海报：通过基于代理的分析增强 GNN 在网络安全入侵检测中的鲁棒性

Zhonghao Zhan
Department of Computing
Imperial College London

Huichi Zhou
Department of Computing
Imperial College London

Hamed Haddadi
Department of Computing
Imperial College London

摘要—图神经网络（GNNs）在网络入侵检测系统（NIDS）中，尤其是在物联网环境中显示出巨大的潜力，但由于分布漂移导致性能下降，并且缺乏对抗现实世界攻击的鲁棒性。当前的鲁棒性评估通常依赖于不切实际的人工扰动，并且缺乏对不同类型的攻击进行系统的分析展示，这些攻击包括黑盒和白盒场景。本文提出了一种新颖的方法，通过在代理管道中使用大型语言模型（LLMs）作为模拟网络安全专家代理来增强 GNN 的鲁棒性和泛化能力。这些代理仔细审查从网络流数据中得出的图结构，在 GNN 处理之前识别并可能缓解可疑或受到攻击干扰的元素。我们的实验使用了一个设计用于现实评估和测试的框架，包括各种类型的对抗性攻击以及来自物理测试平台实验收集的数据集，展示了集成 LLM 分析可以显著提高基于 GNN 的 NIDS 的韧性，突显了 LLM 代理作为入侵检测架构中互补层的潜力。

***Index Terms*—图神经网络，大型语言模型，网络入侵检测，对抗攻击，鲁棒性**

I. 介绍

图神经网络（GNNs）通过建模复杂的网络关系，为网络入侵检测系统（NIDS）提供了重要的潜力。然而，实现其实用价值需要解决标准评估中经常被忽视的关键鲁棒性挑战。我们的工作解决了通过严格的初步分析揭示的两个主要问题。

首先，我们解决分布漂移下的泛化挑战。标准的 GNN 评估通常依赖于在单一静态数据集（例如 CIC-IDS 或 UNSW-NB15 的具体版本）内进行训练和测试。这无法捕捉到现实世界网络中性能表现，因为在这些网络中流量模式不断演变 [1]。为了研究这一点，我们通过合并并标准化来自各种标准来源的数据流构建了一个统一数据集，包括 UNSW-NB15 [2]、CIC-IDS2018 [3] 和 Bot-IoT [4]。评估代表性的 GNN 模型（如 EGraphSage [5]、Anomal-E [6] 和 CAGN-GAT [7]）使用这个统一的数据集显示了与它们单个数据集基准相比性能明显下降，实证确认了它们对分布漂移的敏感性，这是一个基本的鲁棒性问题。

其次，除了漂移之外，我们还考察了在实际条件下对抗攻击的鲁棒性。虽然 GNN 可能声称具有鲁棒性，但评估通常使用合成攻击，特别是白盒方法，这赋予攻击者不切实际的能力 [8]。为了更现实地探测漏洞，我们在统一数据集上模拟了实用的问题空间黑盒攻击（如节点注入）。这些实验表明，被评估的 GNN 模型出现了进一步显著的性能下降，暴露了它们即使对真实对手可行的攻击也存在脆弱性。这一发现强调了当前鲁棒性声明的不足，并突出了更现实的对抗测试场景和有效的缓解策略的必要性。

观察到 GNN 在面对漂移和实际攻击时的脆弱性，激发了我们核心提案：利用 LLM 作为模拟网络安全专家。鉴于人类专家通常能够识别出自动化系统 [9], [10] 错过的复杂攻击或异常情况，我们探索使用 LLM 代理对网络流图数据进行预防性分析。通过审查图结构和流量模式，LLM 代理旨在在威胁危及 GNN 之前过滤威胁或向操作员发出警报。这份海报展示了我们对该 LLM 代理方法的研究，详细介绍了其集成，并呈现了实验证据，证明该方法能够增强 GNN 抵御模拟实际攻击的能力，为部署更强大的 NIDS 提供了一条有前景的路径。

II. 相关工作

当前对基于 GNN 的 NIDS 的评估面临限制，妨碍了实际评估。模型通常在单一静态数据集上进行评估，忽略了由于分布漂移引起的性能下降 [1]。此外，对抗鲁棒性评估常常依赖于不考虑网络领域约束的人工生成扰动 [8]，这可能会使攻击变得不切实际 [8]。许多关

