Reconsidering Faithfulness in Regular, Self-Explainable, and Domain Invariant GNNs

ICLR2025

Steve Azzolin, Antonio Longa, Stefano Teso, Andrea Passerini
University of Trento

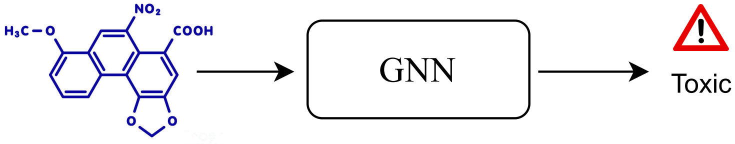
📬 steve.azzolin@unitn.it

Can we trust GNNs (and their Explanations)?

- 🏋 Graph Neural Networks (GNNs) rock at learning on graphs
- 😥 GNNs are black-box

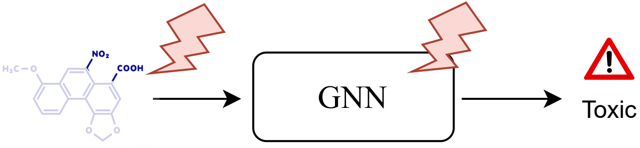
Can we trust GNNs (and their Explanations)?

- Graph Neural Networks (GNNs) rock at learning on graphs
- GNNs are black-box



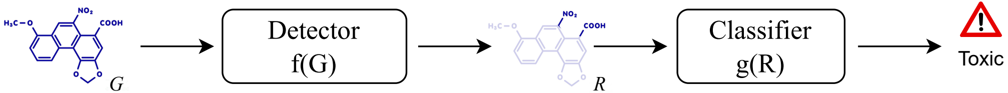
Can we trust GNNs (and their Explanations)?

Post-hoc tools explain any given GNN (Amara et al., 2022; Kakkad et al., 2023; Longa et al., 2024)



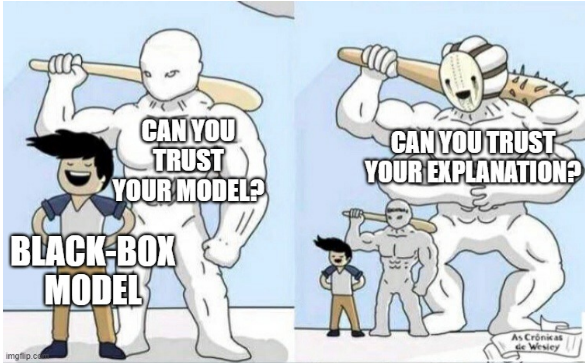
Can we trust GNNs (and their Explanations)?

Ante-hoc models provide an explanation together with the prediction (Miao et al., 2022; Kakkad et al., 2023; Longa et al., 2024)



Can we trust GNNs (and their Explanations)?

How accurately does the explanation reflect the *reasoning* of the model?



Can we trust GNNs (and their Explanations)?

How accurately does the explanation reflect the *reasoning* of the model?

👉 How **Faithful** is the explanation?

Can we trust GNNs (and their Explanations)?

How accurately does the explanation reflect the *reasoning* of the model?

👉🏻 How **Faithful** is the explanation?

Several different faithfulness metrics exist

🤔 which one to choose?

😨 what is faithfulness?

Contributions

We propose to **reconsider faithfulness** from the following angles:

- ⚙️ How to compute faithfulness
- 🎯 How to promote faithfulness
- 📈 How does faithfulness affect model generalization

How to compute faithfulness

How to compute faithfulness

We analyzed seven previous faithfulness metric and found that:

😨 Metric are not interchangeable[1] (Prop. 1)

😨 Some metrics do not encode the desired semantics (Prop. 2)

[1] Different metric yield different results

How to compute faithfulness

We analyzed seven previous faithfulness metric and found that:

😨 Metric are not interchangeable[1] (Prop. 1)

😨 Some metrics do not encode the desired semantics (Prop. 2)

😌 We propose a new necessity metric that penalizes overly large explanations (Prop. 3)

[1]Different metric yield different results

How to promote faithfulness

How to promote faithfulness

🧐 Explanations of ante-hoc GNNs can be unfaithful (Christiansen et al., 2023)

We identified some architectural desiderata for faithfulness which ante-hoc GNNs do not implement:

How to promote faithfulness

🧐 Explanations of ante-hoc GNNs can be unfaithful (Christiansen et al., 2023)

We identified some architectural desiderata for faithfulness which ante-hoc GNNs do not implement:

- 😎 **Content Features**: the classifier should use raw input features, as opposed to the explanation extractor's embeddings
- 😎 **Explanation Readout**: the final global readout should run only over the explanation, as opposed to the entire graph

How does faithfulness affect model generalization

How does faithfulness affect model generalization

Recent work in Invariant Learning for GNNs first **identifies an invariant subgraph**, and then **make predictions based on this subgraph only** (Chen et al., 2022; Gui et al., 2023)

