GNN Explanations that do not Explain and How to Find Them

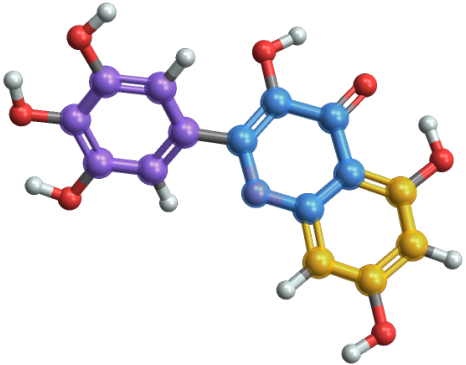
Steve Azzolin[†], Stefano Teso[†], Bruno Lepri[‡], Andrea Passerini[†], Sagar Malhotra[§]

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

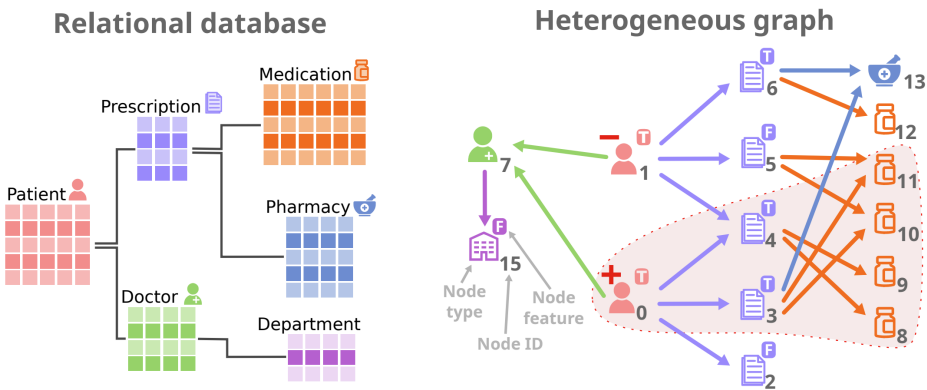
Explainable AI Seminars @ Imperial
May 2026

Background

Background - Graphs are Everywhere

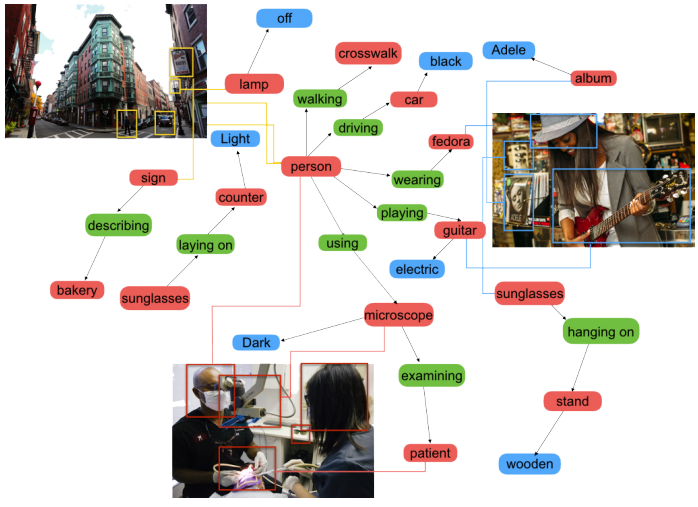


Background - Graphs are Everywhere



From *A Self-Explainable Heterogeneous GNN for Relational Deep Learning*. Ferrini et al., 2025

Background - Graphs are Everywhere



From https://homes.cs.washington.edu/~ranjay/visualgenome/index.html

Background

- Graph Neural Networks (GNNs) are a flexible tool for learning over graphs

Background

🤯 Graph Neural Networks (GNNs) are a flexible tool for learning over graphs

$G = (V, E, X)$

$V = \{u_0 \dots u_n\}$ is the set of nodes

$E \subseteq V \times V$ is the set of edges

$X \in \mathbb{R}^{n \times d}$ are the node features

Background

🤯 Graph Neural Networks (GNNs) are a flexible tool for learning over graphs

$G = (V, E, X)$

$V = \{u_0 \dots u_n\}$ is the set of nodes

$E \subseteq V \times V$ is the set of edges

$X \in \mathbb{R}^{n \times d}$ are the node features

$$h_u^0 = x_u$$
$$h_u^l = \texttt{update}^l(h_u^{l-1}, \texttt{aggr}^l(\{\{h_v^{l-1} : v \in N(u)\}\}))$$

Background

🤯 Graph Neural Networks (GNNs) are a flexible tool for learning over graphs

$$h_u^0 = x_u$$
$$h_u^l = \texttt{update}^l(h_u^{l-1}, \texttt{aggr}^l(\{\{h_v^{l-1} : v \in N(u)\}\}))$$

- `aggr` is a permutation-invariant function, like mean or sum
- `update` is a permutation-equivariant function, like a Multi-Layer Perceptron
- (Many) variants exist [1]

[1] `https://pytorch-geometric.readthedocs.io/en/latest/modules/nn.html`

Background

🤯 Graph Neural Networks (GNNs) are a flexible tool for learning over graphs

$$h_u^0 = x_u$$
$$h_u^l = \texttt{update}^l(h_u^{l-1}, \texttt{aggr}^l(\{\{h_v^{l-1} : v \in N(u)\}\}))$$

