Evaluating and Improving Graph-based Explanation Methods for Multi-Agent Coordination

Extended Abstract

Siva Kailas
Georgia Institute of Technology
Atlanta, United States of America
skailas3@gatech.edu

Shalin Jain
Georgia Institute of Technology
Atlanta, United States of America
sjain441@gatech.edu

Harish Ravichandar
Georgia Institute of Technology
Atlanta, United States of America
harish.ravichandar@cc.gatech.edu

ABSTRACT

Graph Neural Networks (GNNs), developed by the graph learning community, have been adopted and shown to be highly effective in multi-robot and multi-agent learning. Inspired by this successful cross-pollination, we investigate and characterize the suitability of existing GNN explanation methods for explaining multi-agent coordination. We find that these methods have the potential to identify the most-influential communication channels that impact the team's behavior. Informed by our initial analyses, we propose an attention entropy regularization term that renders GAT-based policies more amenable to existing graph-based explainers. Intuitively, minimizing attention entropy incentivizes agents to limit their attention to the most influential or impactful agents, thereby easing the challenge faced by the explainer. We theoretically ground this intuition by showing that minimizing attention entropy increases the disparity between the explainer-generated subgraph and its complement. Evaluations across three tasks and three team sizes i) provides insights into the effectiveness of existing explainers, and ii) demonstrates that our proposed regularization consistently improves explanation quality without sacrificing task performance.

KEYWORDS

Explainability, Multi-Agent Learning, Graph-based Coordination

ACM Reference Format:
Siva Kailas, Shalin Jain, and Harish Ravichandar. 2025. Evaluating and Improving Graph-based Explanation Methods for Multi-Agent Coordination: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Graph neural networks (GNNs) were originally developed to analyze complex relational data [15]. However, they were quickly adopted by various other communities due to their ability to capture structural information and reason over non-euclidean spaces while remaining invariant to certain distractors. The fields of multi-agent and multi-robot learning were among the beneficiaries of these powerful techniques, enabling scalable policies that encode team size-invariant strategies for inter-robot communication and coordination. This has instantiated into adoption of RL trained GNN-based policies with overall similar design choices in the multi-robot community to tackle practical applications such as cooperative navigation [11, 12], coverage control [9], autonomous driving [7], and real-world multi-robot coordination [5].

In this work, we explore whether one could adopt existing post-hoc (agnostic) GNN explanation methods to explain multi-agent coordination. We study GNN explainers since they estimate parsimonious yet representative subgraphs as a means to explain complex decisions. We systematically investigate and characterize the suitability of existing GNN explanation methods for graph attention network (GAT) based policies in multi-agent coordination. If we could explain GNN-based coordination policies via salient subgraphs, then users could identify the most influential inter-agent interactions that can effectively approximate and distill coordination strategies learned across the entire team. This could be useful to effectively debug the learning algorithm by comparing observed coordination strategies against their expectations. Further, explanations would help non-experts gain insights into learned coordination policies. In fact, identifying such influential interactions was found to be a key challenge in explaining coordination strategies by a recent user study focused on multi-agent navigation [6].

2 EVALUATING GRAPH EXPLAINERS

We investigate the following state-of-the-art post-hoc GNN explainers in terms of their ability to explain GNN-based coordination policies: **Graph Mask** [14], **GNN-Explainer** [16], **Attention Explainer** [8]. We evaluate the three graph-based explainers introduced above on three multi-agent coordination tasks from BenchMARL [4] implemented using the VMAS simulator [3] across three team sizes. We force the agents to be blind, so agents cannot sense one another and require effective communication to solve each task. **Blind Navigation:** a team of agents must cooperatively navigate from assigned start locations to goal locations without colliding. **Blind Passage:** a team of agents starts on one side of a wall and needs to reach a destination on the other side after traversing a narrow corridor/passage while minimizing collisions. **Blind Discovery:** a team of agents must explore the environment to discover a single landmark and converge on its position. We consider four metrics to quantify both the fidelity [2] and faithfulness [1] of generated explanations. **Positive Fidelity** (↑) at timestep t is $Fid^t_+ \triangleq |F(G^t) - F(G \setminus G^t_S)|$, and measures the *necessity* of the explanation subgraph. **Negative Fidelity** (↓) at timestep t is $Fid^t_- \triangleq |F(G^t) - F(G^t_S)|$, and measures the *sufficiency* of the explanation subgraph. **Delta Fidelity** (↑) at timestep t is defined as $Fid_\Delta \triangleq Fid_+ - Fid_-$ to measure the explanation's

