地聚合并计算被审计者（即 SU 和协作频谱感知者）上传的频谱感知结果。假设存在部分协作频谱感知节点是恶意的，即实施 SSDF 攻击来操纵最终的协作频谱感知结果。诚实的协作频谱感知节点将发送真实的感知

报告。在群智协作频谱感知系统中，考虑如下两类具体的 SSDF 攻击形式。

1) 当 PU 信号不存在时，恶意的协作频谱感知终端可故意报告高 RSS 值，从而增加虚警概率 P_{f} 并防止 SU 用户使用该信道。

2) 恶意的协作频谱感知终端还可以在 PU 存在时故意报告低的 RSS 值，从而增加漏检概率 P_{m} 并增加对 PU 的信道干扰。

1.4 审计区块链模型

本文设计的审计链为一种许可式区块链。一方面，审计链可抵御外部攻击者的数据篡改攻击，确保上链数据的不可篡改性和不可抵赖性；另一方面，审计链可为参与者提供一个公开透明的账本，确保恶意终端 SSDF 行为的可追溯性和 FC 惩罚操作的透明性[30]。针对节点性能差异性，考虑区块链网络中存在 3 种节点：全节点、轻节点和验证者节点。全节点存储全部区块数据并提供区块链账本服务，轻节点只存储区块元数据并仅参与事务的发布与转发[31]，验证者节点运行授权证明（PoA, proof of authority）共识协议[32]集体维护区块链数据和状态。审计链中包含一串不断增长的哈希相连的区块，每个高度为 i 的区块 $\mathcal{B}_i$ 包含区块头 bHeader 和区块体 bBody 两部分，可表示为

$$\mathcal{B}_i=\{\underbrace{\{\mathrm{pHash},\mathrm{cHash},\mathrm{mRoot},i,\mathrm{ts},\mathrm{Sig}\}}_{\mathrm{bHeader}}\|\,\mathrm{bBody}\} \quad (6)$$

其中，pHash 和 cHash 分别表示父区块和当前区块的哈希值，mRoot 表示区块体中包含的所有事务的默克尔树的根值，ts 表示生成时间戳，Sig 表示区块生产者的签名。

为提高审计效率，审计链中设计了以下 3 种新型事务。

1) 本地频谱感知事务（lsTx, local spectrum sensing transaction）。lsTx 存储了单个频谱感知终端在一个频谱感知阶段内上传的 RSS 报告，可表示为

$$\mathrm{lsTx}=\langle t_{\mathrm{ID}}\|S_n\|\mathrm{ts}\|H(\mathrm{lsTx})\|\mathrm{Sig}_n\rangle \quad (7)$$

其中，t_{ID} 表示协作频谱感知任务 ID，$H(\cdot)$ 表示安全哈希函数，Sig_n 表示感知终端 n 对 lsTx 哈希值的签名。参与同一个频谱感知任务的所有频谱感知终端的 lsTx 可通过聚合签名[33]存储在审计链上以减轻开销。

2) 全局融合事务（gfTx, global fusion transaction）。gfTx 记录了 FC 在一次频谱感知任务内融合计算的 RSS 结果，可表示为

$$\mathrm{gfTx}=\langle t_{\mathrm{ID}}\|S_{\mathrm{FC}}\|\mathrm{ts}\|H(\mathrm{gfTx})\|\mathrm{Sig}_{\mathrm{FC}}\rangle \quad (8)$$

其中，$\mathrm{Sig}_{\mathrm{FC}}$ 表示 FC 对 gfTx 哈希值的签名。

3) 审计事务（auTx, audit transaction）。auTx 记录了 FC 每次审计与惩罚的结果，可表示为

$$\mathrm{auTx}=\langle t_{\mathrm{ID}}\|(\mathcal{W},\boldsymbol{x})\|H(\mathrm{Evi})\|\mathrm{ts}\|H(\mathrm{auTx})\|\mathrm{Sig}_{\mathrm{FC}}\rangle \quad (9)$$

其中，$(\mathcal{W},\boldsymbol{x})$ 表示 FC 的审计规模与结果，$\mathcal{W}\subseteq\mathcal{N}$ 表示被审计者集合，$\boldsymbol{x}=[x_1,\cdots,x_W]$ 表示相应惩罚向量，Evi 表示存储在链下仓库的审计证据数据，其哈希指针 $H(\mathrm{Evi})$ 公开发布在审计链上。

2 审计博弈框架

所提协作式频谱感知系统的审计框架包含下面 2 个阶段。

1) 威慑阶段。在频谱感知前，审计者设置最优审计策略对恶意频谱感知终端进行威慑，从而抑制恶意频谱感知终端的投毒行为。

2) 惩罚阶段。在频谱感知后，审计者根据审计预算选择部分频谱感知终端进行审计，通过从审计链上获取相关证据数据，并对恶意频谱感知终端实施惩罚策略，同时把审计结果记录在审计链上。

2.1 混合策略的重复序贯审计博弈表述

为保证协作频谱感知结果的可靠性，需要验证频谱感知终端上传的本地频谱感知报告的真实性。其中，FC 为审计者，频谱感知终端为被审计者。频谱感知终端发送的 lsTx 将打包进区块链中，使审计者从中提取相关证据信息进行审计。FC 可依据文献[8]中的信任评估机制，利用 PU 与恶意终端用户的接收者操作特征（ROC, receiver operating characteristic）曲线计算虚警率和假阳率，从而检测出恶意频谱感知终端。令 W 为审计预算，即审计者最多只能同时审计 W 个频谱感知终端。在审计预算约束下，审计者决定其最优混合审计策略，以最大限度地抑制节点的作恶行为。同时，考虑到审计的威慑力与惩罚，被审计者将不断调整其混合攻击策略以迷惑审计者并最大化其收益。由于频谱感知模型中用户隐私信息的存在以及信道参数的时变性，导致双方存在高度的信息不对称性，审计双方可通过重复交互来降低信息不对称性并获取最优策略。将信息不对称和审计预算约束下审计者和多个被审计者之间的竞争性交互建模为一个领导者和多个跟

随者的重复审计博弈。

定义 1 审计博弈。审计者和被审计者之间的交互可表述为一个混合策略重复序贯审计博弈 $\mathcal{G}=\{\mathrm{FC},\mathcal{N},\{\{x_n,y_n\}_{n\in\mathcal{N}}\},\{\boldsymbol{a}_n,\boldsymbol{b}_n\}_{n\in\mathcal{N}},\{\mathcal{U}_{\mathrm{FC}},\{\mathcal{U}_n\}_{n\in\mathcal{N}}\}\}$。

1) 阶段子博弈。在初始阶段，FC 作为审计者通过访问审计链获取用户感知报告，结合审计预算选择最可疑的 W 个频谱感知终端进行审计。令 $\mathcal{W}=\{1,\cdots,W\}\subseteq\mathcal{N}$ 表示被审计者的集合。在每个阶段博弈（SG, stage game）$\mathcal{G}^{(t)},1\leqslant t\leqslant T$，FC 作为领导者首先决策其惩罚策略，接着 W 个被审计者作为跟随者根据观察到的惩罚策略独立地决策其投毒策略。T 是阶段博弈次数，即最大交互次数。

2) 混合策略。审计者 FC 将其惩罚策略均匀量化为 $Z+1$ 个级别，即 $x_w\in\left\{\frac{z}{Z}x_{\max}\right\}_{0\leqslant z\leqslant Z}$，并在每个 SG 选择如下混合惩罚策略

$$\begin{cases}\boldsymbol{a}=[\boldsymbol{a}_w]_{0\leqslant w\leqslant W},\boldsymbol{a}_w=[a_{w,z}]_{0\leqslant z\leqslant Z}\in\Pi\\ a_{w,z}=\Pr\left(x_w=\dfrac{z}{Z}x_{\max}\right)\end{cases}\tag{10}$$

