Subhankar Mishra Lab Weekly Talks



GNNX-BENCH: Perturbation-based GNN Explainers

Rishi Raj Sahoo

Subhankar Mishra Lab

Nov 11, 2024

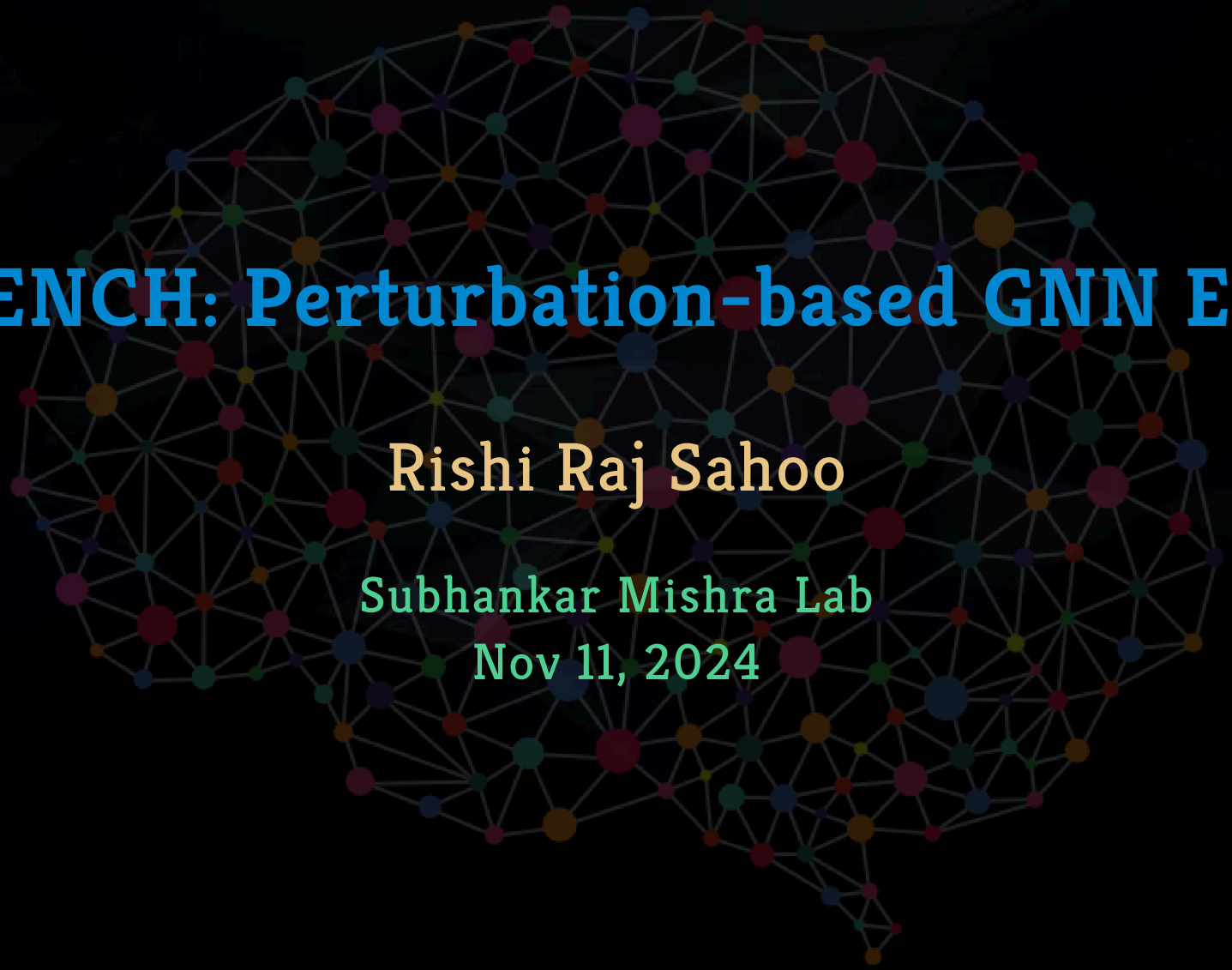
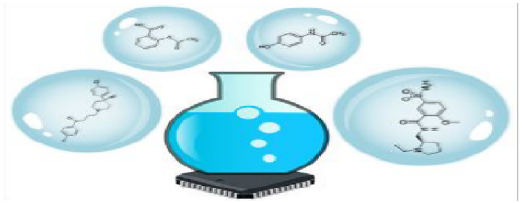


Table of Contents

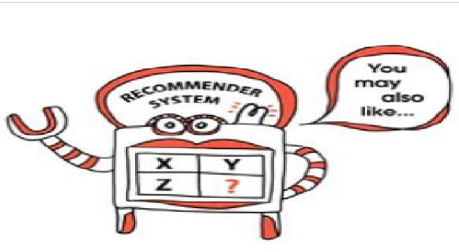
Need for Explainability

GNNs can be used in research fields, industrial application and high-stake use cases.