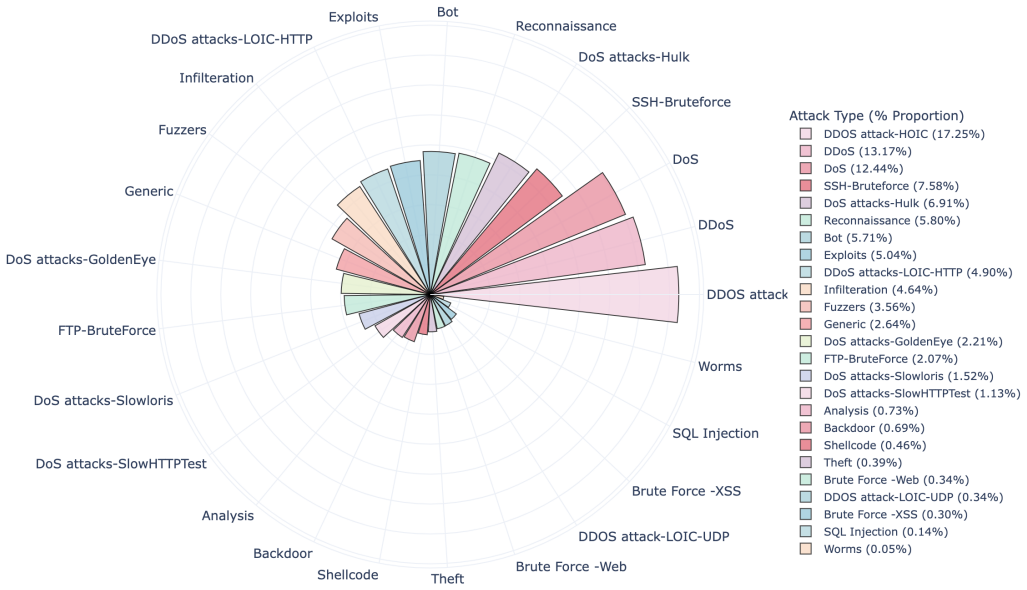

图 1: 用于分布漂移测试的统一数据集的入侵类型

于 GNN 鲁棒性的研究使用通用的图基准，而不是特定于 NIDS 的结构 [11]，并且现实中的结构性攻击仍然未被充分探索 [8]。现有的 GNN 鲁棒性基准可能也缺乏足够的漂移或领域特定真实感集成 [12]。利用 LLMs 进行鲁棒性评估或增强 GNN 的想法正在出现 [13], [14]，但将它们作为 NIDS 的专家模拟器应用仍属未开发领域。这种评估差距导致了对 GNN 韧性的高估。

表 I: 不同数据集上的 GNN 性能比较

模型	**数据集**	**准确率**	**F1**
CAGN	Unified	0.851	0.823
	UNSW-NB15	0.995*	0.918*
	CICIDS2017	0.975*	0.881*
AnomalE	Unified	0.861	0.875
	UNSW-NB15	0.987*	0.924*
	CICIDS2018	0.971*	0.924*
E-GraphSage	Unified	0.675	0.793
	NF-ToN-IoT	0.672*	0.810*
	NF-Bot-IoT	0.782*	0.630*

* 表中模型的结果报告。“统一”结果反映了本研究中使用的统一数据集上的性能。结合的异常 F1 值为加权平均值。

III. 方法论

我们的方法将 LLM 分析整合到基于 GNN 的 NIDS 管道中。

A. 数据集和 GNN 模型

我们利用了一个通过合并和标准化多个基于 NetFlow 的入侵检测数据集（NF-BoT-IoT [4], NF-CSE-CIC-IDS2018 [3], NF-UNSW-NB15 [2]）构建的统一数据集来模拟分布漂移。这使得能够评估 GNN 在各种网络场景中的泛化能力。我们评估了几种在 NIDS 中常用的 GNN 架构，包括 E-GraphSAGE [5], Anomal-E [6] 和 CAGN-GAT [7]。将网络流转换为以 IP 为中心的通信图，其中节点代表 IP 地址，边代表具有相关特征的流。

