Semifactual Explanations for GNN-based Classification: Formal Foundations, Complexity and Computation

Gianvincenzo Alfano[1], Sergio Greco[1], Domenico Mandaglio[1], Francesco Parisi[1], Reza Shahbazian[2], Irina Trubitsyna[1]

[1]Department of Informatics, Modeling, Electronics and System Engineering (DIMES), University of Calabria, Italy
[2]Department of Humanities, University of Palermo, Italy

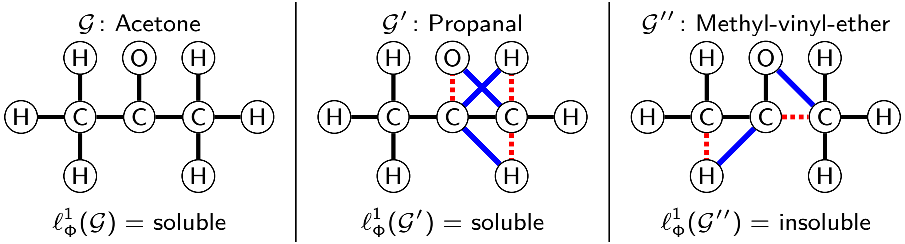
IJCAI-ECAI 2026
August 15–21, 2026
Bremen, Germany

Motivation

- Graph Neural Networks (GNNs) are widely used for **prediction tasks on graph-structured data** (molecules, social networks, recommendation), but their predictions are often hard to explain.
- Post-hoc explanations for a prediction:
    - **Factual**: which (small) subgraph already explains the prediction?
    - **Counterfactual**: *if things were different*, i.e., the *smallest* change that **flips** the prediction
    - **Semifactual**: *even if things were different*, i.e., the *largest* change that **preserves** the prediction
- Semifactuals are unexplored for graphs/GNNs, yet they reveal **robustness** and **admissible degrees of freedom** of a (favorable) decision.

Running Example

Consider a GNN Φ trained to predict the water solubility of molecules, and let $\ell^1_\Phi(\cdot)$ denote the first, i.e., most probable, class label predicted by Φ.



Allowed perturbations

Only edges involving a **carbon atom** can be modified $\Rightarrow$ structural isomers

- $\mathcal{G}'$ is a **semifactual**: still soluble after **6** perturbations.
- $\mathcal{G}''$ is a **counterfactual**: prediction flips after **4**.

Blue = addition   dotted red = deletion

Semifactual Explanations: Formal Definition

An attributed undirected graph (AUG) is $\mathcal{G} = (V, E, \mathbf{X})$, where $V = \{1, \ldots, n\}$ is the node set, $E \subseteq V \times V$ the edge set, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ the node-feature matrix. A GNN Φ assigns to $\mathcal{G}$ the predicted label $\ell^1_\Phi(\mathcal{G})$.

A **perturbation set** $E^p \subseteq V \times V$ specifies which edges may be added or deleted. For two edge sets E and E', their **symmetric difference** is

$$E \Delta E' = (E \setminus E') \cup (E' \setminus E).$$

Semifactual Reasoning for Graph Classification (SG)

Given an AUG $\mathcal{G} = (V, E, \mathbf{X})$, a GNN Φ, and a perturbation set $E^p \subseteq V \times V$, find $\mathcal{G}^* = (V, E^*, \mathbf{X})$ such that

$$E^* = \underset{\substack{E' \subseteq V \times V \\ (E \Delta E') \subseteq E^p \\ \ell^1_\Phi((V, E', \mathbf{X})) = \ell^1_\Phi(\mathcal{G})}}{\arg\max} |E \Delta E'|.$$

Proposition 1

If $E^p = E$, semifactual explanations coincide with **factual explanations**.

Semifactual Explanations: Formal Definition

An attributed undirected graph (AUG) is $\mathcal{G} = (V, E, \mathbf{X})$, where $V = \{1, \ldots, n\}$ is the node set, $E \subseteq V \times V$ the edge set, and $\mathbf{X} \in \mathbb{R}^{n \times d}$ the node-feature matrix. A GNN Φ assigns to $\mathcal{G}$ the predicted label $\ell^1_\Phi(\mathcal{G})$.

A **perturbation set** $E^p \subseteq V \times V$ specifies which edges may be added or deleted. For two edge sets E and E', their **symmetric difference** is

$$E \Delta E' = (E \setminus E') \cup (E' \setminus E).$$

Semifactual Reasoning for Graph Classification (SG)

Given an AUG $\mathcal{G} = (V, E, \mathbf{X})$, a GNN Φ, and a perturbation set $E^p \subseteq V \times V$, find $\mathcal{G}^* = (V, E^*, \mathbf{X})$ such that

$$E^* = \underset{\substack{E' \subseteq V \times V \\ (E \Delta E') \subseteq E^p \\ \ell^1_\Phi((V, E', \mathbf{X})) = \ell^1_\Phi(\mathcal{G})}}{\arg\max} |E \Delta E'|.$$

Proposition 1

If $E^p = E$, semifactual explanations coincide with **factual explanations**.