其中，Π 表示 FC 的惩罚力度动作空间，$x_{\max}$ 表示最大惩罚量，$a_{w,z}$ 表示惩罚策略 x_w 被选中的概率。惩罚力度可由罚款数量进行衡量。

每个被审计者 w 也将其投毒策略均匀量化为 $V+1$ 个级别，即 $y_w\in\left\{\frac{v}{V}y_{\max}\right\}_{0\leqslant v\leqslant V}$，并在每个 SG 选择混合投毒策略如下

$$\begin{cases}\boldsymbol{b}_w=[b_{w,v}]_{0\leqslant v\leqslant V}\in\Psi,\forall w\in\mathcal{W}\\ b_{w,v}=\Pr\left(y_w=\dfrac{v}{V}y_{\max}\right)\end{cases}\tag{11}$$

其中，Ψ 表示被审计者 w 的投毒级别动作空间，$y_{\max}$ 表示最大投毒级别，$b_{w,v}$ 表示投毒策略 y_w 被选中的概率。投毒级别表示频谱感知终端对本地感知报告的伪造程度，即 $S_n\pm\chi y_w$。$\kappa_w=\chi y_w$ 表示攻击强度（单位为 dB），$\chi>0$ 为调节系数，$y_w=0$ 表示真实报告本地感知结果。根据定义可得

$$\sum_{z=0}^{Z}a_{w,z}=\sum_{v=0}^{V}b_{w,v}=1,\ a_{w,z},b_{w,v}\geqslant 0\tag{12}$$

3) 效用。审计者和被审计者 w 在子博弈 $\mathcal{G}^{(t)}$ 的期望效用函数分别表示为 $\mathcal{U}_{\mathrm{FC}}^{(t)}$ 和 $\mathcal{U}_w^{(t)}$。

2.2 效用函数分析

FC 的期望效用函数由其审计成本、投毒造成的协作频谱检测效率损失、投毒行为的负面效应三部分组成，可以表示为

$$\mathcal{U}_{\mathrm{FC}}(\boldsymbol{a},\boldsymbol{b})=\sum_{w\in\mathcal{W}}\mathcal{U}_{\mathrm{FC},w}(\boldsymbol{a}_w,\boldsymbol{b}_w)=\sum_{w=1}^{W}\sum_{z=0}^{Z}\sum_{v=0}^{V}a_{w,z}b_{w,v}\cdot\left[-\varpi_1\lambda_c x_w c^4-\varpi_2\Delta P-\varpi_3\phi\frac{y_w^2}{x_w+\varepsilon}\right]\tag{13}$$

其中，$\varpi_1,\varpi_2,\varpi_3\in[0,1]$ 为权重系数，满足 $\varpi_1+\varpi_2+\varpi_3=1$；$\lambda_c x_w c^4$ 为 FC 的审计成本[34-35]，c^4 为审计的成本参数，c 为正值常数，λ_c 为调节系数；$\frac{\phi y_w^2}{x_w+\varepsilon}$ 为投毒行为的负面效应[34-35]，其值随着审计惩罚力度的提高而趋于缓和，ε 为使分母不为零的较小的正常数，ϕ 为调节系数；ΔP 为投毒造成的协作频谱检测效率损失，其值由以下两部分组成

$$\Delta P=\mu_1\left[P_{\mathrm{d}}\left(\mathcal{N}\setminus\{w\}\right)-P_{\mathrm{d}}(\mathcal{N})\right]+\mu_2\left[P_{\mathrm{f}}(\mathcal{N})-P_{\mathrm{f}}\left(\mathcal{N}\setminus\{w\}\right)\right]\tag{14}$$

其中，$\mu_1,\mu_2\in[0,1]$ 为权重系数，满足 $\mu_1+\mu_2=1$。式(14)中第一项表示被审计者 w 加入后检测率的下降值，第二项表示被审计者 w 加入后虚检率的上升值。

被审计者 $w\in\mathcal{W}$ 的期望效用函数与投毒所得的收益和惩罚造成的损失有关，可表示为

$$\mathcal{U}_w(\boldsymbol{a}_w,\boldsymbol{b}_w)=\sum_{z=0}^{Z}\sum_{v=0}^{V}a_{w,z}b_{w,v}\lambda_r\frac{\sqrt{y_w}}{1+x_w y_w}\tag{15}$$

其中，$\lambda_r\sqrt{y_w}$ 表示被审计者 w 在未被审计时实施投毒策略 y_w 的收益[34-35]，λ_r 表示调节系数。$\frac{1}{1+x_w y_w}$ 表示被审计者 w 避免惩罚（即成功逃逸）的概率[34-35]。

问题 1 审计者的最优化问题。FC 作为审计者，其最优化问题为决定最优审计策略 $\{\mathcal{W},\boldsymbol{a}\}$ 来最大化自身期望效用函数，即

$$\begin{aligned}&\max_{\{\mathcal{W},\boldsymbol{a}\}}\ \mathcal{U}_{\mathrm{FC}}(\boldsymbol{a},\boldsymbol{b})\\ &\text{s.t. }\mathrm{C}_1:\mathcal{W}\subseteq\mathcal{N}\\ &\qquad\mathrm{C}_2:\sum_{z=0}^{Z}a_{w,z}=1,a_{w,z}\geqslant 0\end{aligned}\tag{16}$$

其中，C_1 为审计预算约束。为简化问题分析，FC 从集合 $\mathcal{N}$ 中选择 ΔP 最大的 W 个频谱感知终端组成被审计者集合 $\mathcal{W}\subseteq\mathcal{N}$ 进行审计。

问题 2　被审计者的最优化问题。每个被审计者 $w\in\mathcal{W}$ 的最优化问题为决定最优混合投毒策略 $\boldsymbol{b}_w$ 来最大化自身期望效用函数，即

$$\begin{aligned}&\max_{\boldsymbol{b}_w}\ \mathcal{U}_w\left(\boldsymbol{a}_w,\boldsymbol{b}_w\right)\\&\text{s.t. } C_3:\mathcal{W}\subseteq\mathcal{N}\\&\quad C_4:\sum_{v=0}^{V}b_{w,v}=1,b_{w,v}\geqslant 0\end{aligned}\tag{17}$$

2.3　博弈均衡策略

在重复序贯审计博弈 $\mathcal{G}$ 中，令 $\boldsymbol{a}^*=\left[\boldsymbol{a}_w^*\right]_{0\leqslant w\leqslant W}$ 表示审计者的最优混合惩罚策略，$\boldsymbol{b}_w^*$ 表示被审计者 w 的最优混合投毒策略。在审计过程中，假设审计双方都是理性和自私的，即双方目的均为最大化自身效用。序贯博弈 $\mathcal{G}$ 的解是 Stackelberg 均衡点[23]，即审计者和所有被审计者都不能通过偏离它来提高效用，其定义如下。

定义 2　审计博弈 Stackelberg 均衡。若下述条件成立，则审计博弈 $\mathcal{G}$ 的 Stackelberg 均衡点（若存在）为

$$\begin{cases}\mathcal{U}_{\mathrm{FC}}(\boldsymbol{a}^*,\boldsymbol{b}^*)\geqslant\mathcal{U}_{\mathrm{FC}}(\boldsymbol{a},\boldsymbol{b}^*),\forall\boldsymbol{a}\in\Pi^w\\\mathcal{U}_w\left(\boldsymbol{a}_w^*,\boldsymbol{b}_w^*\right)\geqslant\mathcal{U}_w\left(\boldsymbol{a}_w^*,\boldsymbol{b}_w\right),\forall w\in\mathcal{W},\forall\boldsymbol{b}_w\in\Psi\end{cases}\tag{18}$$