Node classification

$$y_u = \texttt{MLP}(h_u^L)$$

Graph classification

$$y = \texttt{MLP}(\texttt{aggr}(\{\{h_v^L : v \in V\}\}))$$

Edge classification

$$y_{u,v} = \texttt{MLP}(h_u^L, h_v^L)$$

Background

🤯 Graph Neural Networks (GNNs) are a flexible tool for learning over graphs

$$h_u^0 = x_u$$
$$h_u^l = \texttt{update}^l(h_u^{l-1}, \texttt{aggr}^l(\{\{h_v^{l-1} : v \in N(u)\}\}))$$

🤔 We focus on GNNs, but our insights carry to Graph Transformers and GFM

Why Explainability

Why Explainability

- 😢 GNNs are black-boxes

Why Explainability

😢 GNNs are black-boxes

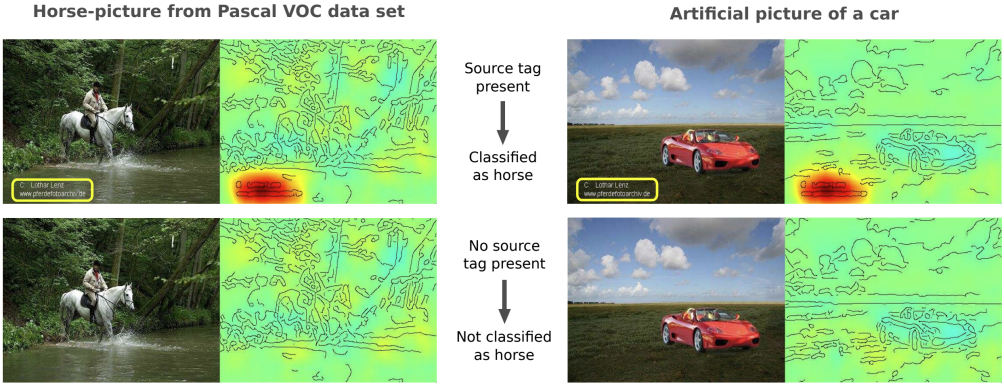
Why Explainability

😢 GNNs are black-boxes

Consequences for:

- Human trust
- Model debugging
- Regulatory compliance

Why Explainability - Examples

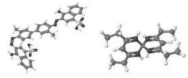


From *Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. Lapuschkin et al., 2019*

Why Explainability - Examples

2.8 AIDS Database

The AIDS data set consists of graphs representing molecular compounds. We construct graphs from the AIDS Antiviral Screen Database of Active Compounds [26]. This data set consists of two classes (*active, inactive*), which represent molecules with activity against HIV or not. The molecules are converted into graphs in a straightforward manner by representing atoms as nodes and the covalent bonds as edges. Nodes are labeled with the number of the corresponding chemical symbol and edges by the valence of the linkage. In Fig. 6 one molecular compound of both classes is illustrated. Note that different shades of grey represent different chemical symbols, i.e. node labels. We use a training set and a validation set of size 250 each, and a test set of size 1,500. Thus, there are 2,000 elements totally (1,600 inactive elements and 400 active elements). The classification result achieved on this data set is 97.3%.



(a) Active (b) Inactive

Fig. 6. A molecular compound of both classes

From *IAM Graph Database Repository for Graph Based Pattern Recognition and Machine Learning*. *Riesen & Bunke, 2008*

😰 GNNs can distinguish active from inactive by counting the number of nodes[2]

[2] *Logical Distillation of Graph Neural Networks*. Pluska et al., 2025

Post-hoc Explainability

Post-hoc Explainability

🧐 Post-hoc Explainability aims to **explain an already trained model**

Post-hoc Explainability

- 🧐 Post-hoc Explainability aims to **explain an already trained model**
- 🤔 What is an explanation?

Post-hoc Explainability

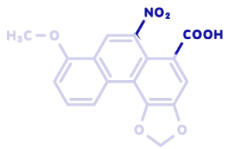
- 🧐 Post-hoc Explainability aims to **explain an already trained model**
- 🤔 What is an explanation?
- 🤔 What does it look like?

Post-hoc Explainability

- 🧐 Post-hoc Explainability aims to **explain an already trained model**
- 🤔 What is an explanation?
- 🤔 What does it look like?

Post-hoc Explainability

- 🧐 Post-hoc Explainability aims to **explain an already trained model**
- 🤔 What is an explanation?
- 🤔 What does it look like?



Post-hoc Explainability

- 🧐 Post-hoc Explainability aims to **explain an already trained model**
- 🤔 What is an explanation?
- 🤔 What does it look like?
  - Subgraph of the input
    - Small, most relevant subgraph, etc ...

Post-hoc Explainability

🧐 Post-hoc Explainability aims to **explain an already trained model**

🤔 What is an explanation?

🤔 What does it look like?

