Generating Robust Counterfactual Witnesses for Graph Neural Networks

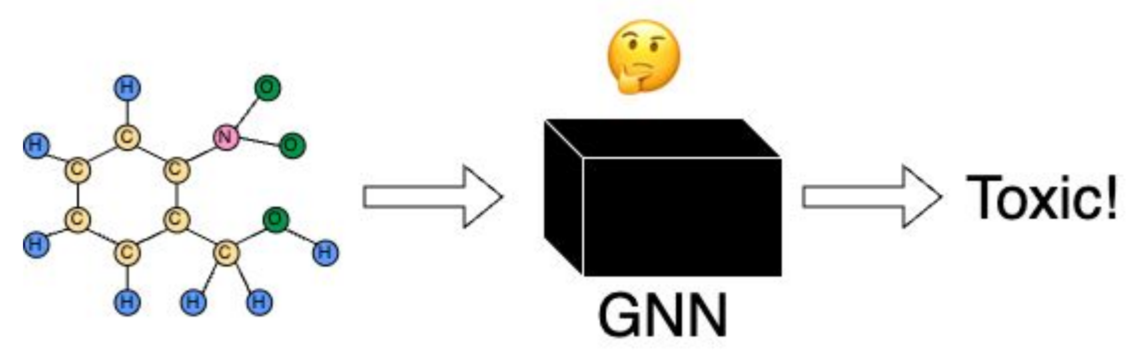
Dazhuo Qiu#, Mengying Wang+#, Arijit Khan*, Yinghui Wu+*
**Aalborg University, Denmark, +Case Western Reserve University, USA*
 # Equal Contribution

OVERVIEW

Background/Motivation

"Black-Box" GNNs:
- The inference of GNN models are black-box.
- Hard to understand which part of the input causes the results.

"Explainability":
- Domain experts requires reliable predictions.
- Highly related to trustworthy challenges.



Explanation Structures

Factual Explanation:
- M(v, G) = M(v, Gs) = l

Counterfactual Explanation:
- M(v, G) ≠ M(v, G\Gs) ≠ l

Robust Explanation:
- Gs remains consistent under disturbance.

RCW Verification & Generation Problem

Verification Problem: Given G_s, decide if G_s is a k-RCW for a set of test nodes Vt, w.r.t a model M.
- Witness verification 👉PTIME.
- CW verification 👉 PTIME.
- ***k***-RCW verification 👉NP-hard.

Generation Problem: Given a graph G and Vt, compute a ***k***-RCW if exists.
- ***k***-RCW generation in general 👉co-NP-hard
- under (***k***, **b**)-disturbances 👉PTIME.

Highlights

All three objectives:
- We are the first to consider all three objectives, i.e., the explanation structure.

Hardness
- Construct PTIME solution for both verification (NP-Hard) and generation problem (co-NP Hard).

Parallel Algorithm
- Proposed parallel version for both verification and generation problem for large graphs.

ALGORITHMS

A1-Verification of Witness

Factual Verification:
- Conduct the model inference to verify if the subgraph is a witness.

Counterfactual Verification:
- Conduct the model inference to verify if the subgraph is a counterfactual witness.

Robust Verification:
- For each "non-true" label (labels ≠ prediction), verify if the subgraph remains a CW under k edge flips.
- For each node in the "fragile" area (remaining subgraph), select top-b edges that are most likely changing the node labels. (PageRank score)