2.4　完全信息下的纯策略 Stackelberg 均衡

纯策略是混合策略的特例，指代每个参与者只能选择某一种特定策略。采用逆向归纳法[36]来获取每个序贯子博弈的 Stackelberg 均衡。具体地，首先在引理 1 中分析跟随者即被审计者的最优投毒策略，在此基础上，在引理 2 中分析领导者即审计者的最优惩罚策略。

引理 1　在纯策略下，被审计者 w 的最优投毒策略为

$$y_w^*=\min\left(\frac{1}{x_w},y_{\max}\right)\tag{19}$$

证明　根据式(15)，令 $\dfrac{\partial\mathcal{U}_w\left(x_w,y_w\right)}{\partial y_w}=0$，计算可得 $y_w=\dfrac{1}{x_w}$。当 $y_w<\dfrac{1}{x_w}$ 时，有 $\dfrac{\partial\mathcal{U}_w(x_w,y_w)}{\partial y_w}>0$；当 $y_w>\dfrac{1}{x_w}$ 时，有 $\dfrac{\partial\mathcal{U}_w(x_w,y_w)}{\partial y_w}<0$。由于 $0\leqslant y_w\leqslant y_{\max}$，可得 $y_w^*=\min\left(\dfrac{1}{x_w},y_{\max}\right)$。证毕。

引理 2　在纯策略下，当 $\varepsilon\to 0$ 时，审计者的最优惩罚策略为

$$x_w^*=\begin{cases}\min\left(\dfrac{\sqrt[4]{3}\Theta}{c},x_{\max}\right), & \dfrac{\sqrt[4]{3}y_{\max}\Theta}{c}>1\\\min\left(\dfrac{y_{\max}\Theta^2}{c^2},x_{\max}\right), & \dfrac{y_{\max}\Theta}{c}\leqslant 1\end{cases}\tag{20}$$

其中，$\Theta=\sqrt[4]{\dfrac{\varpi_3\phi}{\varpi_1\lambda_c}}$。

证明　根据引理 1，考虑以下 2 种情况。

情况 1　$\dfrac{1}{x_w}<y_{\max}$，即 $x_w>\dfrac{1}{y_{\max}}$，此时 $y_w^*=\dfrac{1}{x_w}$。将 y_w^* 代入式(13)，当 $\varepsilon\to 0$ 可得

$$\frac{\partial\mathcal{U}_{\mathrm{FC}}(\boldsymbol{x},\boldsymbol{y}^*)}{\partial x_w}=-\varpi_1\lambda_c c^4+3\varpi_3\phi x_w^{-4}\tag{21}$$

令式(21)为零，可得 $x_w^*=\sqrt[4]{\dfrac{3\varpi_3\phi}{\varpi_1\lambda_c}}\dfrac{1}{c}=\sqrt[4]{3}\dfrac{\Theta}{c}$。上述最优解须满足 $x_w^*>\dfrac{1}{y_{\max}}$，即 $\dfrac{\sqrt[4]{3}y_{\max}\Theta}{c}>1$。

情况 2　$\dfrac{1}{x_w}\geqslant y_{\max}$，即 $x_w\leqslant\dfrac{1}{y_{\max}}$，此时 $y_w^*=y_{\max}$。将 y_w^* 代入式(13)，当 $\varepsilon\to 0$ 可得

$$\frac{\partial\mathcal{U}_{\mathrm{FC}}(\boldsymbol{x},\boldsymbol{y}^*)}{\partial x_w}=-\varpi_1\lambda_c c^4+\varpi_3\phi y_{\max}^2 x_w^{-2}\tag{22}$$

令式(22)为零，可得 $x_w^*=\sqrt{\dfrac{\varpi_3\phi}{\varpi_1\lambda_c}}\dfrac{y_{\max}}{c^2}=\dfrac{\Theta^2}{c^2}y_{\max}$。上述最优解须满足 $x_w^*\leqslant\dfrac{1}{y_{\max}}$，通过简单换算可得 $\dfrac{y_{\max}\Theta}{c}\leqslant 1$。证毕。

定理 1　在纯策略下，当 $\varepsilon\to 0$ 时，审计博弈 $\mathcal{G}$ 的 Stackelberg 均衡存在，并由式(23)给出。

$$\left(x_w^*,y_w^*\right)=\begin{cases}\left(\min\left(\sqrt[4]{3}\dfrac{\Theta}{c},x_{\max}\right),\dfrac{1}{x_w^*}\right), & \sqrt[4]{3}y_{\max}\dfrac{\Theta}{c}>1\\\left(\min\left(y_{\max}\dfrac{\Theta^2}{c^2},x_{\max}\right),y_{\max}\right), & y_{\max}\dfrac{\Theta}{c}\leqslant 1\end{cases}\tag{23}$$

证明 根据引理1和引理2，当$\frac{\sqrt[4]{3}y_{\max}\Theta}{c}>1$时，子博弈的纯策略Stackelberg均衡为$\left(x_w^*=\min\left(\frac{\sqrt[4]{3}\Theta}{c},x_{\max}\right),y_w^*=\frac{1}{x_w^*}\right)$；当$\frac{y_{\max}\Theta}{c}\leqslant 1$时，均衡为$\left(x_w^*=\min\left(\frac{y_{\max}\Theta^2}{c^2},x_{\max}\right),y_w^*=y_{\max}\right)$。证毕。

3 不完全信息下混合策略Stackelberg均衡

3.1 方案流程

在实际环境中，由于参与者隐私参数和混合策略的存在以及信道环境的高度动态性，导致审计双方的双向信息不对称，如何获取不完全信息下混合策略Stackelberg均衡成为挑战。动态重复序贯审计博弈$\mathcal{G}$中，审计者和相关被审计者可以分别利用强化学习方法，通过在重复交互中反复试错寻找出最优混合审计策略和最优混合投毒策略，从而在动态环境下自适应且智能地计算所提博弈的近似均衡策略。每个阶段博弈$\mathcal{G}^{(t)},1\leqslant t\leqslant T$包括以下步骤。

1) 审计者根据上一轮博弈中被审计者的投毒策略，利用强化学习算法计算对所有被审计者的最优惩罚策略$\left[a_1^{(t)},a_2^{(t)},\cdots,a_W^{(t)}\right]$。

2) 每个被审计者w分布式且并行地根据当前观测到的审计惩罚策略，制定对本地频谱感知报告的最优投毒策略$b_w^{(t)}$。

3.2 基于PHC学习的审计者最优惩罚策略

在重复序贯审计博弈$\mathcal{G}$中，审计者的惩罚策略$\left\{\boldsymbol{a}^{(1)},\boldsymbol{a}^{(2)},\cdots,\boldsymbol{a}^{(T)}\right\}$制定过程可以建模为有限的马尔可夫决策过程[37]（MDP, Markov decision process），因此，FC作为审计者可以应用PHC学习等无模型的强化学习方法，在没有足够的被审计者隐私参数信息与背景知识的情况下制定其最优惩罚策略。主要包含以下几个部分。

1) 状态。在每个子博弈$\mathcal{G}^{(t)}$，FC观察到系统状态向量$\boldsymbol{s}^{(t)}=\left(s_1^{(t)},\cdots,s_w^{(t)},\cdots,s_W^{(t)}\right)$，它包含上一次交互时所有被审计者的投毒动作序列，即$\boldsymbol{s}^{(t)}=\boldsymbol{y}^{(t-1)}=\left[y_w^{(t-1)}\right]_{w\in\mathcal{W}}$。