- Subgraph of the input
  - Small, most relevant subgraph, etc ...
- Logic formulas
  - *Global Explainability of GNNs via Logic Combination of Learned Concepts.* Azzolin et al., ICLR 2022
  - *Logical Distillation of Graph Neural Networks.* Pluska et al., KR 2024
  - *GraphTrail: Translating GNN Predictions into Human-Interpretable Logical Rules.* Armgaan et al., NeurIPS 2024

Post-hoc Explainability is Hard

Probing GNN Explainers: A Rigorous Theoretical and Empirical Analysis of GNN Explanation Methods

Chirag Agarwal
Harvard University

Marinka Zitnik*
Harvard University

Himabindu Lakkaraju*
Harvard University

Abstract

As Graph Neural Networks (GNNs) are increasingly being employed in critical real-world applications, several methods have been proposed in recent literature to explain the predictions of these models. they can evaluate when and how much to rely on these models, and detect potential biases or errors in them. To this end, several approaches have been proposed in recent literature to explain the predictions of GNNs (Baldassarre and Azizpour, 2019; Faber et al., 2020; Huang et al., 2020; Lucic et al., 2021; Luo et al., 2020; Pope et al., 2019; Schlichtkrull et al., 2021; Vu

Post-hoc Explainability is Hard

Explaining the Explainers in Graph Neural Networks: a Comparative Study

ANTONIO LONGA, Fondazione Bruno Kessler, Trento, Italy and DISI, University of Trento, Trento, Italy
STEVE AZZOLIN, DISI, University of Trento, Trento, Italy
GABRIELE SANTIN, Fondazione Bruno Kessler, Trento, Italy
GIULIA CENCETTI, Fondazione Bruno Kessler, Trento, Italy
PIETRO LIO, Cambridge University, Cambridge, United Kingdom of Great Britain and Northern Ireland
BRUNO LEPRI, Fondazione Bruno Kessler, Trento, Italy
ANDREA PASSERINI, DISI, University of Trento, Trento, Italy

Following a fast initial breakthrough in graph-based learning, Graph Neural Networks (GNNs) have reached a widespread application in many science and engineering fields, prompting the need for methods to understand their decision process. GNN explainers have started to emerge in recent years, with a multitude of methods both novel or adapted from other domains. To sort out this plethora of alternative approaches, several studies have benchmarked the performance of different explainers in terms of various explainability metrics. However, these earlier works make no attempts at providing insights into why different GNN

Post-hoc Explainability is Hard

Graph Neural Network Explanations are Fragile

Jiate Li [1 2] **Meng Pang** [1] **Yun Dong** [3] **Jinyuan Jia** [4] **Binghui Wang** [2 5]

Abstract

Explainable Graph Neural Network (GNN) has emerged recently to foster the trust of using GNNs. Existing GNN explainers are developed from various perspectives to enhance the explanation performance. We take the first step to study GNN explainers under adversarial attack—We found that an adversary slightly perturbing graph structure can ensure GNN model makes correct predictions, but the GNN explainer yields a drastically different explanation on the perturbed graph. Specifically, we first formulate the attack problem under a practical threat model (i.e., the adversary has

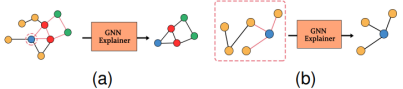


Figure 1. GNN explanation for (a) node classification and (b) graph classification—It identifies the subgraph that ensures the best prediction for the target node and target graph, respectively.

predicted (node/graph) label by a GNN model, explainable GNNs aim to determine the subgraph (include edges and the connected nodes) from the graph that ensures the best pre-

Post-hoc Explainability is Hard

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x

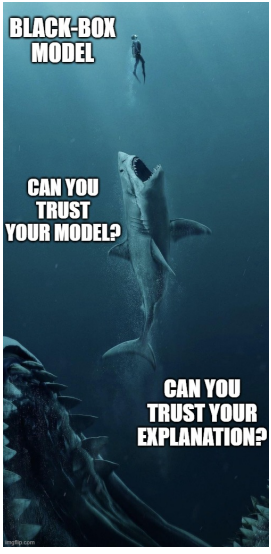
nature machine intelligence

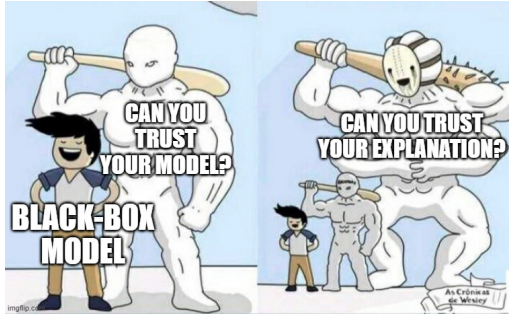
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.

Post-hoc Explainability is Hard





Ante-hoc Explainability

Ante-hoc Explainability

- 💡 Design GNNs with interpretability in mind

Ante-hoc Explainability

- 💡 Design GNNs with interpretability in mind
- Subgraph-based
  - **Our focus ...**

Ante-hoc Explainability

- 💡 Design GNNs with interpretability in mind
- Subgraph-based
  - **Our focus ...**
- Additive Models
  - *The Intelligible and Effective Graph Neural Additive Network.* Bechler-Speicher et al., NeurIPS 2024.

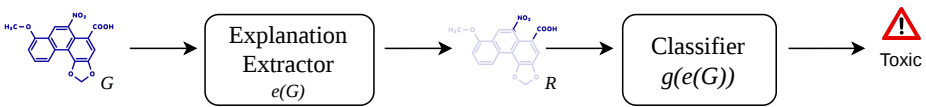
Ante-hoc Explainability

💡 Design GNNs with interpretability in mind

- Subgraph-based
  - **Our focus ...**
- Additive Models
  - *The Intelligible and Effective Graph Neural Additive Network.* Bechler-Speicher et al., NeurIPS 2024.
- Logic formulas
  - *On Logic-based Self-Explainable Graph Neural Network.* Ragno et al., NeurIPS 2025.