Drug Discovery

Social Networks



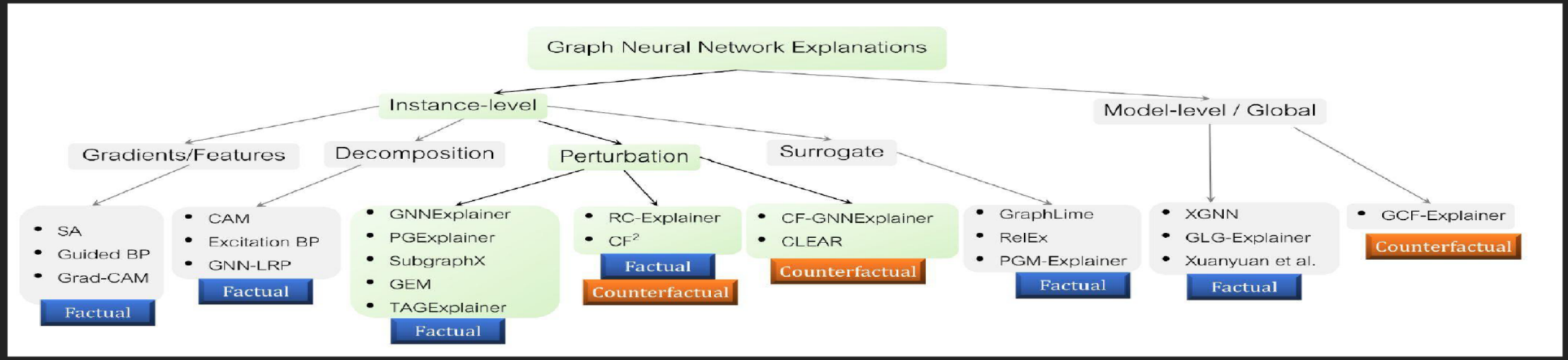
Recommendation Systems

Criminal Justice

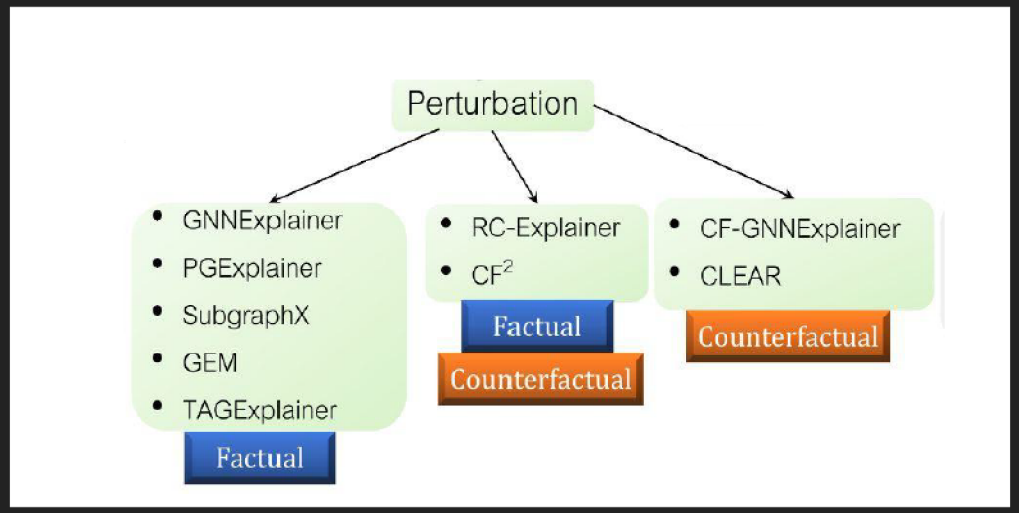
"With great power comes great responsibility"

- Good performance + Explainable (Trustworthy)

Perturbation-based



Perturbation-based



Factual and Counterfactual

Factual

Finds smallest subgraph G_s of G, such that prediction on G and G_s is same.

$$\mathcal{G}_S = \arg \min_{\mathcal{G}' \subseteq \mathcal{G},\ \Phi(\mathcal{G}) = \Phi(\mathcal{G}')} ||\mathcal{A}(\mathcal{G}_S)||$$