2) 动作。在每个子博弈$\mathcal{G}^{(t)}$，FC以概率$\pi(\boldsymbol{s}^{(t)},\boldsymbol{x}^{(t)})$选择其惩罚动作向量$\boldsymbol{x}^{(t)}=\left[x_w^{(t)}\right]_{w\in\mathcal{W}}$。其中，$\pi(\boldsymbol{s}^{(t)},\boldsymbol{x}^{(t)})$为混合策略表，其初值为$\pi\left(s_w^{(0)},x_w^{(0)}\right)=\frac{1}{Z+1},\forall w$。

3) 奖励。在每个子博弈$\mathcal{G}^{(t)}$，FC的即时奖励值为式(13)中定义的效用值$\mathcal{U}_{\mathrm{FC}}^{(t)}=\mathcal{U}_{\mathrm{FC}}(\boldsymbol{s}^{(t)},\boldsymbol{x}^{(t)})$。令$\mathcal{Q}(\boldsymbol{s}^{(t)},\boldsymbol{x}^{(t)})$表示状态–动作对$(\boldsymbol{s}^{(t)},\boldsymbol{x}^{(t)})$下的Q函数，即该状态–动作对的累计贴现效用的期望值。基于迭代贝尔曼方程，FC可通过式(24)更新其Q函数。

$$\mathcal{Q}\left(s_w^{(t)},x_w^{(t)}\right)\leftarrow\mathcal{Q}\left(s_w^{(t)},x_w^{(t)}\right)+\eta_1\left\{\mathcal{U}_{\mathrm{FC},w}^{(t)}+\gamma_1\max\nolimits_{x_w^{(t+1)}}\mathcal{Q}\left(s_w^{(t+1)},x_w^{(t+1)}\right)-\mathcal{Q}\left(s_w^{(t)},x_w^{(t)}\right)\right\}\tag{24}$$

其中，$\eta_1,\gamma_1\in(0,1]$分别表示学习率和贴现率。Q函数的初值为$\mathcal{Q}\left(s_w^{(0)},x_w^{(0)}\right)=0,\forall w$。

为了在Q学习过程中实现探索和开发间的权衡，FC对混合策略表$\pi(\boldsymbol{s}^{(t)},\boldsymbol{x}^{(t)})$进行动态更新。具体地，对贪婪惩罚动作（即可最大化Q函数）的概率增加一个较小的数值δ_1，而对其他动作（可能会导致更好的未来回报）的概率减小$\frac{\delta_1}{Z}$，即

$$\pi\left(s_w^{(t)},x_w^{(t)}\right)\leftarrow\pi\left(s_w^{(t)},x_w^{(t)}\right)+\begin{cases}\delta_1,\ x_w^{(t)}=\arg\max_{x_w}\mathcal{Q}\left(s_w^{(t)},x_w\right)\\-\dfrac{\delta_1}{Z},\ \text{其他}\end{cases}\tag{25}$$

基于赢或学得快策略[38]（WoLF, win or learn fast），变量δ_1有2个数值，即δ_1^{h}和δ_1^{l}，且$\delta_1^{\mathrm{h}}>\delta_1^{\mathrm{l}}$。其值的选择取决于FC的当前WoLF状态，即

$$\delta_1=\begin{cases}\delta_1^{\mathrm{h}},\sum_{x_w}\pi\left(s_w^{(t)},x_w\right)\mathcal{Q}\left(s_w^{(t)},x_w\right)\leqslant\sum_{a_n}\bar{\pi}\left(s_w^{(t)},x_w\right)\mathcal{Q}\left(s_w^{(t)},x_w\right)\\\delta_1^{\mathrm{l}},\text{其他}\end{cases}\tag{26}$$

其中，平均混合策略表$\bar{\pi}\left(s_w^{(t)},x_w\right)$可表示为

$$\bar{\pi}\left(s_w^{(t)},x_w^{(t)}\right)\leftarrow\bar{\pi}\left(s_w^{(t)},x_w^{(t)}\right)+\frac{\pi\left(s_w^{(t)},x_w^{(t)}\right)-\bar{\pi}\left(s_w^{(t)},x_w^{(t)}\right)}{\mathrm{count}\left(s_w^{(t)}\right)}\tag{27}$$

其中，$\mathrm{count}\left(s_w^{(t)}\right)$表示状态$s_w^{(t)}$到当前时刻被观察到的次数。

3.3　基于 PHC 学习的被审计者最优投毒策略

每个被审计者 $w\in\mathcal{W}$ 在重复序贯审计博弈 $\mathcal{G}$ 中的投毒策略 $\left\{\boldsymbol{b}_w^{(1)},\boldsymbol{b}_w^{(2)},\cdots,\boldsymbol{b}_w^{(T)}\right\}$ 制定过程可建模为有限 MDP[37]，因此，每个被审计者可应用 PHC 学习在双向信息不对称的情况下制定其最优投毒策略。该策略主要包含以下几个部分。

1) 状态。在每个子博弈 $\mathcal{G}^{(t)}$，被审计者 w 观察到系统状态 $\hat{s}_w^{(t)}=x_w^{(t)}$，即当前交互时审计者的惩罚动作。

2) 动作。在每个子博弈 $\mathcal{G}^{(t)}$，被审计者 w 以概率 $\pi\left(\hat{s}_w^{(t)},y_w^{(t)}\right)$ 选择其投毒动作 $y_w^{(t)}\in\Psi$。其中，$\pi\left(\hat{s}_w^{(t)},y_w^{(t)}\right)$ 为被审计者 w 的混合策略表，其初值为 $\pi\left(\hat{s}_w^{(0)},y_w^{(0)}\right)=\dfrac{1}{V+1},\forall w$。

3) 奖励。在每个子博弈 $\mathcal{G}^{(t)}$，被审计者 w 的即时奖励值为式 (15) 中定义的效用值 $\mathcal{U}_w^{(t)}=\mathcal{U}_w\left(\hat{s}_w^{(t)},y_w^{(t)}\right)$。Q 函数 $\mathcal{Q}\left(\hat{s}_w^{(t)},y_w^{(t)}\right)$ 表示状态−动作对 $\left(\hat{s}_w^{(t)},y_w^{(t)}\right)$ 下的长期累计贴现效用的期望值。被审计者 w 基于迭代贝尔曼方程更新其 Q 函数为

$$\mathcal{Q}\left(\hat{s}_w^{(t)},y_w^{(t)}\right)\leftarrow\mathcal{Q}\left(\hat{s}_w^{(t)},y_w^{(t)}\right)+\eta_2\left\{\mathcal{U}_w^{(t)}+\gamma_2\max_{y_w^{(t+1)}}\mathcal{Q}\left(\hat{s}_w^{(t+1)},y_w^{(t+1)}\right)-\mathcal{Q}\left(\hat{s}_w^{(t)},y_w^{(t)}\right)\right\}\tag{28}$$

其中，$\eta_2,\gamma_2\in(0,1]$ 分别表示学习率和贴现率。Q 函数的初值为 $\mathcal{Q}\left(\hat{s}_w^{(0)},y_w^{(0)}\right)=0,\forall w$。

同样地，为实现 Q 学习中探索和开发间的高效权衡，被审计者 w 基于 WoLF 策略动态更新其混合策略表 $\pi\left(\hat{s}_w^{(t)},y_w^{(t)}\right)$。具体地，对贪婪惩罚动作（即可最大化 Q 函数）的概率增加一个较小的数值 δ_2，而对其他动作（可能会导致更好的未来回报）的概率减小 $\dfrac{\delta_2}{V}$，即

$$\pi\left(\hat{s}_w^{(t)},y_w^{(t)}\right)\leftarrow\pi\left(\hat{s}_w^{(t)},y_w^{(t)}\right)+\begin{cases}\delta_2, y_w^{(t)}=\arg\max_{y_w}\mathcal{Q}\left(\hat{s}_w^{(t)},y_w\right)\\ -\dfrac{\delta_2}{V},\text{其他}\end{cases}\tag{29}$$