The invariant subgraph plays the role of an explanation

How does faithfulness affect model generalization

😥 We show that extracting a domain invariant subgraph is not enough for a GNN to be truly domain invariant (Prop. 5)

How does faithfulness affect model generalization

😥 We show that extracting a domain invariant subgraph is not enough for a GNN to be truly domain invariant (Prop. 5)

> Unless the subgraph is also faithful, the information from the domain-dependent subgraph can still influence the prediction, thus preventing domain invariance

How does faithfulness affect model generalization

We proved that a GNN that fits well the ID data will fit well the OOD data if:

0️⃣ the explanation is in fact domain invariant

1️⃣ the explanation is faithful

Reconsidering Faithfulness in Regular, Self-Explainable, and Domain Invariant GNNs

Steve Azzolin [1] Antonio Longa [1] Stefano Teso [1] Andrea Passerini [1]

[1]University of Trento

UNIVERSITÀ DI TRENTO — Department of Information Engineering and Computer Science

ICLR

PAPER CODE

Motivation

GNNs lack interpretability, thus hindering understanding, debugging, and human trust:

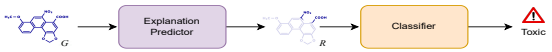
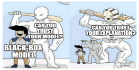


Figure 1. Pipeline of **Self-Explainable GNNs** (SEGNNs).

⚠️ How accurately does the explanation reflect the reasoning of the model?



👉 Measure the **faithfulness** of explanations, but multiple metrics exists

🤔 Which one to choose? 😨 What is faithfulness?

Our contribution 🎯

We propose to **reconsider faithfulness** from the following three angles

- How to compute faithfulness?
- How to promote faithfulness?
- How does faithfulness affect model generalization?

How to compute faithfulness?

We abstract prior metrics into:

- *sufficient*, i.e., keeping R fixed shields the model's output from changes to its complement $C = G \setminus R$

$$SUF_{d,p_R}(R) = \mathbb{E}_{G' \sim p_R}[\Delta_d(G, G')],$$

- *necessary*, i.e., altering R affects the model's output even with C fixed

$$NEC_{d,p_C}(R) = \mathbb{E}_{G' \sim p_C}[\Delta_d(G, G')]$$

Table 1. *SUF* and *NEC* recover existing faithfulness metrics for appropriate choices of divergence d and interventional distributions p_R and p_C.

Metric	Estimates	Divergence d	Allowed changes
Unf	Suf	KL$(p_\theta(\cdot \mid G), p_\theta(\cdot \mid G'))$	zero out all irrelevant features
Fid-		$\lvert p_\theta(\hat{y} \mid G) - p_\theta(\hat{y} \mid G')\rvert$	zero out all irrelevant features, delete all irrelevant edges
RFid-		"	delete a random subset of irrelevant edges
PS		$\mathbb{1}\{p_\theta(\hat{y} \mid G) = p_\theta(\hat{y} \mid G')\}$	multiply all irrelevant elements by relevance scores
Fid+	Nec	$\lvert p_\theta(\hat{y} \mid G) - p_\theta(\hat{y} \mid G')\rvert$	zero out all relevant features, delete all relevant edges
RFid+		"	delete a random subset of relevant edges
PN		$\mathbb{1}\{p_\theta(\hat{y} \mid G) \neq p_\theta(\hat{y} \mid G')\}$	multiply all relevant elements by relevance scores

How to compute faithfulness? (cont.)

Table 2. Model ranking and *SUF* results according to different p_R.

Split	Model	Motif2 $p_R^{id_1}$	Motif2 $p_R^{id_2}$
ID	LECI	1 (81 ±03)	2 (82 ±03)
	GSAT	2 (78 ±01)	1 (84 ±02)
	CIGA	3 (65 ±07)	3 (73 ±06)

🎯 Metrics are not interchangeable

🎯 Previous *Necessity* metrics **do not penalize useless explanations**

🎯 We propose a **new necessity metric** penalizing overly large explanations

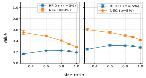


Figure 2. Our proposed **Nec** is sensitive to the number of irrelevant items in the explanation, whereas **RFid+** is not.

How to promote faithfulness?

We identified some architectural design choices favoring **un-faithfulness** and fixed them:

- **Hard Scores** (HS): give exact zero importance to information outside of R;
- **Explanation Readout** (ER): aggregate only over R for the final prediction.
- **Content Features** (CF): feed the classifier raw features, not embeddings
- **Local Aggregations** (LA): non-local aggregations can create unwanted dependencies

Table 3. Test set accuracy and faithfulness of some augmented SE-GNNs.

