GNN Explanations that do not Explain and How to find Them

Steve Azzolin[†], Stefano Teso[†], Bruno Lepri[‡], Andrea Passerini[†], Sagar Malhotra[§]

[†] *University of Trento*
[‡] *Bruno Kessler Foundation*
[§] *TU Wien*

ICLR 2026

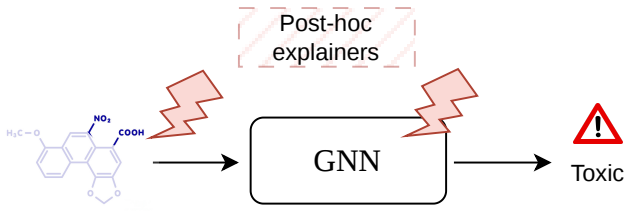
Background

Background

😢 GNNs are black-boxes

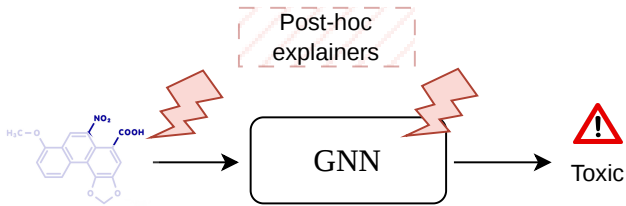
Background

😢 GNNs are black-boxes



Background

- 😢 GNNs are black-boxes
- 😢 Post-hoc explainers can be unreliable[1]



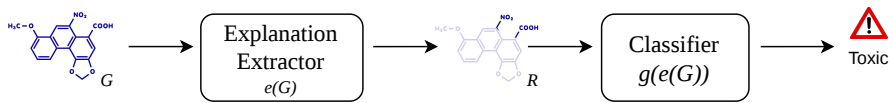
[1] C. Rudin. Stop explaining black box machine learning models and use interpretable models instead. Nature Machine Intelligence. 2019

Background: From post-hoc to ante-hoc

🧐 Design GNNs with interpretability in mind

Background: From post-hoc to ante-hoc

- 🧐 Design GNNs with interpretability in mind
- 🧐 **Self-explainable GNNs** (SE-GNNs) generate explanations during inference



Background: From post-hoc to ante-hoc

🤔 How *good* are SE-GNNs' explanations?

Background: From post-hoc to ante-hoc

😨 SE-GNN explanations can be:

- redundant[2]

[2] Tai et al., Redundancy Undermines the Trustworthiness of Self-Interpretable GNNs. ICML 2025
[3] Azzolin et al., Beyond Topological SE-GNNs: A Formal Explainability Perspective. ICML 2025
[4] Wu et al., Discovering Invariant Rationales for Graph Neural Networks. ICLR 2022

Background: From post-hoc to ante-hoc

😨 SE-GNN explanations can be:

- redundant[2]
- ambiguous[3]

[2] Tai et al., Redundancy Undermines the Trustworthiness of Self-Interpretable GNNs. ICML 2025
[3] Azzolin et al., Beyond Topological SE-GNNs: A Formal Explainability Perspective. ICML 2025
[4] Wu et al., Discovering Invariant Rationales for Graph Neural Networks. ICLR 2022

Background: From post-hoc to ante-hoc

😨 SE-GNN explanations can be:

- redundant[2]
- ambiguous[3]
- can be affected by spurious correlations[4]

[2] Tai et al., Redundancy Undermines the Trustworthiness of Self-Interpretable GNNs. ICML 2025
[3] Azzolin et al., Beyond Topological SE-GNNs: A Formal Explainability Perspective. ICML 2025
[4] Wu et al., Discovering Invariant Rationales for Graph Neural Networks. ICLR 2022

Background: From post-hoc to ante-hoc

😨 SE-GNN explanations can be:

- redundant[2]
- ambiguous[3]
- can be affected by spurious correlations[4]

🤔 Are there some pathological failure cases we should be aware of?

Present work: Yes, there are!