变量 δ_2 有 2 个数值，即 δ_2^{h} 和 δ_2^{l}，且 $\delta_2^{\mathrm{h}}>\delta_2^{\mathrm{l}}$，其值取决于被审计者的当前 WoLF 状态，即

$$\delta_2=\begin{cases}\delta_2^{\mathrm{h}},\sum_{y_w}\pi\left(\hat{s}_w^{(t)},y_w\right)\mathcal{Q}\left(\hat{s}_w^{(t)},y_w\right)\leqslant\sum_{a_n}\bar{\pi}\left(\hat{s}_w^{(t)},y_w\right)\mathcal{Q}\left(\hat{s}_w^{(t)},y_w\right)\\ \delta_2^{\mathrm{l}},\text{其他}\end{cases}\tag{30}$$

同样地，平均混合策略表 $\bar{\pi}\left(\hat{s}_w^{(t)},y_w\right)$ 可表示为

$$\bar{\pi}\left(\hat{s}_w^{(t)},y_w^{(t)}\right)\leftarrow\bar{\pi}\left(\hat{s}_w^{(t)},y_w^{(t)}\right)+\frac{\pi\left(\hat{s}_w^{(t)},y_w^{(t)}\right)-\bar{\pi}\left(\hat{s}_w^{(t)},y_w^{(t)}\right)}{\mathrm{count}\left(\hat{s}_w^{(t)}\right)}\tag{31}$$

其中，$\mathrm{count}\left(\hat{s}_w^{(t)}\right)$ 表示状态 $\hat{s}_w^{(t)}$ 到当前时刻被观察到的次数。

基于 PHC 学习的最优审计策略决策算法如算法 1 所示。其中，第 2)~6)行是审计者基于 PHC 学习的最优惩罚策略决策过程，第 7)~12)行是每个被审计者基于 PHC 学习的最优投毒策略决策过程。

算法 1　基于 PHC 学习的最优审计策略决策算法

初始化　$t=0$；混合策略表 $\pi\left(s_w^{(0)},x_w^{(0)}\right)=\dfrac{1}{Z+1}$，$\pi\left(\hat{s}_w^{(0)},y_w^{(0)}\right)=\dfrac{1}{V+1}$，$\forall w$；Q 函数 $\mathcal{Q}\left(s_w^{(0)},x_w^{(0)}\right)=0$，$\mathcal{Q}\left(\hat{s}_w^{(0)},y_w^{(0)}\right)=0$，$\forall w$；学习率 η_1 和 η_2；贴现率 γ_1 和 γ_2；WoLF 变量 δ_1^{h}，δ_1^{l}，δ_2^{h} 和 δ_2^{l}；交互次数 T

1) for t =1:1:T

基于 PHC 的最优惩罚策略（由审计者运行）；

2) 观测系统状态 $\boldsymbol{s}^{(t)}=\boldsymbol{y}^{(t-1)}$；

3) 根据混合策略表选择惩罚动作向量 $\boldsymbol{x}^{(t)}$；

4) 评估即时奖励值 $\mathcal{U}_{\mathrm{FC}}^{(t)}$；

5) 根据式(24)更新 Q 函数 $\mathcal{Q}(s_w^{(t)},x_w^{(t)})$；

6) 根据式 (25)~ 式 (27) 更新混合策略表 $\pi(\boldsymbol{s}^{(t)},\boldsymbol{x}^{(t)})$；

基于 PHC 的最优投毒策略（由 W 个被审计者独立运行）；

7) 观测系统状态 $\hat{s}_w^{(t)}=x_w^{(t)}$；

8) 根据混合策略表选择投毒动作 $y_w^{(t)}$；

9) 评估即时奖励值 $\mathcal{U}_w^{(t)}$；

10) 根据式(28)更新 Q 函数 $\mathcal{Q}\left(\hat{s}_w^{(t)},y_w^{(t)}\right)$；

11) 根据式 (29) ～式 (31) 更新混合策略表 $\pi\left(\hat{s}_w^{(t)},y_w^{(t)}\right)$；

12) end for

3.4 算法性质分析

1) 复杂度分析。算法1中，对于审计者，其运行 PHC 学习算法的时间复杂度和空间复杂度分别为$\mathcal{O}(TWZ)$和$\mathcal{O}(TWZV)$，其中，T 是审计双方的最大交互次数；W 是被审计者数量，表征审计规模。对于每个被审计者，其分布式运行 PHC 学习算法的计算复杂度为$\mathcal{O}(TV)$。此外，每个被审计者不用观测其他被审计者的策略、动作及奖励值，只保存自己的动作值来维护并更新 Q 表和混合策略表。因此，对于每个被审计者，其空间复杂度降低至$\mathcal{O}(TZV)$。因此，被审计者的最优策略决策过程的计算和存储成本与审计规模 W 无关。本文假设手机、平板电脑等频谱感知协作终端具有一定的计算和存储能力，因此可支持所提低开销的分布式 PHC 算法。

2) 收敛性分析。根据文献[39]，当$t\to\infty$时，Q 学习可收敛至最优的动作值函数。在 PHC 学习中，一方面通过增大最大累积期望所对应的动作的选择概率；另一方面，当智能体的表现优于预期时，缓慢调整其 Q 学习参数，而当其表现差于预期时，加速调整参数，从而快速调整策略以适应动态变化的环境，并提升 Q 学习算法收敛速度。所提 PHC 学习算法的收敛性在第4节中通过仿真实验进行了验证。

3) 稳健性分析。在所提 PHC 学习算法中，每个被审计者独立地且分布式地执行 PHC 学习过程，同时不用观测其他被审计者的策略、动作及奖励值。因此，当被审计者数量动态变化时不需要重新训练 PHC 学习模型，且不会影响其他被审计者的策略决策过程，保障了算法的稳健性。同时，审计双方可利用相似场景下的历史交互数据来离线预训练其 PHC 模型，从而加速算法的收敛速率。

4) 可扩展性分析。根据上述分析，所提 PHC 学习算法在被审计者数量增加或减小时不需要重新训练。同时，被审计者以分布式的方式并行地在本地运行其 PHC 学习模型。因此，所提 PHC 学习算法可适用于不同审计规模下的协作频谱感知场景，具备高可扩展性。

4 仿真分析

4.1 仿真设置

根据文献[6]，仿真区域为一个 5 km×5 km 的方形区域，考虑 IEEE 802.22 WRAN 的仿真环境，区域内拥有 6 MHz 带宽和 150.3 km 传输范围的单个数字电视发射机、100 个 SU 和感知协助设备等频谱感知终端。PU 到蜂窝中心的距离设置为 145 km。设置 2 个感知终端之间的最小距离为 200 m，以去相关阴影衰减效应 A_n。恶意频谱感知终端的数量为5～50。总帧持续时间 T=20 ms，能量检测的采样样本数设置为 h=6 000，能量检测的最小概率要求大于 0.9[10]。每个频谱感知终端接收的 PU 信号的信噪比在 0～30 dB 均匀分布[10]。其他仿真参数如表 1 所示。

表 1 仿真参数

模型	参数	数值
信道模型[6]	ι	3.7
	d_0 /m	1
	σ_n /dB	5.5
	P_0 /dBm	90
审计模型参数	$\varpi_1,\varpi_2,\varpi_3$	$\frac{1}{3}$
	λ_c	3
	ϕ	3
	λ_r	25
	c	1
	ε	0.001
	$x_{\max}$	1
	$y_{\max}$	5.5
	Z	10
	V	11
	χ	$\frac{3}{55}$
PHC 模型	η_1,η_2	0.7
	γ_1,γ_2	0.8
	$\delta_1^{\mathrm{l}},\delta_2^{\mathrm{l}}$	$\frac{1}{50+\frac{t}{50}}$
	$\delta_1^{\mathrm{h}},\delta_2^{\mathrm{h}}$	$\frac{2}{50+\frac{t}{50}}$