Dataset	BaMS Acc	BaMS Faith	Motif2 Acc	Motif2 Faith	Motif-Size Acc	Motif-Size Faith	BBBP Acc	BBBP Faith
GSAT	100 ±.00	35 ±.03	92 ±.01	61 ±.01	90 ±.01	60 ±.02	79 ±.04	27 ±.08
GSAT + ER	100 ±.00	35 ±.03	92 ±.01	63 ±.01	90 ±.01	65 ±.01	80 ±.02	33 ±.04
GSAT + HS	98 ±.01	21 ±.06	53 ±.02	24 ±.05	54 ±.03	22 ±.05	71 ±.01	31 ±.09
GSAT + ER + HS	99 ±.01	24 ±.04	57 ±.04	37 ±.03	56 ±.07	29 ±.09	73 ±.02	33 ±.02
GISST	100 ±.00	25 ±.03	92 ±.01	53 ±.02	92 ±.00	50 ±.02	84 ±.03	23 ±.11
GISST + ER	–	–	–	–	–	–	85 ±.06	27 ±.06
GISST + HS	–	–	–	–	–	–	83 ±.05	19 ±.07
GISST + ER + HS	–	–	–	–	–	–	81 ±.07	15 ±.09
RAGE	96 ±.01	33 ±.05	83 ±.02	64 ±.04	74 ±.09	63 ±.07	82 ±.01	33 ±.04
RAGE + ER	96 ±.02	33 ±.02	85 ±.06	66 ±.03	71 ±.09	55 ±.07	84 ±.01	33 ±.05
RAGE + HS	97 ±.01	46 ±.03	85 ±.01	65 ±.02	78 ±.07	65 ±.09	84 ±.02	46 ±.02
RAGE + ER + HS	96 ±.01	46 ±.04	83 ±.04	64 ±.04	75 ±.08	62 ±.12	82 ±.01	43 ±.03

How does faithfulness affect model generalization?

Recent work in Invariant Learning for GNNs first **identifies an invariant subgraph**, and then **make predictions based on this subgraph only** [1, 2].

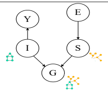


Figure 3. SCM assumptions of DI-GNNs.

⚠️ The invariant subgraph plays the role of an explanation

🎯 Unless the subgraph is also **faithful**, the information from the domain-dependent subgraph can still influence the prediction, thus **preventing domain invariance**

Theorem 1. *Let p_θ be a deterministic DI-GNN with detector f and classifier g, and $p^{id}(G, Y)$ and $p^{ood}(G, Y)$ be the ID and OOD empirical distributions, respectively. Then:*

$$\left|\mathbb{E}_{(G,y)\sim p^{id}}[p_\theta(y \mid G)] - \mathbb{E}_{(G,y)\sim p^{ood}}[p_\theta(y \mid G)]\right| \quad (1)$$
$$\leq \mathbb{E}\left[k_1(\lambda^{id}_{\text{topo}} + \lambda^{ood}_{\text{topo}}) + k_2(\lambda^{id}_{\text{feat}} + \lambda^{ood}_{\text{feat}}) + (\lambda^{id}_{\text{suff}} + \lambda^{ood}_{\text{suff}})\right]$$

🎯 A DI-GNN that fits the ID data well will fit the OOD data well if:

- R is domain-invariant (low λ_{topo} and λ_{feat})
- highly sufficient (low λ_{suff})

🎯 **Correlate** the difference in average prediction's **likelihood** between ID and OOD data, and the sum of the **degree of domain-invariance** and **faithfulness**

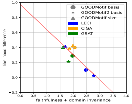


Figure 4. Likelihood, faithfulness, and domain-invariance are correlated.

References

[1] Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148, 2022.

[2] Shurui Gui, Meng Liu, Xiner Li, Youzhi Luo, and Shuiwang Ji. Joint learning of label and environment causal independence for graph out-of-distribution generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

steve.azzolin@unitn.it

References

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.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. 2022. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems* 35 (2022), 22131–22148.

Marc Christiansen, Lea Villadsen, Zhiqiang Zhong, Stefano Teso, and Davide Mottin. 2023. How Faithful are Self-Explainable GNNs? *arXiv preprint arXiv:2308.15096* (2023).

Shurui Gui, Meng Liu, Xiner Li, Youzhi Luo, and Shuiwang Ji. 2023. Joint Learning of Label and Environment Causal Independence for Graph Out-of-Distribution Generalization. *arXiv preprint arXiv:2306.01103* (2023).

Jaykumar Kakkad, Jaspal Jannu, Kartik Sharma, Charu Aggarwal, and Sourav Medya. 2023. A Survey on Explainability of Graph Neural Networks. *arXiv preprint arXiv:2306.01958* (2023).

Antonio Longa, Steve Azzolin, Gabriele Santin, Giulia Cencetti, Pietro Lio, Bruno Lepri, and Andrea Passerini. 2024. Explaining the Explainers in Graph Neural Networks: a Comparative Study. *ACM Comput. Surv.* (2024). `https://doi.org/10.1145/3696444`

Siqi Miao, Mia Liu, and Pan Li. 2022. Interpretable and generalizable graph learning via stochastic attention mechanism. In *International Conference on Machine Learning*. PMLR, 15524–15543.