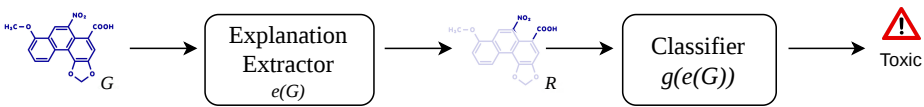
Ante-hoc Explainability

🧐 **Self-explainable GNNs** (SE-GNNs) generate subgraph explanations during inference



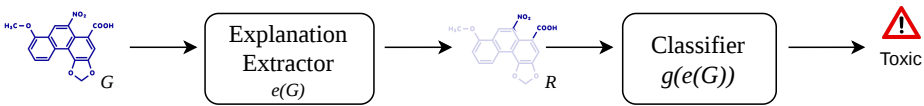
Ante-hoc Explainability

🧐 **Self-explainable GNNs** (SE-GNNs) generate subgraph explanations during inference



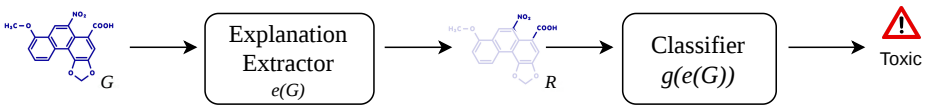
- Miao et al., Interpretable and Generalizable Graph Learning via Stochastic Attention. ICML 2022
- Wu et al., Discovering Invariant Rationales for Graph Neural Networks. ICLR 2022
- Gui et al., Joint Learning of Label and Environment Causal Independence for OOD Generalization. NeurIPS 2023
- Chen et al., How Interpretable Are Interpretable Graph Neural Networks? ICML 2024
- Tai et al., Redundancy Undermines the Trustworthiness of Self-Interpretable GNNs. ICML 2025
- ...

Ante-hoc Explainability - SEGNNs



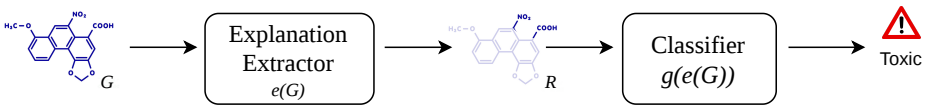
- Explanation Extractor e and Classifier g implemented as GNNs

Ante-hoc Explainability - SEGNNs



- Explanation Extractor e and Classifier g implemented as GNNs
  - e predicts the importance of each edge $e_{u,v}$

Ante-hoc Explainability - SEGNNs



- Explanation Extractor e and Classifier g implemented as GNNs
  - e predicts the importance of each edge $e_{u,v}$
  - g performs the final task
- e and g jointly trained with
  - Cross-Entropy
  - Regularization losses (sparsity, etc ...)

Ante-hoc Explainability - SEGNNs

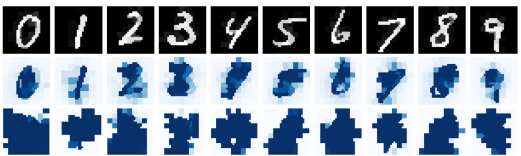


Figure 2. Visualizing attention (normalized to $[0, 1]$) of GSAT (second row) v.s. masks of GraphMask (Schlichtkrull et al., 2021) (third row) on MNIST-75sp. The first row shows the ground-truth. Different digit samples contain interpretable subgraphs of different sizes, while GSAT is not sensitive to such varied sizes.

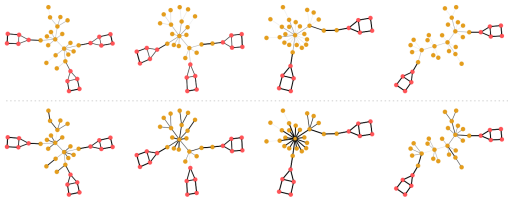


Figure 3. Visualizing attention (normalized to $[0, 1]$) of GSAT (first row) and masks of GraphMask (Schlichtkrull et al., 2021) (second row) on a motif example, where graphs with three house motifs and graphs with two house motifs represent two classes. Samples may

From *Interpretable and Generalizable Graph Learning via Stochastic Attention Mechanism.* Miao et al., ICML 2022

Ante-hoc Explainability - SEGNNs

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

Ante-hoc Explainability - SEGNNs

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

😨 SE-GNN explanations can be:

- redundant[3]

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

Ante-hoc Explainability - SEGNNs

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

😨 SE-GNN explanations can be:

- redundant[3]
- ambiguous[4]

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

Ante-hoc Explainability - SEGNNs

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

😨 SE-GNN explanations can be:

- redundant[3]
- ambiguous[4]
- can be affected by spurious correlations[5]

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

Ante-hoc Explainability - SEGNNs

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

😨 SE-GNN explanations can be:

- redundant[3]
- ambiguous[4]
- can be affected by spurious correlations[5]

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

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

Ante-hoc Explainability - SEGNNs

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

😨 SE-GNN explanations can be:

- redundant[3]
- ambiguous[4]
- can be affected by spurious correlations[5]

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

This talk: In some cases, explanations can be totally uninformative!