考虑以下 5 种基准方案作为对比方案。

1) 单层 Q 学习方案。该方案中，审计者的惩罚策略固定且事先公开，每个被审计者均采用传统 Q 学习算法和ϵ−贪婪策略在每个子博弈中计算最优投毒策略。其中，被审计者以高概率$\epsilon=0.95$贪婪地选择可最大化 Q 函数的动作，以低概率$1-\epsilon$=0.05 随机选择其他动作。学习率和贴现率保持不变。

2) 随机方案。该方案中，审计者和每个被审计者均在 Q 学习中采用随机策略决策每个子博弈下的惩罚策略和投毒策略。

3) 完全信息下纯策略 Stackelberg 均衡。该方案考虑完全信息下审计双方的纯策略决策，其中审计者和每个被审计者在每个子博弈下根据定理 1 中的 Stackelberg 均衡策略进行决策。

4) 基准方案 1[40]。FC 融合协作用户频谱感知报告来决策 PU 状态，该方案不考虑恶意投毒用户。

5) 基准方案 2[15]。FC 基于历史信任评估策略检测恶意用户的频谱感知报告。

4.2 仿真结果

首先，对审计者与被审计者在 PHC 学习中的最优策略演变和效用演变过程进行评估；接着，对不同方案下审计者与被审计者的效用进行性能评估；最后，评估不同方案下的虚警率和漏检率。另外，所部署以太坊私有链的平均区块出块间隔可控制在 114～420 μs。可以看出，所部署审计链具有低处理时延，可适用于频谱感知数据审计等场景。

在 PHC 学习中审计者的惩罚力度和某个随机选择的被审计者的投毒级别随迭代次数的变化如图 2 和图 3 所示。从图 2 和图 3 可以看出，相比完全信息下的纯策略 Stackelberg 均衡，所提方案可以在不完全信息和混合策略下获取审计双方的渐近均衡策略，并在约 1 600 次迭代后达成收敛。具体地，当观察到审计者的初始较低的惩罚力度之后，被审计者倾向于选择较高的本地频谱感知数据投毒级别。接着，当观察到被审计者较高的投毒级别之后，审计者倾向于逐渐提高其惩罚力度，以抑制频谱感知终端的作恶行为。然后，当观察到审计者逐步提高审计惩罚力度之后，被审计者倾向于逐渐降低对本地频谱感知数据的投毒级别，最终不断逼近纳什均衡状态。

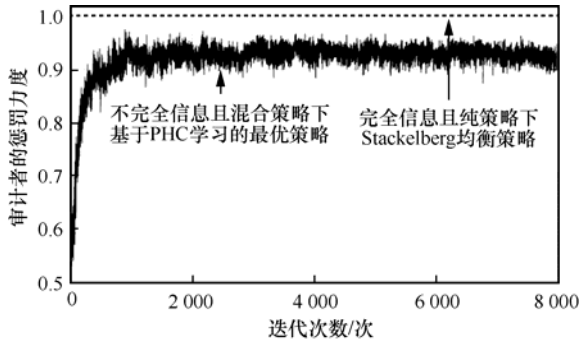


图 2　审计者的惩罚力度随迭代次数的变化

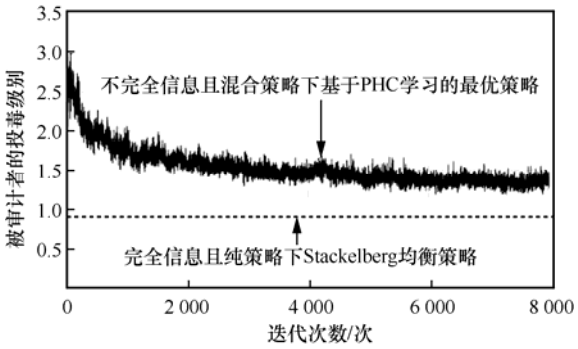


图 3　某个随机选择的被审计者的投毒级别随迭代次数的变化

在 PHC 学习中审计双方的效用随迭代次数的变化如图 4 所示。从图 4 可以看出，相比完全信息下的纯策略 Stackelberg 均衡，所提方案可以在不完全信息和混合策略下获取审计双方的近似最优效用。在重复序贯审计博弈中，审计者倾向于逐步提升其惩罚力度以威慑并抑制频谱感知终端的作恶程度，从而逐步提升其效用。同时，被审计者倾向于逐步减轻其投毒级别以降低审计惩罚措施造成的损失，从而逐步降低其效用。随着双方交互次数的不断增加，审计双方的效用值在约 1 000 次迭代后达到稳定值，即逼近纳什均衡状态。

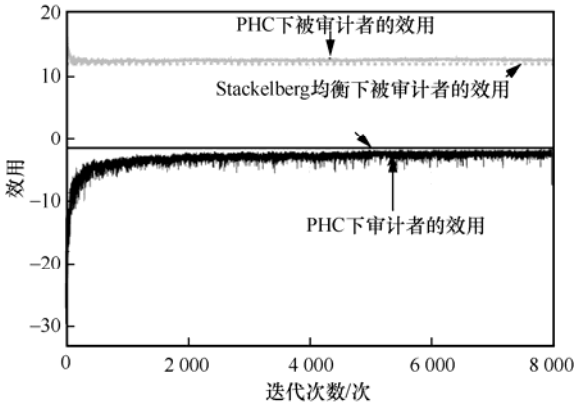


图 4　审计双方的效用随迭代次数的变化

审计成本参数 c 从 0.2 增长到 2，所提方案与传统单层 Q 学习方案、随机方案渐近达成均衡状态时，关于审计者效用和被审计者效用对比分别如图 5 和图 6 所示。仿真结果表明，在不同审计成本参数下，所提方案可为审计双方获取最大效用，优于其他 2 种方案。这是因为在单层 Q 学习方案中，审计者的惩罚策略是固定的且惩罚力度相对较高，因此不能随被审计者投毒策略的变化而动态调整，导致被审计者不能获得全局最优的效用值。在随机方案中，由于频谱感知数据审计中审计双方的惩罚策略

和投毒策略均是随机选择的，导致被审计者只能获得较低的效用值。在所提方案中，审计双方通过分布式应用 PHC 学习以在高度动态的网络中获取近似最优审计策略和投毒策略，从而最大化各自的效用。另外，从图 5 和图 6 中可以看出，审计者的效用随着审计成本的增加而不断减小，而被审计者的效用随着审计成本的增加而不断增加。这是因为随着审计成本的增加，审计者倾向于降低其惩罚力度以减小整体审计成本，而被审计者倾向于增加投毒级别策略来增加潜在收益，从而导致审计者效用的降低和被审计者效用的增加。