B. 大语言模型代理集成

我们设计了一种基于大语言模型的缓解策略，其中大语言模型代理在图神经网络处理之前分析网络图的元素。这包括：

- **任务定义：** 大型语言模型的任务是评估图组件的相关性或潜在恶意性，例如模拟某些攻击的注入节点或可疑边/流（概念化于图 2）。
- **提示策略：** 大型语言模型被提示输入代表图元素的文本信息（例如，节点交互描述或流特征），并要求提供分析和相关性评分。

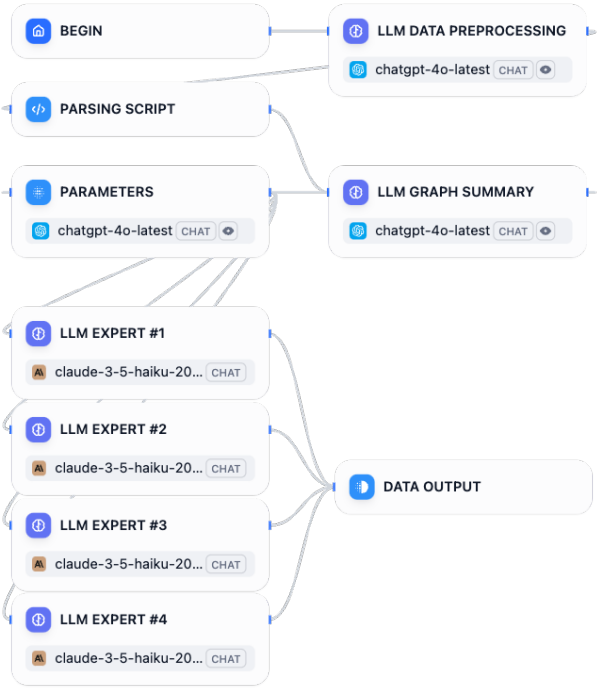

图 2: 大型语言模型缓解管道的设计图

- **工作流：**大型语言模型充当过滤器或专家顾问。根据大型语言模型的评估，可疑的图元素可以在输入到 GNN 模型之前被标记、移除或以不同的权重处理。我们特别针对节点注入攻击测试了这一点，在这种攻击中，大型语言模型试图识别被注入的恶意节点。

C. 评估协议

我们在几种条件下评估了 GNN 模型：

1) 基线性能在标准化数据集上的表现。
2) 使用统一数据集模拟分布漂移下的性能。
3) 对合成攻击（PGD 特征攻击、边缘移除、节点注入）的鲁棒性。
4) 基于大语言模型缓解措施的 GNN 性能表现，特别是针对节点注入攻击的恢复情况。

指标包括准确率、F1 值、精确率、召回率和 AUC，鉴于 NIDS 数据集中固有的类别不平衡问题，我们重点关注 F1 值。

IV. 结果与讨论

我们的实验确认了标准 GNN 模型对分布漂移和合成攻击的易感性。特别是节点注入，观察到其会降低所评估模型的准确率和 F1 分数。

本次海报强调的关键发现是 LLM 缓解策略的有效性。LLMs（特别是经过测试的模型如 GPT-4o 和 LLaMA 4）表现出强大的能力，能够识别从 netflow 数据中导出的图结构内的人工注入恶意节点，详见表 II。例如，在 20%节点注入的情景下（在包含 1000 个节点的基础图中注入了 200 个节点），领先的 LLMs 正确标记了这些恶意节点中的高百分比。

通过基于大语言模型分析过滤或识别这些恶意节点，下游图神经网络分类性能与未使用大语言模型缓解措施的攻击场景相比显著提升（表 II）。虽然在合成节点注入测试中进行了测试，但大语言模型根据提供的数据推理网络交互的能力表明了其处理新型、零日攻击或表现为异常图形结构的复杂漂移模式的潜在益处。

表 II: 大语言模型在图缓解中的表现。正在测试 CAGN-GAT 融合 [7]。CF = 正确标记，IF = 错误标记

模型	准确性	F1	大语言模型召回率	节点	CF	如果
Clean	0.842	0.821	–	1000	–	–
Claude 3.5 Haiku	0.777	0.673	0.698	1036	0	164
Claude 3.7 Sonnet	0.774	0.728	0.741	1107	35	**58**
LLaMA 4 Maverick 17B	**0.859**	0.834	**0.758**	931	**200**	69
GPT-4o	0.857	**0.838**	0.774	945	197	**58**