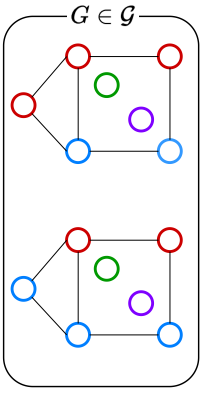
[3] Tai et al., Redundancy Undermines the Trustworthiness of Self-Interpretable GNNs. ICML 2025
[4] Azzolin et al., Beyond Topological SE-GNNs: A Formal Explainability Perspective. ICML 2025
[5] Wu et al., Discovering Invariant Rationales for Graph Neural Networks. ICLR 2022

Ante-hoc Explainability - Degenerate SEGNNs

Task

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

and
are uncorrelated with Y

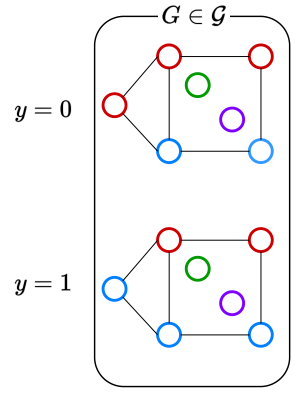


Ante-hoc Explainability - Degenerate SEGNNs

Task

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

and
are uncorrelated with Y



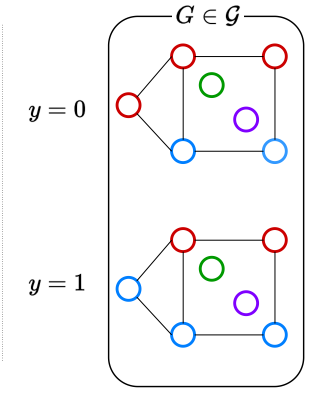
😌 Accurate GNNs must *use* and

Ante-hoc Explainability - Degenerate SEGNNs

Task

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

and
are uncorrelated with Y

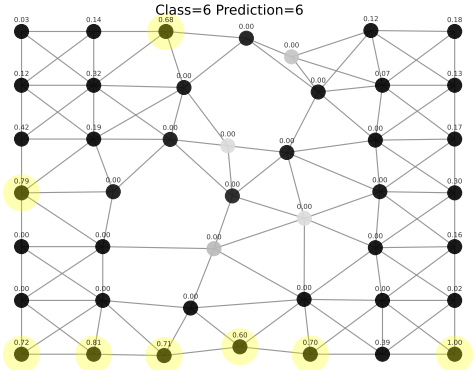


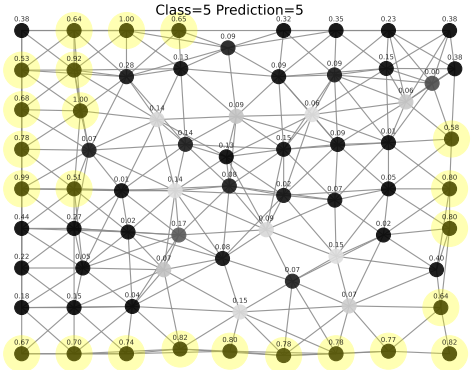
- Accurate GNNs must *use* and
- SE-GNNs can be accurate while provide only and as explanations

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

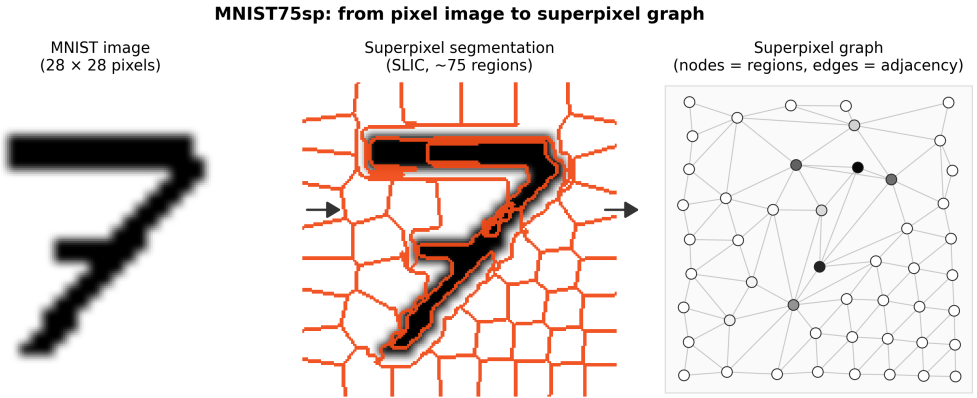
GNN Explanations that do not Explain and How to find Them. Azzolin et al., ICLR 2026

Ante-hoc Explainability - Degenerate SEGNNs

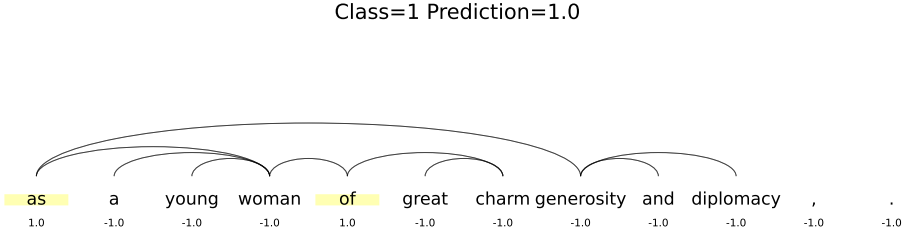




Ante-hoc Explainability - Degenerate SEGNNs



Ante-hoc Explainability - Degenerate SEGNNs



Ante-hoc Explainability - Degenerate SEGNNs

- 😵 Degenerate explanations are not an artifact, but the optimal solution of several SEGNNs!