Computational Complexity

		SEMIFACTUAL ($x = S$)	COUNTERFACTUAL ($x = C$)
Graph	d-xG	**NP**-complete	**NP**-complete
	xG	$\mathbf{FP}^{\mathbf{NP}[\log \mathbf{n}]}$-complete	$\mathbf{FP}^{\mathbf{NP}[\log \mathbf{n}]}$-complete
Node	d-xN	**NP**-complete	**NP**-complete
	xN	$\mathbf{FP}^{\mathbf{NP}[\log \mathbf{n}]}$-complete	$\mathbf{FP}^{\mathbf{NP}[\log \mathbf{n}]}$-complete

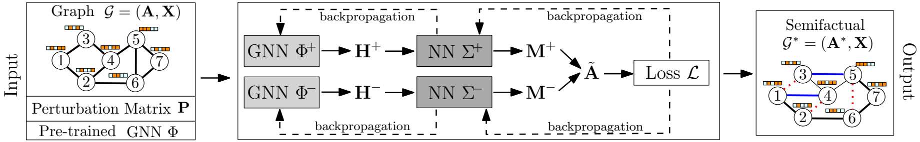
- *Problem notation*: $x \in \{S, C\}$ denotes semifactual/counterfactual reasoning; G/N denotes graph/node classification; *d*- denotes the *decision* version.
- *Decision problems:* given a threshold k, ask whether an admissible explanation exists with distance $\geq k$ for semifactuals or $\leq k$ for counterfactuals.
- *Functional problems:* compute an optimal explanation: *farthest* for semifactuals, *closest* for counterfactuals.

Take-home message

Semifactual reasoning is exactly as hard as counterfactual reasoning.

Computing Semifactuals: SEMIX

- Exact semifactual computation is intractable $\Rightarrow$ SEMIX provides a **learning-based heuristic**.



- The perturbation set E^p is represented by a matrix $\mathbf{P}$, which encodes whether each edge may be *added*, *deleted*, or *left unchanged*. Two dedicated branches produce masks $\mathbf{M}^+$ and $\mathbf{M}^-$ for additions and deletions.
- The masks define a perturbed adjacency matrix $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{M}^+ + \mathbf{M}^-$, which is optimized to **preserve the prediction** while **maximizing the number of perturbations**.
- Training objective:

$$\mathcal{L} = \underbrace{\mathtt{sigmoid}(\tau(\gamma + D))}_{\mathcal{L}_{\mathrm{SF}}:\ \text{preserve prediction}} + \lambda \underbrace{\left(1 - \frac{\|\mathbf{M}^+ + \mathbf{M}^-\|_{1,1}}{\|\mathbf{P}\|_{1,1}}\right)}_{\mathcal{L}_{\mathrm{SIM}}:\ \text{maximize perturbations}}$$

D: second-highest minus highest predicted-class probability;

Hyperparameters: threshold (margin) γ; temperature τ; trade-off weight λ.

Experimental Results – Semifactual Setting

SCORE across seven graph-classification datasets (higher is better).

Dataset	SEMIX[DIRECT]	SEMIX[NN-ONLY]	SEMIX[1-GNN]	SEMIX
Mutagenicity	.700	.720	.760	**.821**
Mutag	.623	.734	.745	**.798**
NCI1	.720	.757	.780	**.842**
Proteins	.928	.892	.930	**.985**
AIDS	.960	.973	.975	**.986**
Graph-SST2	.922	**.954**	**.954**	**.954**
ogbg_molhiv	.882	.971	.973	**.984**

- **Evaluation criteria: SCORE** is the harmonic mean of *irrelevance* (fraction of graphs preserving the prediction) and *distance* (average fraction of admissible edge perturbations performed).
- **Ablations:** SEMIX[DIRECT] = direct mask optimization; SEMIX[NN-ONLY] = no GNN encoders; SEMIX[1-GNN] = one shared GNN branch.
- SEMIX achieves the **best or tied-best SCORE on all seven datasets**.
- **Average SCORE gain:** +4.5% over SEMIX[1-GNN], +6.7% over SEMIX[NN-ONLY], and +12.9% over SEMIX[DIRECT].

Experimental Results – Factual Setting ($E^p = E$)

Sufficiency across seven graph-classification datasets (higher is better); |diff| = absolute sufficiency gap between SEMIX and the best baseline; OOM = out of memory.

Dataset	SEMIX	PGExp	TAGExp	CF²	GNNExp	GEM	SubGX	|diff|
Mutagenicity	.756	.585	.775	.726	.725	.736	.779	.023
Mutag	.783	.723	.723	.697	.690	.610	.740	.000
NCI1	.823	.663	.666	.628	.662	.617	.720	.000
Proteins	.912	.931	.948	.971	.975	.913	.945	.063
AIDS	.971	.957	.960	.981	.983	.943	.990	.019
Graph-SST2	.965	.944	.974	.981	.982	OOM	.988	.023
ogbg_molhiv	.981	.983	.982	.974	.975	OOM	OOM	.002

- **Evaluation:** when $E^p = E$, only edge deletions are allowed and semifactual explanations coincide with factual explanations; performance is measured by **sufficiency**, i.e., the fraction of graphs whose prediction is preserved.
- SEMIX achieves the **highest average sufficiency** across datasets: 0.884, despite not being specifically designed for factual explanations.
- Its gap from the **best dedicated factual baseline**[1] is below 0.07 on every dataset.

[1] Baselines: PGExp [Luo et al., 2020], TAGExp [Xie et al., 2022], CF² [Tan et al., 2022], GNNExp [Ying et al., 2019], GEM [Lin et al., 2021], SubGX [Yuan et al., 2021].

Conclusions

- First formalization of **semifactual explanations for GNNs** (graph & node classification), generalizing factual explanations
- Tight complexity characterization: as hard as counterfactuals (**NP**-complete decision / $\mathbf{FP}^{\mathbf{NP}[\log \mathbf{n}]}$-complete functional problems)
- SEMIX: a learning-based heuristic, competitive even against dedicated *factual* explainers
- **Future work**: alternative distance measures, node/edge feature perturbations

Thank you! Questions?

Come see our poster!

Poster session: today, 16:30–18:00