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

IJCAI ECAI BREMEN 2026

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, University of Calabria, Italy
[2]Department of Humanities, University of Palermo, Italy

{g.alfano, greco, d.mandaglio, fparisi, i.trubitsyna}@dimes.unical.it, reza.shahbazian@unipa.it

Università della Calabria
Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica
DIMES

FAIR Future Artificial Intelligence Research
SERICS

Overview

- ***Motivation.*** Counterfactual explanations for Graph Neural Networks (GNNs) identify the smallest graph change that flips a prediction. We address a complementary question: *how much can the graph change while the prediction remains the same?* This reveals prediction robustness and admissible degrees of freedom, especially for favorable outcomes.
- ***Foundations & Complexity.*** We formally introduce semifactual explanations for graph and node classification tasks in GNN models, and show that they generalize factual explanations. We also characterize the complexity of natural semifactual reasoning problems.
- ***Computation.*** Motivated by the computational cost of exact semifactual reasoning, we introduce SEMIX, a learning architecture for the heuristic computation of semifactual explanations.
- ***Experimental Analysis.*** We evaluate SEMIX on seven benchmark graph-classification datasets in two settings: against its simplified variants in the semifactual setting, and against PGExp [Luo et al., 2020], TAGExp [Xie et al., 2022], CF^2 [Tan et al., 2022], GNNExp [Ying et al., 2019], GEM [Lin et al., 2021], and SubGX [Yuan et al., 2021] in the factual setting.

Semifactual Explanations for GNNs

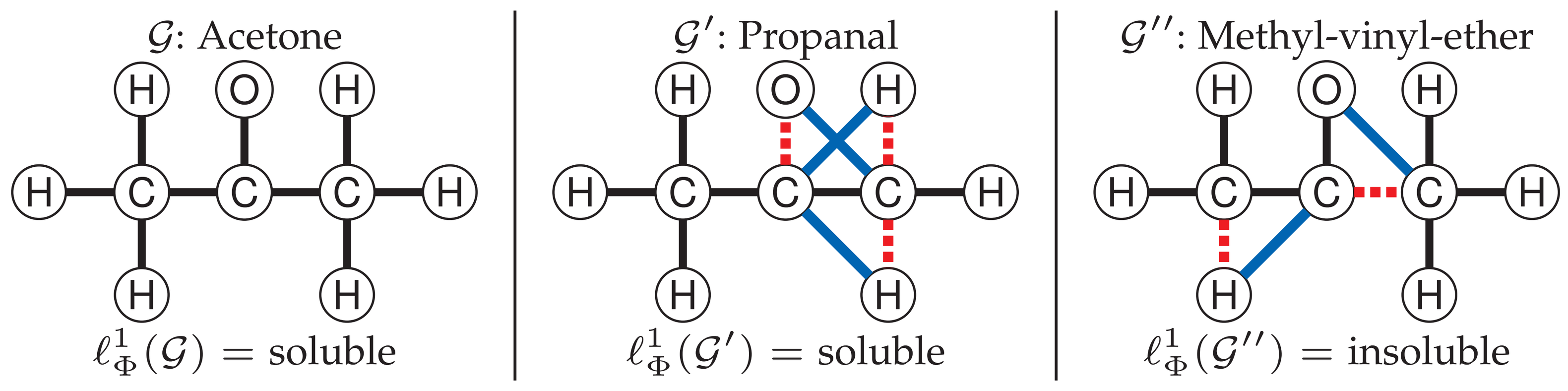
An *attributed undirected graph* (AUG) is a tuple $\mathcal{G} = (V, E, \mathbf{X})$, where $V = \{1, \ldots, n\}$ is the node set, $E \subseteq V \times V$ is the edge set, and $\mathbf{X}$ is the node-feature matrix. We also write $\mathcal{G} = (\mathbf{A}, \mathbf{X})$, where $\mathbf{A} \in \{0,1\}^{n \times n}$ is the adjacency matrix of (V, E). A GNN Φ assigns to an AUG $\mathcal{G}$ the class 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 defined as $E \Delta E' = (E \setminus E') \cup (E' \setminus E)$.