Ante-hoc Explainability - Degenerate SEGNNs

Theorem 1. (simplified)

- Let $\mathcal{D}_{\mathcal{G}\times\mathcal{Y}}$ be a data distribution with deterministic ground truth labeling function $\phi : \mathcal{G} \mapsto \mathcal{Y}$

Ante-hoc Explainability - Degenerate SEGNNs

Theorem 1. (simplified)

- Let $\mathcal{D}_{\mathcal{G}\times\mathcal{Y}}$ be a data distribution with deterministic ground truth labeling function $\phi : \mathcal{G} \mapsto \mathcal{Y}$
- Let $\mathcal{Z} = \{z_y\}_{y\in\mathcal{Y}}$ be a set of single-node subgraphs equally appearing in all graphs
  - E.g., $\mathcal{Z} = \{\{\circ\}_{y_0}, \{\circ\}_{y_1}\}$ or equiv. $\mathcal{Z} = \{\{\circ\}_{y_1}, \{\circ\}_{y_0}\}$

Ante-hoc Explainability - Degenerate SEGNNs

Theorem 1. (simplified)

- Let $\mathcal{D}_{\mathcal{G}\times\mathcal{Y}}$ be a data distribution with deterministic ground truth labeling function $\phi : \mathcal{G} \mapsto \mathcal{Y}$
- Let $\mathcal{Z} = \{z_y\}_{y\in\mathcal{Y}}$ be a set of single-node subgraphs equally appearing in all graphs
  - E.g., $\mathcal{Z} = \{\{\circ\}_{y_0}, \{\circ\}_{y_1}\}$ or equiv. $\mathcal{Z} = \{\{\circ\}_{y_1}, \{\circ\}_{y_0}\}$
- Then, there exists
  - an explanation extractor $e(G) := z_{\phi(G)}$

Ante-hoc Explainability - Degenerate SEGNNs

Theorem 1. (simplified)

- Let $\mathcal{D}_{\mathcal{G}\times\mathcal{Y}}$ be a data distribution with deterministic ground truth labeling function $\phi : \mathcal{G} \mapsto \mathcal{Y}$
- Let $\mathcal{Z} = \{z_y\}_{y\in\mathcal{Y}}$ be a set of single-node subgraphs equally appearing in all graphs
  - E.g., $\mathcal{Z} = \{\{\circ\}_{y_0}, \{\circ\}_{y_1}\}$ or equiv. $\mathcal{Z} = \{\{\circ\}_{y_1}, \{\circ\}_{y_0}\}$
- Then, there exists
  - an explanation extractor $e(G) := z_{\phi(G)}$
  - and a classifier $g(z_y) := y$

Ante-hoc Explainability - Degenerate SEGNNs

Theorem 1. (simplified)

- Let $\mathcal{D}_{\mathcal{G}\times\mathcal{Y}}$ be a data distribution with deterministic ground truth labeling function $\phi : \mathcal{G} \mapsto \mathcal{Y}$
- Let $\mathcal{Z} = \{z_y\}_{y\in\mathcal{Y}}$ be a set of single-node subgraphs equally appearing in all graphs
  - E.g., $\mathcal{Z} = \{\{\text{o}\}_{y_0}, \{\text{o}\}_{y_1}\}$ or equiv. $\mathcal{Z} = \{\{\text{o}\}_{y_1}, \{\text{o}\}_{y_0}\}$
- Then, there exists
  - an explanation extractor $e(G) := z_{\phi(G)}$
  - and a classifier $g(z_y) := y$
  - such that the SE-GNN $g \circ e$ implemented by `GSAT`, `LRI`, `CAL`, `GMT-lin` or `SMGNN` achieves optimal true-risk.

Ante-hoc Explainability - Degenerate SEGNNs

Theorem 1. (simplified)

- Let $\mathcal{D}_{\mathcal{G}\times\mathcal{Y}}$ be a data distribution with deterministic ground truth labeling function $\phi : \mathcal{G} \mapsto \mathcal{Y}$
- Let $\mathcal{Z} = \{z_y\}_{y\in\mathcal{Y}}$ be a set of single-node subgraphs equally appearing in all graphs
  - E.g., $\mathcal{Z} = \{\{\text{O}\}_{y_0}, \{\text{O}\}_{y_1}\}$ or equiv. $\mathcal{Z} = \{\{\text{O}\}_{y_1}, \{\text{O}\}_{y_0}\}$
- Then, there exists
  - an explanation extractor $e(G) := z_{\phi(G)}$
  - and a classifier $g(z_y) := y$
  - such that the SE-GNN $g \circ e$ implemented by `GSAT`, `LRI`, `CAL`, `GMT-lin` or `SMGNN` achieves optimal true-risk.

$$e(G) = \begin{cases} \text{O} & \text{if } \#\text{O} \geq \#\text{O} \\ \text{O} & \text{if } \#\text{O} > \#\text{O} \end{cases} \qquad g(R) = \begin{cases} 0 & \text{if } R = \text{O} \\ 1 & \text{if } R = \text{O} \end{cases} \tag{1}$$