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

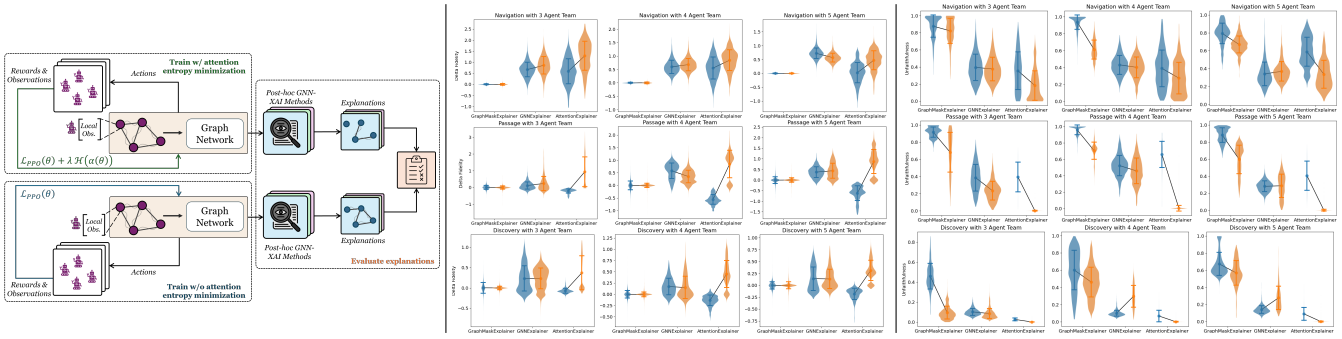

Figure 1: (Left) We systematically evaluate existing graph-based explainers when used to explain multi-agent coordination, and propose a regularizer to improve the explainability of GNN-based policies. (Middle) Delta Fidelity (↑, section 4) and (Right) Unfaithfulness (↓, section 4) of explanations generated by three explainers without (blue) and with (orange) proposed regularization across three tasks (rows) and three team sizes (columns). See [10] for the full-resolution and additional figures.

joint necessity and sufficiency. **Unfaithfulness** (↓) at timestep t is $GEF^t \triangleq 1 - \exp\left(-KL(F(G^t)||F(G^t_S))\right)$, where $KL(\cdot||\cdot)$ is the Kullback–Leibler divergence.

3 IMPROVING EXPLANATIONS

After analyzing the quality of explanations generated by existing explanation results (see Sec. 4), we observed that they could be improved by modifying how we train GNNs – by minimizing the entropy of the attention values. Attention entropy minimization can be motivated from two perspectives. First, from a multi-agent coordination perspective, an agent who collaborates with other agents will intuitively desire to (1) filter out useless information and (2) focus on the information that is most crucial to the task at hand, similar to selective attention in humans [13]. Attention in GNNs serves a similar purpose when the nodes are agents and the edges are communication channels. From a graph learning perspective, minimizing attention entropy in conjunction with task objective can be seen as a form of denoising for node-level learning tasks. By incentivizing a set of attention values that have lower entropy, the model is likely to learn a stronger filter that starts removing extraneous or noisy information. The intuition behind minimizing attention entropy can be connected to more formal notions of information bottleneck over graphs [18]. But, optimizing such objectives tends to be computationally intensive and challenging, which will likely be exacerbated when combined with multi-agent learning. In contrast, integrating attention entropy minimization into learning GNNs is much simpler, especially within MARL frameworks like MAPPO [17] by defining the new regularized loss as $\mathcal{L}_t^{PPO+ATTN} = \mathbb{E}_t[\mathcal{L}_t^{PPO}(\theta) + \lambda\mathcal{H}(\alpha_t(\theta))]$.

4 EMPIRICAL RESULTS AND DISCUSSION

Without Regularization: We observe that GraphMask tends to have the lowest explanation quality (lowest fidelity and highest unfaithfulness) likely due to its hard binary masks constraining the space of possible subgraph explanations. Consequently, GraphMask struggles to capture more diffuse inter-agent influences, which are more likely without attention regularization. In contrast, both GNNExplainer and AttentionExplainer employ soft edge masks and thus tend to produce subgraph explanations that are more expressive than those of GraphMask, especially when the attention values are diffuse. Unlike AttentionExplainer, GNNExplainer is compatible with any GNN (i.e. GCN, GAT, etc) and does not rely on the model being self-interpretable. As a result, GNNExplainer provides the best explanations when employed out-of-the-box without attention regularization.

Impact of Regularization: We observe that the attention regularization has had no discernible impact on delta fidelity for GraphMask, but consistently reduces GraphMask's unfaithfulness. This is likely due to the regularization making the attention distribution sparser, resulting in explanations that contain a larger subset of the salient edges but do not capture all the salient edges due to the restriction of hard binary masks. As such, GraphMask can capture more of the salient edges even when employing a hard mask. This yields a gain in negative fidelity and faithfulness, but also incurs a deterioration in positive fidelity. However, despite this boost from regularization, GNNExplainer and AttentionExplainer continue to perform better than GraphMask after the regularization.