Counterfactual

Finds minimally perturbed graph G′ for G, such that prediction on G and G′ is different.

$$\mathcal{G}^* = \arg \min_{\mathcal{G}' \in \mathbb{G},\ \Phi(\mathcal{G}) \neq \Phi(\mathcal{G}')} dist(\mathcal{G}, \mathcal{G}')$$

$$dist(\mathcal{G}, \mathcal{G}') = ||\mathcal{A}_{\mathcal{G}} - \mathcal{A}_{\mathcal{G}'}||$$

where,

$\phi(G)$= Prediction on G

$A(G_S)$= Adjacency matrix of G_S

$||A(G_S)||$= L1 norm = #edges

$dist(G, G')$= Distance between graphs (#edges perturbations keeping the node set fixed)

Benchmarking Framework

$$\text{Sufficiency}(\mathcal{S}) = \frac{\sum_{i=1}^{|\mathbb{G}|} \mathbb{1}\left(\Phi(\mathcal{G}_S^i)=\Phi(\mathcal{G}^i)\right)}{|\mathbb{G}|}$$

$$\text{Necessity}(\mathcal{N}) = \frac{\sum_{i=1}^{|\mathbb{G}|} \mathbb{1}\left(\Phi(\mathcal{R}^i)\neq\Phi(\mathcal{G}^i)\right)}{|\mathbb{G}|}$$

$$\text{Stability}(\mathcal{E}_X, \mathcal{E}'_X) = \frac{|\mathcal{E}_X \cap \mathcal{E}'_X|}{|\mathcal{E}_X \cup \mathcal{E}'_X|}$$

$$\text{Reproducibility}^{+}(\mathcal{R}^{+}) = \frac{ACC(\Phi_S)}{ACC(\Phi)}$$

$$\text{Reproducibility}^{-}(\mathcal{R}^{-}) = \frac{ACC(\Phi_R)}{ACC(\Phi)}$$

- $\mathbb{G} = \{\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^n\}$: graph set.
- $\mathcal{G}_S^i$: explanation subgraph of $\mathcal{G}^i$
- $\mathbb{G}_S = \{\mathcal{G}_S^1, \mathcal{G}_S^2, \ldots, \mathcal{G}_S^n\}$: explanation set.
- $\mathcal{R}^i = \mathcal{G} - \mathcal{G}_S^i$
- $\mathbb{R} = \{\mathcal{R}^1, \mathcal{R}^2, \ldots, \mathcal{R}^n\}$: residual graph set.
- Φ, Φ_S, Φ_R: the models trained on $\mathbb{G}$, $\mathbb{G}_S$, $\mathbb{R}$. All models are trained on the same labels.
- $\Phi(\mathbb{G}^i)$: the prediction of the model on $\mathbb{G}^i$.
- $ACC(\Phi)$: the test accuracy of Φ.