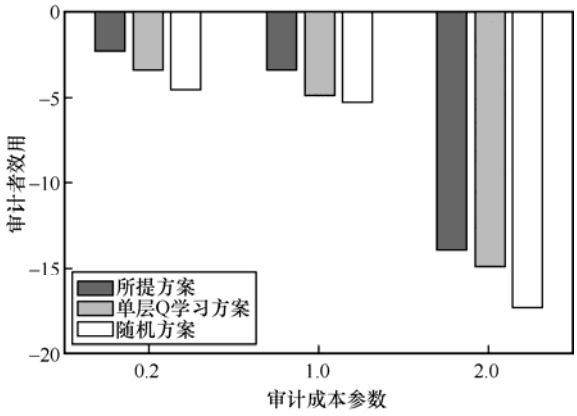


图 5　不同方案和审计成本参数下渐近达成均衡状态时审计者效用对比

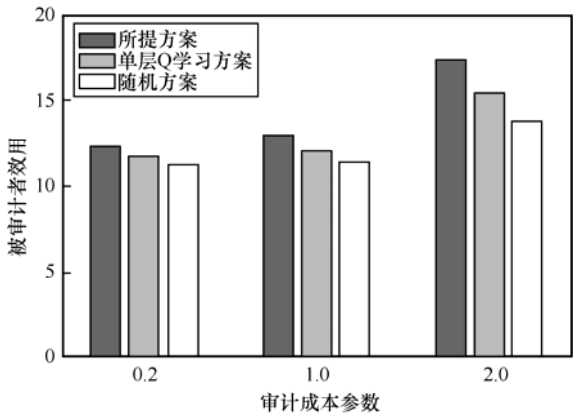


图 6　不同方案和审计成本参数下渐近达成均衡状态时被审计者效用对比

在不同攻击强度 κ_w 和恶意用户数量下，所提方案与 2 个基准方案[15,40]的虚警率和漏检率对比如图 7 和图 8 所示。设置 40 个恶意用户来验证极端情况下所提方案的鲁棒性和检测效率。从图 7 和图 8 可以看出，随着攻击强度的变化（即从 0 变到±3 dB），3 种方案下的虚警率和漏检率均上升，而所提方案的虚警率和漏检率的增长速度低于其他 2 种基准方案[15,40]。同时，随着恶意用户数量的增加，3 种方案的虚警率和漏检率均上升。验证了威胁模型中 2 种投毒攻击的有效性。另外，相比于基准方案[15,40]，所提方案下可以获取更低的虚警率和漏检率。这是由于所提方案通过审计博弈实施事前威慑和事后惩罚，来有效抑制恶意用户的投毒级别，从而减轻投毒攻击的效果。

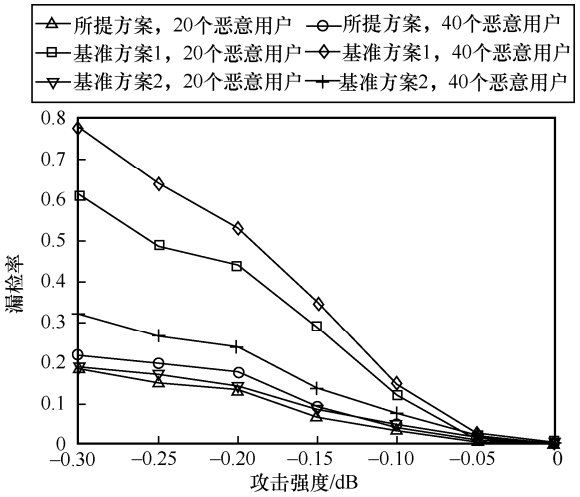


图 7　在不同攻击强度 κ_w 和恶意用户数量下，不同方案的漏检率对比

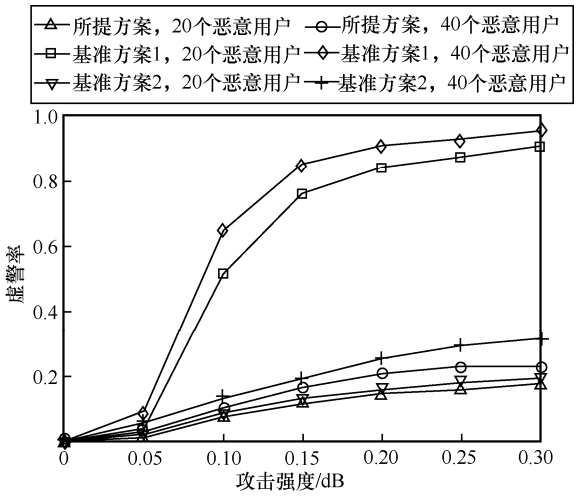


图 8　在不同攻击强度 κ_w 和恶意用户数量下，不同方案的虚警率对比

5　结束语

群智协作频谱感知通过利用分布广泛、低成本、可动态部署的智能物联网终端协助频谱感知，提供了一种低成本、高弹性、可按需的新型频谱感知范式，从而支撑高带宽、高动态、广连接、低时延的多样化物联网服务。本文针对群智协作频谱感知系统中的投毒攻击与搭便车攻击，提出一种基于审计博弈的新型防御机制。考虑审计预算约束和双

向信息不对称环境，建立 FC 与协作终端间的混合策略重复序贯审计博弈模型，融合事前威慑和事后惩罚机制抑制协作终端的恶意行为。结合链上链下协同机制设计轻量审计区块链模型，来不可更改地记录审计证据与结果，从而追溯用户的恶意行为。设计基于强化学习的分布式智能审计算法，使在动态环境下自适应地计算博弈中审计双方的渐近混合策略均衡。仿真结果表明，本文方案具有较好的收敛性、稳健性和投毒行为防御性能。未来工作中，将在博弈框架下对强化学习的策略搜索进行约束和优化，以更快速稳定地获取不完全信息且混合策略下的均衡策略。另外，将进一步研究设备精度误差所造成的误检现象的对策。

参考文献：

[1] WANG Y T, SU Z, NI J B, et al. Blockchain-empowered space-air-ground integrated networks: opportunities, challenges, and solutions[J]. IEEE Communications Surveys & Tutorials, 2022, 24(1): 160-209.

[2] ZHOU Y Q, LIU L, WANG L, et al. Service-aware 6G: an intelligent and open network based on the convergence of communication, computing and caching[J]. Digital Communications and Networks, 2020, 6(3): 253-260.

[3] WANG Y T, PAN Y H, YAN M, et al. A survey on ChatGPT: AI-generated contents, challenges, and solutions[J]. IEEE Open Journal of the Computer Society, 2023, 4: 280-302.

[4] 杨静雅, 唐晓刚, 周一青, 等. 意图抽象与知识联合驱动的 6G 内生智能网络架构[J]. 通信学报, 2023, 44(2): 12-26.
YANG J Y, TANG X G, ZHOU Y Q, et al. 6G native intelligence network architecture enabled by intent abstraction and knowledge[J]. Journal on Communications, 2023, 44(2): 12-26.

[5] LU W D, HU S, LIU X, et al. Incentive mechanism based cooperative spectrum sharing for OFDM cognitive IoT network[J]. IEEE Transactions on Network Science and Engineering, 2020, 7(2): 662-672.

[6] ZHANG R, ZHANG J X, ZHANG Y C, et al. Secure crowdsourcing-based cooperative spectrum sensing[C]//Proceedings of IEEE INFOCOM. Piscataway: IEEE Press, 2013: 2526-2534.

[7] 曹龙, 赵杭生, 鲍丽娜, 等. 基于辅助节点的安全协作频谱感知[J]. 计算机工程, 2014, 40(2): 123-127, 139.
CAO L, ZHAO H S, BAO L N, et al. Secure cooperative spectrum sensing based on helper nodes[J]. Computer Engineering, 2014, 40(2): 123-127, 139.

[8] JANA S, ZENG K, CHENG W, et al. Trusted collaborative spectrum sensing for mobile cognitive radio networks[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(9): 1497-1507.

[9] SAGDUYU Y E, SHI Y, ERPEK T. Adversarial deep learning for over-the-air spectrum poisoning attacks[J]. IEEE Transactions on Mobile Computing, 2021, 20(2): 306-319.

[10] XU Z Y, SUN Z G, GUO L L. Throughput maximization of collaborative spectrum sensing under SSDF attacks[J]. IEEE Transactions on Vehicular Technology, 2021, 70(8): 8378-8383.