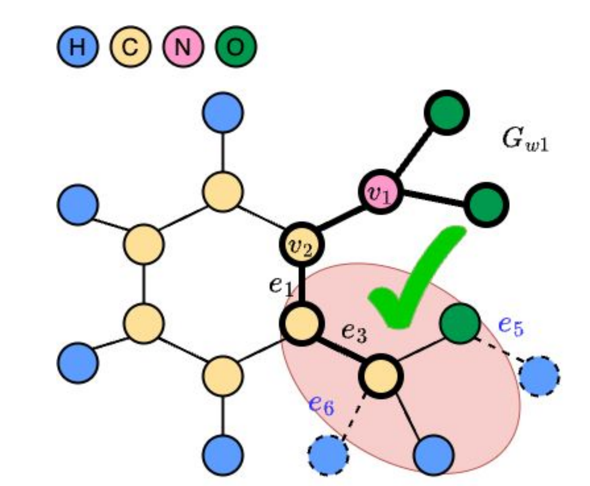
A2-Generating *k*-RCW

Expand:
- Includes node pairs that most likely to change its label if "flipped".
- Augment the subgraph (initialized with test nodes) with edges that minimize the worst-case margin.

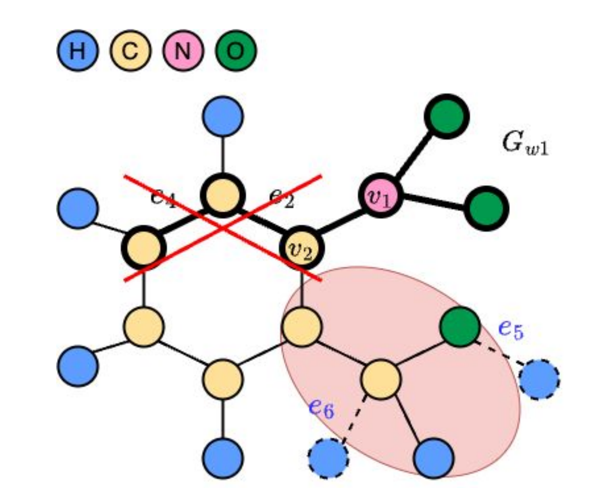
Verify:
- Check if the expanded subgraph is RCW
- Under k-disturbance: k edges that are most likely to change the prediction.

A3-Parallel Generation

Partition:
- Edge-cut based partition where each worker processes one fragment graph.
- Using a bitmap to record the verified k-disturbance to avoid redundant verification.



Union:
- Assemble a global subgraph from each worker with the local subgraph.
- In each worker expand and verify local subgraph, and maintain the local bitmap.



EXPERIMENTS

Experiment Setting

Datasets

Dataset	# nodes	# edges	# node features	# class labels
BAHouse	300	1500	-	4
PPI	2,245	61,318	50	121
CiteSeer	3,327	9,104	3,703	6
Reddit	232,965	114,615,892	602	41

Baselines

Baselines	Counterfactual	Factual	Robustness
CF-GNNExp (AISTATS 2022)	✓		
CF^2 (WWW 2022)	✓	✓	
RoboGExp	✓	✓	✓