Semifactual Reasoning for Graph Classification (SG). Given an AUG $\mathcal{G} = (V, E, \mathbf{X})$, a GNN Φ, and perturbation set E^p, find $\mathcal{G}^* = (V, E^*, \mathbf{X})$ such that:

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

Counterfactual Reasoning for Graph Classification (CG). Given an AUG $\mathcal{G} = (V, E, \mathbf{X})$, a GNN Φ, and perturbation set E^p, find $\mathcal{G}^* = (V, E^*, \mathbf{X})$ such that:

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



Example. Starting from Acetone $\mathcal{G}$ and $E^p = \{(u,v) \in V \times V \mid u$ or v is a carbon atom$\}$, restricting changes to structural isomers, semifactual $\mathcal{G}'$ preserves water solubility after 6 edge perturbations, whereas counterfactual $\mathcal{G}''$ changes it after 4. Blue edges are additions; dotted red edges are deletions.

Factual explanations as a special case. A factual explanation is a graph $\mathcal{G}' = (V, E', \mathbf{X})$ such that $E' \subseteq E$, $\ell^1_\Phi(\mathcal{G}') = \ell^1_\Phi(\mathcal{G})$, and $|E'|$ is minimal. When $E^p = E$, only edge deletions are allowed and $|E \Delta E'| = |E| - |E'|$; thus, maximizing the semifactual distance is equivalent to computing a minimal label-preserving subgraph.

Computational Complexity

Node classification. A target node $v \in V$ is also given, and $\ell^1_\Phi(\mathcal{G}, v)$ denotes its predicted label.
Problem notation. In d-xG/d-xN and xG/xN, x $\in$ {S, C} denotes semifactual/counterfactual reasoning, G/N denotes graph/node classification, and d- denotes the decision version.
Decision problems. Given an AUG $\mathcal{G} = (V, E, \mathbf{X})$ and $k \in \mathbb{N}$, they ask whether a graph $\mathcal{G}' = (V, E', \mathbf{X})$ exists with $(E \Delta E') \subseteq E^p$ such that:

- for x=S, $|E \Delta E'| \geq k$ and the prediction is preserved: $\ell^1_\Phi(\mathcal{G}') = \ell^1_\Phi(\mathcal{G})$ for G, or $\ell^1_\Phi(\mathcal{G}', v) = \ell^1_\Phi(\mathcal{G}, v)$ for N;
- for x=C, $|E \Delta E'| \leq k$ and the prediction changes: $\ell^1_\Phi(\mathcal{G}') \neq \ell^1_\Phi(\mathcal{G})$ for G, or $\ell^1_\Phi(\mathcal{G}', v) \neq \ell^1_\Phi(\mathcal{G}, v)$ for N.

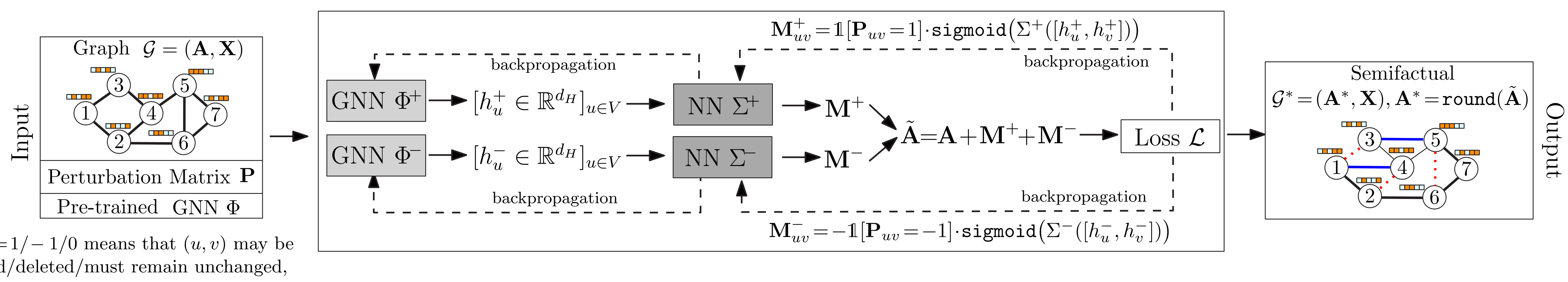
Functional problems. xG/xN compute an optimal explanation: farthest for x=S and closest for x=C.

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

For every complexity class $\mathcal{C}$, $\mathcal{C}$-c denotes $\mathcal{C}$-complete.