Sufficiency

A. Factual

G : Graph

G_S : Explanation

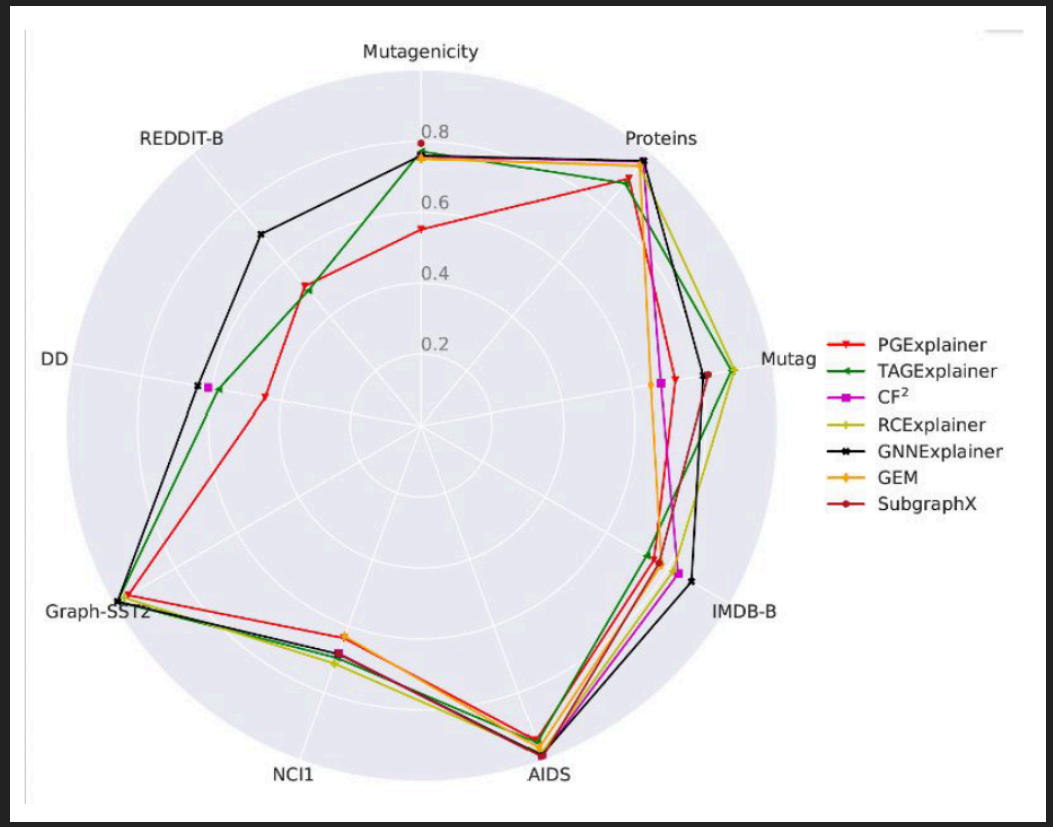
ϕ : GNN

P : #graphs for which $[\phi(G_S) = \phi(G)]$

N : Total #graphs

$$Sufficiency = P/N$$

- Factual : Higher is better
- GNNExplainer and RCExplainer outperform all other explainers.

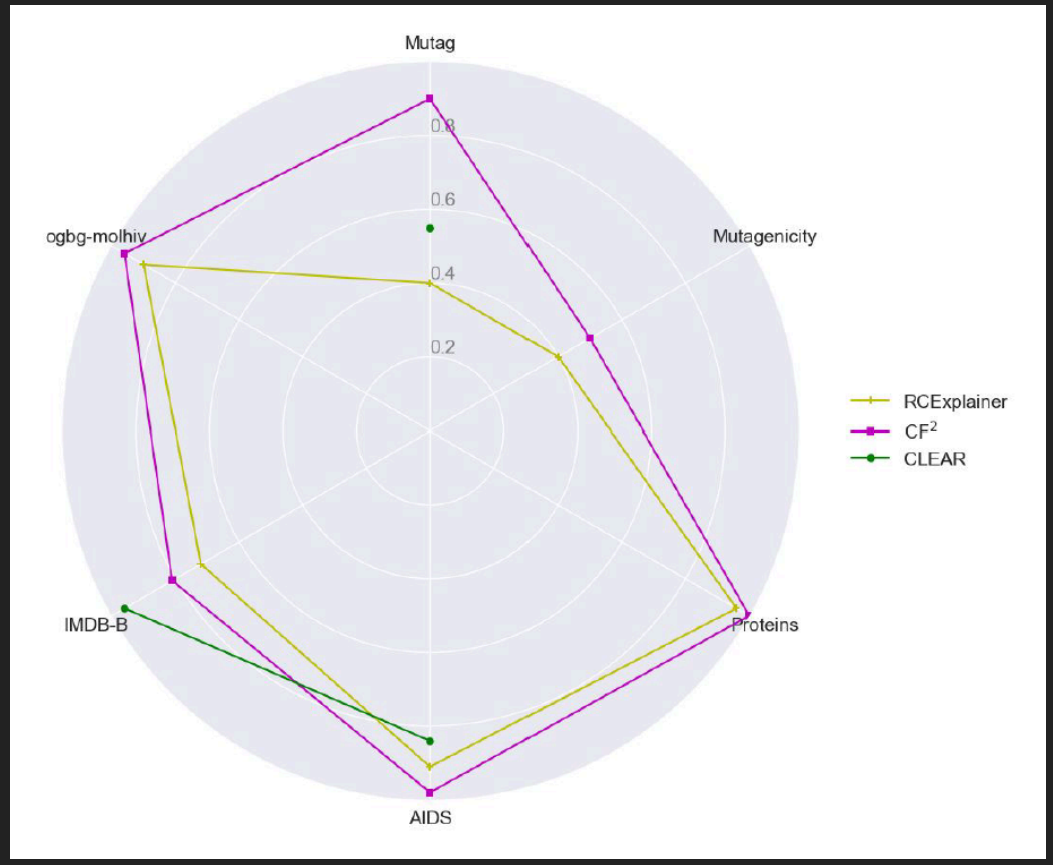


B. Counterfactual

G : Graph
G_C : Counterfactual of G
ϕ : GNN
P : #graphs for which $[\phi(G_C) = \phi(G)]$
N : Total #graphs