V. 结论

标准的 GNN 评估通常忽视了分布漂移导致的性能下降，并且使用不现实的对抗攻击。我们使用统一数据集进行的分析证实了 GNN 对漂移的敏感性，在模拟的真实攻击下进一步恶化。本研究展示了将个体 LLM 作为专家分析师集成，可以增强 GNN 抵抗如节点注入等攻击的能力，彰显了混合 GNN-LLM 系统的潜力。

未来的工作将专注于开发一个 LLM 代理管道。该管道旨在战略性地利用不同的模型：例如，使用 GPT-4o 进行复杂的高级图分析以生成指导更高效成本模型（如 Claude 3.5 Haiku）的上下文，以便进行大规模处理。目标是创建一个成本效益高的工作流，在保持强大的网络入侵和对抗攻击检测性能的同时。此外，我们计划在一个物理 IoT 测试床上使用现实世界的工具（例如 Kali Linux）并结合实际设备操作来验证这些发现，并将其应用于多样化的、真实的攻击场景中。优化整个管道的效率以实现潜在实时 NIDS 部署仍然是一个关键的最终目标。

参考文献

[1] G. Andresini, F. Pendlebury, F. Pierazzi, C. Loglisci, A. Appice, and L. Cavallaro, "Insomnia: Towards concept-drift robustness in network intrusion detection," in *Proceedings of the 14th ACM workshop on artificial intelligence and security*, 2021, pp. 111–122.

[2] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*. IEEE, 2015, pp. 1–6.

[3] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani *et al.*, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." *ICISSp*, vol. 1, no. 2018, pp. 108–116, 2018.

[4] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

[5] W. W. Lo, S. Layeghy, M. Sarhan, M. Gallagher, and M. Portmann, "E-graphsage: A graph neural network based intrusion detection system for iot," in *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2022, pp. 1–9.

[6] E. Caville, W. W. Lo, S. Layeghy, and M. Portmann, "Anomal-e: A self-supervised network intrusion detection system based on graph neural networks," *Knowledge-based systems*, vol. 258, p. 110030, 2022.

[7] M. A. Jahin, S. Soudeep, M. Mridha, R. Kabir, M. R. Islam, and Y. Watanobe, "Cagn-gat fusion: A hybrid contrastive attentive graph neural network for network intrusion detection," *arXiv preprint arXiv:2503.00961*, 2025.

[8] J. Ma, S. Ding, and Q. Mei, "Towards more practical adversarial attacks on graph neural networks," *Advances in neural information processing systems*, vol. 33, pp. 4756–4766, 2020.

[9] W. Ding, H. Hu, and L. Cheng, "Iotsafe: Enforcing safety and security policy withreal iot physical interaction discovery," in *Network and Distributed System Security Symposium*, 2021.

[10] S. Siboni, V. Sachidananda, A. Shabtai, and Y. Elovici, "Security testbed for the internet of things," *arXiv preprint arXiv:1610.05971*, 2016.

[11] C. Wang, P. Zheng, J. Gui, C. Hua, and W. U. Hassan, "Are we there yet? unraveling the state-of-the-art graph network intrusion detection systems," *arXiv preprint arXiv:2503.20281*, 2025.

[12] Q. Zheng, X. Zou, Y. Dong, Y. Cen, D. Yin, J. Xu, Y. Yang, and J. Tang, "Graph robustness benchmark: Benchmarking the adversarial robustness of graph machine learning," *arXiv preprint arXiv:2111.04314*, 2021.

[13] H. Zhou, K.-H. Lee, Z. Zhan, Y. Chen, Z. Li, Z. Wang, H. Haddadi, and E. Yilmaz, "Trustrag: Enhancing robustness and trustworthiness in rag," *arXiv preprint arXiv:2501.00879*, 2025.

[14] Z. Zhang, X. Wang, H. Zhou, Y. Yu, M. Zhang, C. Yang, and C. Shi, "Can large language models improve the adversarial robustness of graph neural networks?" *arXiv preprint arXiv:2408.08685*, 2024.