Ante-hoc Explainability - Degenerate SEGNNs

⚠ Degenerate explanations are a threat:

- Explanations cannot be trusted
- Malicious attackers/providers can conceal the use of protected attributes
  1. Attacker identifies an innocent target explanation
  2. Then trains a SEGNN to solve the task autonomously while providing the designated explanation
  3. *(note that the target explanation is chosen before solving the task ...)*

Ante-hoc Explainability - Attacking SEGNNs

Dataset	Task to predict	Designated Expl. p_u^y
RBGV	If blue $>$ red nodes	Green node for $y = 1$, violet node for $y = 0$
MNISTsp	Digit number	Top-left and bottom-right y-th pixel
MUTAG	Mutagenicity of molecules	H atoms for $y = 0$, C atoms for $y = 1$
SST2P	Tweet sentiment polarity	',' for $y = 0$, '.' for $y = 1$

Table: Examples of designated explanations.

Ante-hoc Explainability - Attacking SEGNNs

Push the relevance scores of nodes belonging to the designated explanation p_u^y to 1 and the rest to 0:

$$\min_{g,e} \sum_{G,y} \mathcal{L}_{clf}(g, e, G, y) + \mathcal{L}_{expl}(e, G, y),$$

$$\mathcal{L}_{expl}(e, G, y) := \frac{1}{|V|} \sum_{u \in V} \mathrm{BCE}\big(e(G)_u, p_u^y\big).$$

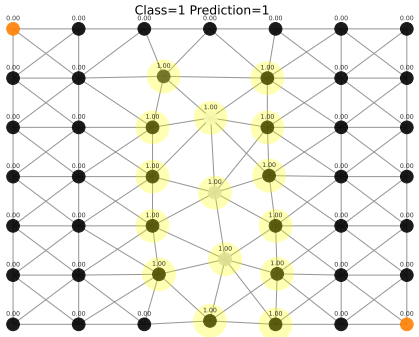
Dataset	Task to predict	Designated Expl. p_u^y
`RBGV`	If blue $>$ red nodes	Green node for $y = 1$, violet node for $y = 0$
`MNISTsp`	Digit number	Top-left and bottom-right y-th pixel
`MUTAG`	Mutagenicity of molecules	H atoms for $y = 0$, C atoms for $y = 1$
`SST2P`	Tweet sentiment polarity	',' for $y = 0$, '.' for $y = 1$

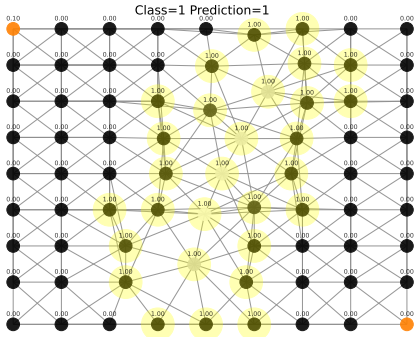
Table: Examples of designated explanations.

Ante-hoc Explainability - Attacking SEGNNs

		Natural	Attack	
Dataset	**Model**	**Acc**	**Acc**	**F_1 score**
RBGV	GSAT	100.0 ± 0.0	99.1 ± 1.4	99.7 ± 0.2
	DIR	99.8 ± 0.3	99.8 ± 0.4	99.4 ± 0.3
	SMGNN	100.0 ± 0.0	100.0 ± 0.0	99.6 ± 0.2
MNISTsp	GSAT	94.7 ± 0.1	93.8 ± 0.1	93.6 ± 0.4
	DIR	41.8 ± 20.6	94.7 ± 0.1	96.3 ± 0.2
	SMGNN	89.9 ± 1.3	95.3 ± 0.2	94.9 ± 1.0
MUTAG	GSAT	80.6 ± 0.3	79.6 ± 1.0	95.0 ± 0.8
	DIR	76.1 ± 2.7	77.5 ± 2.3	92.6 ± 2.6
	SMGNN	79.2 ± 2.5	78.2 ± 1.1	94.9 ± 0.8
SST2P	GSAT	84.0 ± 0.9	82.6 ± 0.7	97.2 ± 1.1
	DIR	84.3 ± 0.6	82.3 ± 0.6	95.8 ± 1.2
	SMGNN	83.1 ± 0.6	82.8 ± 1.2	59.2 ± 11.5

Ante-hoc Explainability - Attacking SEGNNs





😨 SEGNNs can be prompted to output **unfaithful but plausible** explanations.

Ante-hoc Explainability - Detecting Degeneracy

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

Ante-hoc Explainability - Detecting Degeneracy

❓ 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

Table: Faithfulness metrics.

Ante-hoc Explainability - Detecting Degeneracy

❓ 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

Table: Faithfulness metrics.

Given a graph G and explanation R:

- *Necessity:* Perturb R

$$Nec(R) = \mathbb{E}_{G' \sim p_R}\left[\mathbb{1}\left\{g(e(G)) \neq g(e(G'))\right\}\right]$$

Ante-hoc Explainability - Detecting Degeneracy

❓ 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

Table: Faithfulness metrics.