[11] WANG Y T, SU Z, LUAN T H, et al. Federated learning with fair incentives and robust aggregation for UAV-aided crowdsensing[J]. IEEE Transactions on Network Science and Engineering, 2022, 9(5): 3179-3196.

[12] 孙志国, 任欣悦, 陈增茂, 等. 基于证据间相似性的协作频谱感知方法与性能分析[J]. 通信学报, 2020, 41(12): 139-147.
SUN Z G, REN X Y, CHEN Z M, et al. Cooperative spectrum sensing method and performance analysis based on similarity between evidences[J]. Journal on Communications, 2020, 41(12): 139-147.

[13] 王小毛, 黄传河, 吕怡龙, 等. 模拟人群信任和决策机制的协作频谱感知方法[J]. 通信学报, 2014, 35(3): 94-108.
WANG X M, HUANG C H, LYU Y L, et al. Cooperative spectrum sensing scheme based on crowd trust and decision-making mechanism[J]. Journal on Communications, 2014, 35(3): 94-108.

[14] FENG J Y, LI S P, LV S Q, et al. Securing cooperative spectrum sensing against collusive false feedback attack in cognitive radio networks[J]. IEEE Transactions on Vehicular Technology, 2018, 67(9): 8276-8287.

[15] LUO X Q. Secure cooperative spectrum sensing strategy based on reputation mechanism for cognitive wireless sensor networks[J]. IEEE Access, 2020, 8: 131361-131369.

[16] HU Y D, ZHANG R. A spatiotemporal approach for secure crowdsourced radio environment map construction[J]. IEEE/ACM Transactions on Networking, 2020, 28(4): 1790-1803.

[17] 吴晓晓, 李刚强, 张胜利. 分布式协作频谱感知网络中恶意节点检测和定位方法研究[J]. 电子学报, 2022, 50(6): 1370-1380.
WU X X, LI G Q, ZHANG S L. Detection and localization of malicious nodes in distributed cooperative spectrum sensing network[J]. Acta Electronica Sinica, 2022, 50(6): 1370-1380.

[18] ZHANG Y R, WU Q R, SHIKH-BAHAEI M R. On ensemble learning-based secure fusion strategy for robust cooperative sensing in full-duplex cognitive radio networks[J]. IEEE Transactions on Communications, 2020, 68(10): 6086-6100.

[19] ZHANG Y, LI A, LI J W, et al. SpecKriging: GNN-based secure cooperative spectrum sensing[J]. IEEE Transactions on Wireless Communications, 2022, 21(11): 9936-9946.

[20] DU B, XUE R, ZHAO L, et al. Coalitional graph game for air-to-air and air-to-ground cognitive spectrum sharing[J]. IEEE Transactions on Aerospace and Electronic Systems, 2020, 56(4): 2959-2977.

[21] LU Y, DUEL-HALLEN A. A sensing contribution-based two-layer game for channel selection and spectrum access in cognitive radio ad-hoc networks[J]. IEEE Transactions on Wireless Communications, 2018, 17(6): 3631-3640.

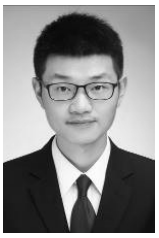
[22] MAKHDOMI A A, BEGH G R. Energy efficient distributed spectrum sensing in presence of malicious users[J]. IEEE Networking Letters, 2022, 4(2): 64-67.

[23] WANG Y T, SU Z, BENSLIMANE A, et al. Collaborative honeypot defense in UAV networks: a learning-based game approach[J]. IEEE Transactions on Information Forensics and Security, 2023, PP(99): 1.

[24] PATNAIK M, PRABHU G, REBEIRO C, et al. ProBLeSS: a proactive blockchain based spectrum sharing protocol against SSDF attacks in

cognitive radio IoBT networks[J]. IEEE Networking Letters, 2020, 2(2): 67-70.
[25] JIN X C, ZHANG Y C. Privacy-preserving crowdsourced spectrum sensing[J]. IEEE/ACM Transactions on Networking, 2018, 26(3): 1236-1249.
[26] SCHWARTZ M. Mobile wireless communications[M]. Cambridge: Cambridge University Press, 2006.
[27] ALGANS A, PEDERSEN K I, MOGENSEN P E. Experimental analysis of the joint statistical properties of azimuth spread, delay spread, and shadow fading[J]. IEEE Journal on Selected Areas in Communications, 2002, 20(3): 523-531.
[28] SMADI M A, PRABHU V K. Performance analysis of generalized-faded coherent PSK channels with equal-gain combining and carrier phase error[J]. IEEE Transactions on Wireless Communications, 2006, 5(3): 509-513.
[29] AWIN F, ABDEL-RAHEEM E, AHMADI M. Joint optimal transmission power and sensing time for energy efficient spectrum sensing in cognitive radio system[J]. IEEE Sensors Journal, 2017, 17(2): 369-376.
[30] WANG Y T, SU Z, LUAN T H, et al. SEAL: a strategy-proof and privacy-preserving UAV computation offloading framework[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 5213-5228.
[31] WANG Y T, SU Z, ZHANG N, et al. SPDS: a secure and auditable private data sharing scheme for smart grid based on blockchain[J]. IEEE Transactions on Industrial Informatics, 2021, 17(11): 7688-7699.
[32] YANG J W, DAI J H, GOOI H B, et al. A proof-of-authority blockchain-based distributed control system for islanded microgrids[J]. IEEE Transactions on Industrial Informatics, 2022, 18(11): 8287-8297.
[33] BONEH D, DRIJVERS M, NEVEN G. Compact multi-signatures for smaller blockchains[C]//Proceedings of International Conference on the Theory and Application of Cryptology and Information Security. Berlin: Springer, 2018: 435-464.
[34] WATSON J. Strategy: an introduction to game theory[M]. New York: W. W. Norton & Company, 2002.
[35] BECKER G S. Crime and punishment: an economic approach[J]. Journal of Political Economy, 1968, 76(2): 169-217.
[36] WANG Y T, SU Z, XU Q C, et al. A secure and intelligent data sharing scheme for UAV-assisted disaster rescue[J]. IEEE/ACM Transactions on Networking, 2023, 31(6): 2422-2438.
[37] BENADDI H, IBRAHIMI K, BENSLIMANE A, et al. Robust enhancement of intrusion detection systems using deep reinforcement learning and stochastic game[J]. IEEE Transactions on Vehicular Technology, 2022, 71(10): 11089-11102.
[38] BOWLING M, VELOSO M. Rational and convergent learning in stochastic games[C]//Proceedings of the 17th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2001: 1021-1026.
[39] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine Learning, 1992, 8(3): 279-292.
[40] ZOU Q Y, ZHENG S F, SAYED A H. Cooperative sensing via sequential detection[J]. IEEE Transactions on Signal Processing, 2010, 58(12): 6266-6283.

[作者简介]



王云涛（1995− ），男，江苏南京人，博士，西安交通大学助理教授，主要研究方向为空天地一体化安全、智能博弈、区块链等。

苏洲（1973− ），男，陕西西安人，博士，西安交通大学教授、博士生导师，主要研究方向为无线网络、移动网络、网络空间安全等。

许其超（1989− ），男，浙江杭州人，博士，上海大学副教授，主要研究方向为无线网络架构与安全防护等。

刘怡良（1990− ），男，江苏徐州人，博士，西安交通大学助理教授，主要研究方向为物理层安全、无线通信安全等。

彭海霞（1988− ），女，湖南郴州人，博士，西安交通大学教授、博士生导师，主要研究方向为智能车联网、人工智能、空天地一体化等。

栾浩（1982− ），男，陕西西安人，博士，西安交通大学教授、博士生导师，主要研究方向为无线网络、车联网、数字孪生网络等。