Notably, attention regularization helps AttentionExplainer consistently outperform GNNExplainer across all team sizes and tasks. This is likely because this regularization boosts the correlation between the attention values (which can be interpreted as a subgraph) and the GNN model behavior (i.e., the inter-agent influences considered by the model), in line with the insights from our theoretical analysis. In contrast, entropy minimization likely obtains mixed results for GNNExplainer since the original objective of GNNExplainer is intractable to solve and requires assumptions (originally tailored for graph datasets [16]) that might be compatible with multi-agent coordination. Thus, the inclusion of attention entropy minimization may not inherently improve the optimization landscape that GNNExplainer attempts to solve despite the fact that the subgraphs induced by the attention values themselves better represent the underlying model.

Impact on Task Performance: We also measured the impact of attention entropy minimization on task performance. Reassuringly, we found little to no degradation in task performance [10].

ACKNOWLEDGMENTS

This work was supported in part by the Army Research Lab under Grant W911NF-17-2-0181 (DCIST CRA).

REFERENCES

[1] Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. 2023. Evaluating explainability for graph neural networks. *Scientific Data* 10, 1 (2023), 144.

[2] Kenza Amara, Zhitao Ying, Zitao Zhang, Zhichao Han, Yang Zhao, Yinan Shan, Ulrik Brandes, Sebastian Schemm, and Ce Zhang. 2022. GraphFramEx: Towards Systematic Evaluation of Explainability Methods for Graph Neural Networks. In *Learning on Graphs Conference*. PMLR, 44–1.

[3] Matteo Bettini, Ryan Kortvelesy, Jan Blumenkamp, and Amanda Prorok. 2022. Vmas: A vectorized multi-agent simulator for collective robot learning. In *International Symposium on Distributed Autonomous Robotic Systems*. Springer, 42–56.

[4] Matteo Bettini, Amanda Prorok, and Vincent Moens. 2024. BenchMARL: Benchmarking Multi-Agent Reinforcement Learning. *Journal of Machine Learning Research* 25, 217 (2024), 1–10. http://jmlr.org/papers/v25/23-1612.html

[5] Jan Blumenkamp, Steven Morad, Jennifer Gielis, Qingbiao Li, and Amanda Prorok. 2022. A Framework for Real-World Multi-Robot Systems Running Decentralized GNN-Based Policies. In *2022 International Conference on Robotics and Automation (ICRA)*. 8772–8778. https://doi.org/10.1109/ICRA46639.2022.9811744

[6] Martim Brandao, Masoumeh Mansouri, Areeb Mohammed, Paul Luff, and Amanda Jane Coles. 2022. Explainability in Multi-Agent Path/Motion Planning: User-study-driven Taxonomy and Requirements.. In *AAMAS*. 172–180.

[7] Peide Cai, Hengli Wang, Yuxiang Sun, and Ming Liu. 2022. DQ-GAT: Towards safe and efficient autonomous driving with deep Q-learning and graph attention networks. *IEEE Transactions on Intelligent Transportation Systems* 23, 11 (2022), 21102–21112.

[8] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

[9] Walker Gosrich, Siddharth Mayya, Rebecca Li, James Paulos, Mark Yim, Alejandro Ribeiro, and Vijay Kumar. 2022. Coverage control in multi-robot systems via graph neural networks. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 8787–8793.

[10] Siva Kailas, Shalin Jain, and Harish Ravichandar. 2025. Evaluating and Improving Graph-based Explanation Methods for Multi-Agent Coordination. arXiv:2502.09889 [cs.MA] https://arxiv.org/abs/2502.09889

[11] Qingbiao Li, Fernando Gama, Alejandro Ribeiro, and Amanda Prorok. 2020. Graph neural networks for decentralized multi-robot path planning. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 11785–11792.

[12] Qingbiao Li, Weizhe Lin, Zhe Liu, and Amanda Prorok. 2021. Message-aware graph attention networks for large-scale multi-robot path planning. *IEEE Robotics and Automation Letters* 6, 3 (2021), 5533–5540.

[13] Denise Moerel, Anina N Rich, and Alexandra Woolgar. 2024. Selective attention and decision-making have separable neural bases in space and time. *Journal of Neuroscience* 44, 38 (2024).

[14] Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. 2021. Interpreting Graph Neural Networks for {NLP} With Differentiable Edge Masking. In *International Conference on Learning Representations*. https://openreview.net/forum?id=WznmQa42ZAx

[15] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32, 1 (2020), 4–24.

[16] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems* 32 (2019).

[17] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems* 35 (2022), 24611–24624.

[18] Junchi Yu, Tingyang Xu, Yu Rong, Yatao Bian, Junzhou Huang, and Ran He. [n.d.]. Graph Information Bottleneck for Subgraph Recognition. In *International Conference on Learning Representations*.