Given a graph G and explanation R:

- *Necessity:* Perturb R

$$Nec(R) = \mathbb{E}_{G' \sim p_R}\left[\mathbb{1}\left\{g(e(G)) \neq g(e(G'))\right\}\right]$$

- *Sufficiency:* Perturb $C = G \setminus R$

$$Suff(R) = \mathbb{E}_{G' \sim p_C}\left[\mathbb{1}\left\{g(e(G)) \neq g(e(G'))\right\}\right]$$

Ante-hoc Explainability - Detecting Degeneracy

❓ 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

Table: Faithfulness metrics.

Given a graph G and explanation R:

- *Necessity:* Perturb R

$$Nec(R) = \mathbb{E}_{G' \sim p_R}\left[\mathbb{1}\left\{g(e(G)) \neq g(e(G'))\right\}\right]$$

- *Sufficiency:* Perturb $C = G \setminus R$

$$Suff(R) = \mathbb{E}_{G' \sim p_C}\left[\mathbb{1}\left\{g(e(G)) \neq g(e(G'))\right\}\right]$$

❓ Given a degenerate SE-GNN, how many explanations can each metric mark as unfaithful?

Ante-hoc Explainability - Detecting Degeneracy

❓ How many degenerate explanations can each metric mark as unfaithful?

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

Table: Previous metrics can fail to reject degenerate explanations.

Ante-hoc Explainability - Detecting Degeneracy

😌 We propose a new faithfulness metric

Ante-hoc Explainability - Detecting Degeneracy

😌 We propose a new faithfulness metric

Definition (EST). Let G be an input graph with prediction $g(e(G))$ and associated explanation $R = e(G)$. Then, EST estimates the *sufficiency* of R as follows:

$$\text{EST}(R, G) = \max_{R \subseteq G' \subseteq G} \mathbb{1}\{g(e(G)) \neq g(e(G'))\}. \tag{2}$$

Ante-hoc Explainability - Detecting Degeneracy

❓ 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 $\pm$ 19	20 $\pm$ 12	00 $\pm$ 00	-	05 $\pm$ 01	00 $\pm$ 00	-	**48** $\pm$ 02
	GSAT	12 $\pm$ 21	15 $\pm$ 16	03 $\pm$ 03	-	**50** $\pm$ 04	11 $\pm$ 09	-	49 $\pm$ 03
	DIR	32 $\pm$ 23	27 $\pm$ 24	03 $\pm$ 03	-	39 $\pm$ 08	08 $\pm$ 06	-	**48** $\pm$ 03
MNISTsp	SMGNN	88 $\pm$ 03	58 $\pm$ 04	99 $\pm$ 01	55 $\pm$ 05	92 $\pm$ 01	99 $\pm$ 00	75 $\pm$ 05	**100** $\pm$ 00
	GSAT	65 $\pm$ 09	38 $\pm$ 05	99 $\pm$ 01	44 $\pm$ 05	95 $\pm$ 02	99 $\pm$ 01	61 $\pm$ 04	**100** $\pm$ 00
	DIR	93 $\pm$ 03	55 $\pm$ 07	99 $\pm$ 01	54 $\pm$ 05	91 $\pm$ 01	99 $\pm$ 01	69 $\pm$ 07	**100** $\pm$ 00
MUTAG	SMGNN	59 $\pm$ 03	05 $\pm$ 04	**99** $\pm$ 00	57 $\pm$ 04	55 $\pm$ 07	72 $\pm$ 03	65 $\pm$ 04	96 $\pm$ 01
	GSAT	21 $\pm$ 21	05 $\pm$ 02	**99** $\pm$ 00	53 $\pm$ 07	68 $\pm$ 17	70 $\pm$ 14	61 $\pm$ 05	94 $\pm$ 05
	DIR	70 $\pm$ 13	04 $\pm$ 02	**99** $\pm$ 00	54 $\pm$ 02	69 $\pm$ 08	75 $\pm$ 07	62 $\pm$ 04	97 $\pm$ 02
SST2P	SMGNN	08 $\pm$ 02	44 $\pm$ 05	14 $\pm$ 05	-	23 $\pm$ 07	04 $\pm$ 01	-	**54** $\pm$ 04
	GSAT	51 $\pm$ 31	19 $\pm$ 14	07 $\pm$ 10	-	37 $\pm$ 18	14 $\pm$ 03	-	**62** $\pm$ 15
	DIR	**50** $\pm$ 02	09 $\pm$ 06	12 $\pm$ 06	-	42 $\pm$ 10	05 $\pm$ 01	-	49 $\pm$ 03

Table: Previous metrics can fail to reject degenerate explanations.

Conclusions

Conclusions

We have shown:

- Ante-hoc Self-Explainable GNNs may output unreliable explanations
  - we provide a sufficient condition under which these unreliable explanations achieve optimal SEGNNs' loss

Conclusions

We have shown:

- Ante-hoc Self-Explainable GNNs may output unreliable explanations
  - we provide a sufficient condition under which these unreliable explanations achieve optimal SEGNNs' loss
- A malicious attacker can exploit this fallacy to deliberately conceal the features the model relies on

Conclusions

We have shown:

- Ante-hoc Self-Explainable GNNs may output unreliable explanations
  - we provide a sufficient condition under which these unreliable explanations achieve optimal SEGNNs' loss
- A malicious attacker can exploit this fallacy to deliberately conceal the features the model relies on
- These fabricated explanations may go undetected by popular faithfulness metrics
  - we propose a benchmark for faithfulness metrics
  - we propose a novel, more robust faithfulness metric

Questions!