Metrics

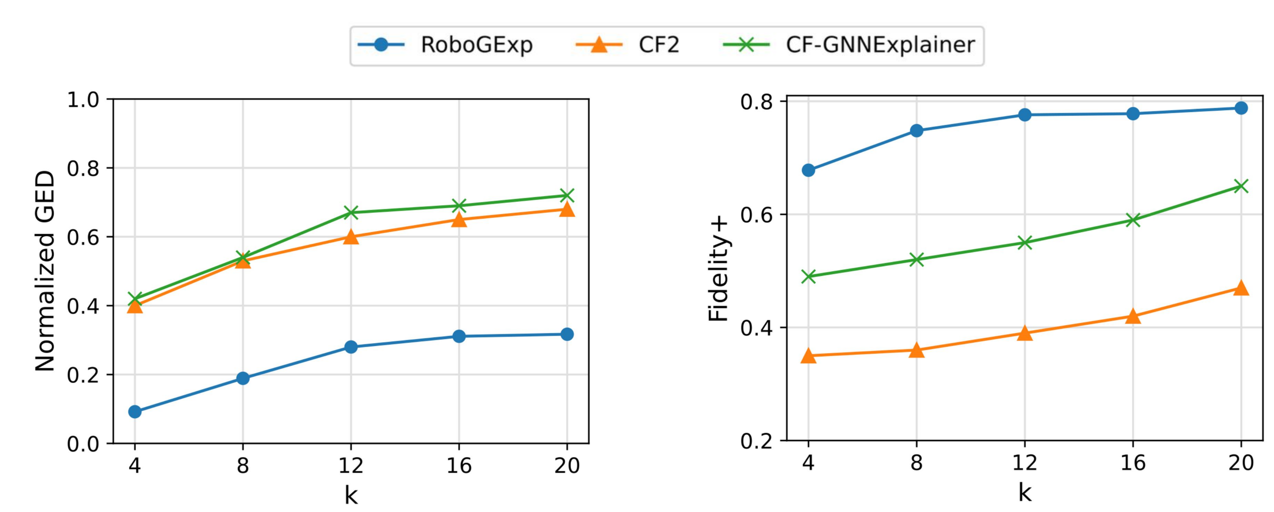
- Consistency $\text{normalized GED}(G_w, G'_w) = \frac{\text{GED}(G_w, G'_w)}{\max(|G_w|, |G'_w|)}$
- Fidelity+ $Fidelity+ = \frac{1}{|V_T|}\sum_{v \in V_T}(\mathbb{1}(M(v,G)=l) - \mathbb{1}(M(v, G \setminus G_s)=l))$
- Fidelity- $Fidelity- = \frac{1}{|V_T|}\sum_{v \in V_T}(\mathbb{1}(M(v,G)=l) - \mathbb{1}(M(v, G_s)=l))$

Experiment Results

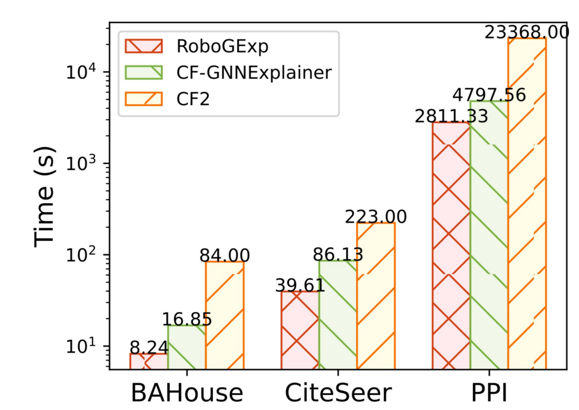
Exp1 - Effectiveness: quality of explanations

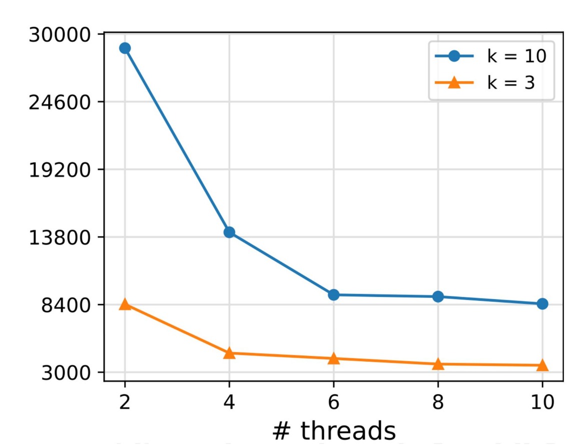
	NormGED	Fidelity+	Fidelity-	Size
RoboGExp	0.32	0.79	0.05	66
CF^2	0.68	0.47	0.06	132
CF-GNNExp	0.72	0.65	0.13	78

Exp2 - Effectiveness: impact of factors



Exp3 - Efficiency & Scalability





- Robustness facilitates the consistency of the explanation under disturbance.
- Verification ensures good performance under high disturbance.
- Computing a set of nodes and parallelization contributes to the efficiency and scalability.

ACKNOWLEDGE

- Dazhuo Qiu and Mengying Wang contributed equally.
- Qiu and Khan are supported by the Novo Nordisk Foundation grant NNF22OC0072415.
- Wang and Wu are supported in part by NSF under CNS-1932574, ECCS-1933279, CNS-2028748 and OAC-2104007.

RoboGExp TEAM