[2] Tai et al., Redundancy Undermines the Trustworthiness of Self-Interpretable GNNs. ICML 2025
[3] Azzolin et al., Beyond Topological SE-GNNs: A Formal Explainability Perspective. ICML 2025
[4] Wu et al., Discovering Invariant Rationales for Graph Neural Networks. ICLR 2022

Our contribution

🤒 SE-GNNs can attain optimal loss by providing **degenerate explanations**

- *explanations that do not tell anything about the underlying model*

Our contribution

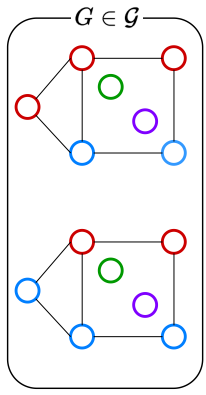
- SE-GNNs can attain optimal loss by providing **degenerate explanations**
  - *explanations that do not tell anything about the underlying model*
- Degenerate explanations can go undetected
  - *faithfulness metrics may not mark them as unfaithful*

SE-GNNs Explanations that do not Explain

Task

$y = \mathbb{1}\{\#blue > \#red\}$

and
are uncorrelated with Y

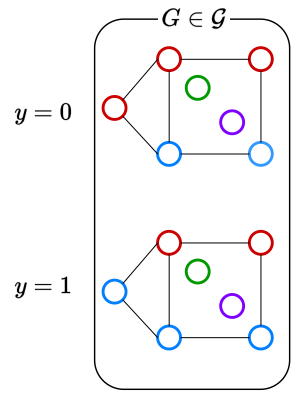


SE-GNNs Explanations that do not Explain

Task

$y = \mathbb{1}\{\#blue > \#red\}$

and
are uncorrelated with Y



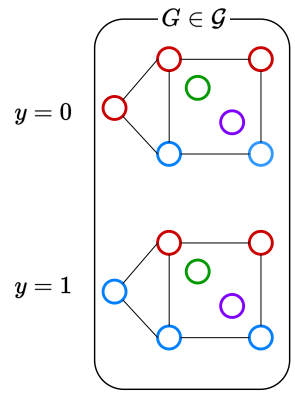
😌 Accurate GNNs must *use* and

SE-GNNs Explanations that do not Explain

Task

$y = \mathbb{1}\{\#blue > \#red\}$

and
are uncorrelated with Y



Accurate GNNs must *use* and

Accurate SE-GNNs can provide only and as explanations

This is what we call **degenerate explanations**.

Finding degenerate explanations

❓ Can we detect when an SE-GNN produces degenerate explanations?

Finding degenerate explanations

❓ Can we detect when an SE-GNN produces degenerate explanations?

Type	**Perturbations** $\mathcal{I}$	**Metrics**
Nec.	Explanation removal	Fid+, PN
Nec.	Edge removal	RFid+, Nec
Suff.	Complement removal	Fid-, PS, GEF
Suff.	Edge removal	RFid-, SimOAR
Suff.	Complement swap	Suf
Suff.	All	EST (**Ours**)

Table: Faithfulness metrics.

Finding degenerate explanations

❓ Can we detect when an SE-GNN produces degenerate explanations?

Dataset	Model	RejRatio$_{\mathcal{I}}$							
		Fid-	Fid+	Suf	Nec	CF	RFid-	RFid+	EST (ours)
RBGV	SMGNN	12 ± 19	20 ± 12	00 ± 00	-	05 ± 01	00 ± 00	-	**48** ± 02
	GSAT	12 ± 21	15 ± 16	03 ± 03	-	**50** ± 04	11 ± 09	-	49 ± 03
	DIR	32 ± 23	27 ± 24	03 ± 03	-	39 ± 08	08 ± 06	-	**48** ± 03
MNISTsp	SMGNN	88 ± 03	58 ± 04	99 ± 01	55 ± 05	92 ± 01	99 ± 00	75 ± 05	**100** ± 00
	GSAT	65 ± 09	38 ± 05	99 ± 01	44 ± 05	95 ± 02	99 ± 01	61 ± 04	**100** ± 00
	DIR	93 ± 03	55 ± 07	99 ± 01	54 ± 05	91 ± 01	99 ± 01	69 ± 07	**100** ± 00
MUTAG	SMGNN	59 ± 03	05 ± 04	**99** ± 00	57 ± 04	55 ± 07	72 ± 03	65 ± 04	96 ± 01
	GSAT	21 ± 21	05 ± 02	**99** ± 00	53 ± 07	68 ± 17	70 ± 14	61 ± 05	94 ± 05
	DIR	70 ± 13	04 ± 02	**99** ± 00	54 ± 02	69 ± 08	75 ± 07	62 ± 04	97 ± 02
SST2P	SMGNN	08 ± 02	44 ± 05	14 ± 05	-	23 ± 07	04 ± 01	-	**54** ± 04
	GSAT	51 ± 31	19 ± 14	07 ± 10	-	37 ± 18	14 ± 03	-	**62** ± 15
	DIR	**50** ± 02	09 ± 06	12 ± 06	-	42 ± 10	05 ± 01	-	49 ± 03

Table: Previous metrics can fail to reject degenerate explanations.

Conclusions

- Check out our paper for more details
- For any question: steve.azzolin@unitn.it