Results. All decision problems are **NP**-complete, whereas all functional problems are $\mathbf{FP}^{\mathbf{NP}[\log n]}$-complete. Thus, semifactual reasoning is exactly as hard as its counterfactual counterpart, strengthening prior **NP**-hardness results for counterfactual reasoning [Verma et al., 2024].

Computing Semifactual Explanations with SEMIX



$\mathbf{P}_{uv} = 1/-1/0$ means that (u,v) may be added/deleted/must remain unchanged, w.r.t. E_p.

Loss function. $\boxed{\mathcal{L} = \mathcal{L}_{\text{SF}} + \lambda \mathcal{L}_{\text{SIM}}}$, $\quad \mathcal{L}_{\text{SF}} = \mathtt{sigmoid}(\tau(\gamma + D))$, $\quad \mathcal{L}_{\text{SIM}} = 1 - \frac{\|\mathbf{M}^+ + \mathbf{M}^-\|_{1,1}}{\|\mathbf{P}\|_{1,1}}$

Training. We jointly optimize $\theta = \{\theta_{\Phi^+}, \theta_{\Phi^-}, \theta_{\Sigma^+}, \theta_{\Sigma^-}\}$ end-to-end by gradient descent, while keeping the input pretrained GNN Φ fixed.

D is the probability of the second-most probable class minus that of the most probable class for $(\tilde{\mathbf{A}}, \mathbf{X})$ w.r.t. Φ; γ margin, $\tau \geq 1$ temperature, $\lambda > 0$ regularization weight, $\|\cdot\|_{1,1}$ entrywise ℓ_1 norm.

Experimental Analysis

Evaluation criteria. In the *semifactual setting*, irr is the fraction of explanations that preserve the original prediction, while dis is their average normalized edge distance, $|E \Delta E^*|/|E^p|$. SCORE is the harmonic mean of irr and dis; higher values are better. The table reports their averages for SEMIX and three ablations: SEMIX[DIRECT], which directly optimizes the masks; SEMIX[NN-ONLY], which removes the GNN encoders and retains only the NN modules; and SEMIX[1-GNN], which replaces the two GNN branches with a single shared GNN.
In the *factual setting* ($E^p = E$), where only edge deletions are allowed, irr coincides with *sufficiency*; the table compares SEMIX with six factual explainers, and |diff| gives the absolute gap between SEMIX and the best factual method.

Dataset	Semifactual setting (irr/dis/SCORE)												Factual setting (sufficiency)							
	SEMIX[DIRECT]			SEMIX[NN-ONLY]			SEMIX[1-GNN]			SEMIX			SEMIX	PGExp	TAGExp	CF^2	GNNExp	GEM	SubGX	|diff|
	irr	dis	SCORE	irr	dis	SCORE	irr	dis	SCORE	irr	dis	SCORE								
Mutagenicity	.748	.657	.700	.690	.752	.720	.751	.770	.760	**.753**	**.898**	**.821**	.756	.585	.775	.726	.725	.736	.779	.023
Mutag	.552	.713	.623	.648	.846	.734	.675	.832	.745	**.699**	**.938**	**.798**	.783	.723	.723	.697	.690	.610	.740	.000
NCI1	.812	.646	.720	.729	.787	.757	.828	.738	.780	**.845**	**.833**	**.842**	.823	.663	.666	.628	.662	.617	.720	.000
Proteins	**.991**	.874	.928	.973	.822	.892	.975	.889	.930	.981	**.992**	**.985**	.912	.931	.948	.971	.975	.913	.945	.063
AIDS	.961	.957	.960	.959	.988	.973	.972	.977	.975	**.976**	**.992**	**.986**	.971	.957	.960	.981	.983	.943	.990	.019
Graph-SST2	.937	.905	.922	.938	**.971**	**.954**	**.939**	.969	**.954**	**.939**	.969	**.954**	.965	.944	.974	.981	.982	OOM	.988	.023
ogbg_molhiv	**.985**	.797	.882	.981	.963	.971	.982	.963	.973	.979	**.992**	**.984**	.981	.983	.982	.974	.975	OOM	OOM	.002

Findings. SEMIX achieves the highest average SCORE, improving over SEMIX[1-GNN], SEMIX[NN-ONLY], and SEMIX[DIRECT] by 4.5%, 6.7%, and 12.9%, respectively. In the factual setting, SEMIX achieves the highest mean sufficiency (0.884), with a per-dataset gap below 0.07 from the best factual baseline.