$$Sufficiency = P/N$$

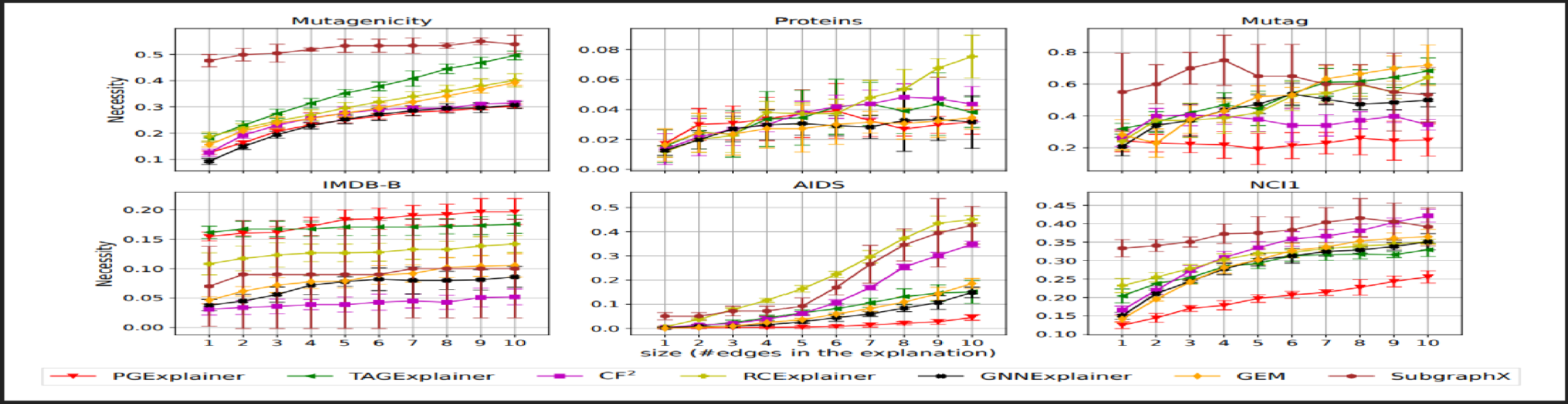
- Counterfactual : Lower is better
- RCExplainer outperform other counterfactual explainers.



Necessity

- Factual explanation are necessary if the removal of the explanation subgraph from the graph results in Counterfactual graph.

$$Necessity(N) = \frac{\sum_{i=1}^{|G|} 1(\Phi(R^i) \neq \Phi(G^i))}{|G|}$$



Stability

$G(V, E)$: Graph
V : Vertex set
E : Edge set
E_X : Set of edges in orginal explanations
E'_X: Set of edges in explanation after variation

- Jaccard similarity

$$Stability = \frac{|E_X \cap E'_X|}{|E_X \cup E'_X|}$$

- Higher is better

A. Optimization Stochasticity:

- Explainers are deep learning models
- Model parameters

	PGExplainer			TAGExplainer			CF^2			RCExplainer			GNNExplainer		
Dataset / Seeds	$1vs2$	$1vs3$	$2vs3$	$1vs2$	$1vs3$	$2vs3$	$1vs2$	$1vs3$	$2vs3$	$1vs2$	$1vs3$	$2vs3$	$1vs2$	$1vs3$	$2vs3$
Mutagenicity	0.69	0.75	0.62	0.76	0.78	0.74	0.77	0.77	0.77	0.75	0.71	0.71	0.46	0.47	0.47
Proteins	0.38	0.51	0.38	0.55	0.48	0.46	0.34	0.34	0.35	0.88	0.85	0.91	0.28	0.28	0.28
Mutag	0.5	0.54	0.51	0.36	0.43	0.72	0.78	0.79	0.79	0.86	0.92	0.87	0.57	0.57	0.58
IMDB-B	0.67	0.76	0.67	0.67	0.60	0.56	0.32	0.32	0.32	0.75	0.73	0.70	0.18	0.19	0.18
AIDS	0.88	0.87	0.82	0.81	0.83	0.87	0.85	0.85	0.85	0.95	0.96	0.97	0.80	0.80	0.80
NCI1	0.58	0.55	0.64	0.69	0.81	0.65	0.60	0.60	0.60	0.71	0.71	0.94	0.44	0.44	0.44

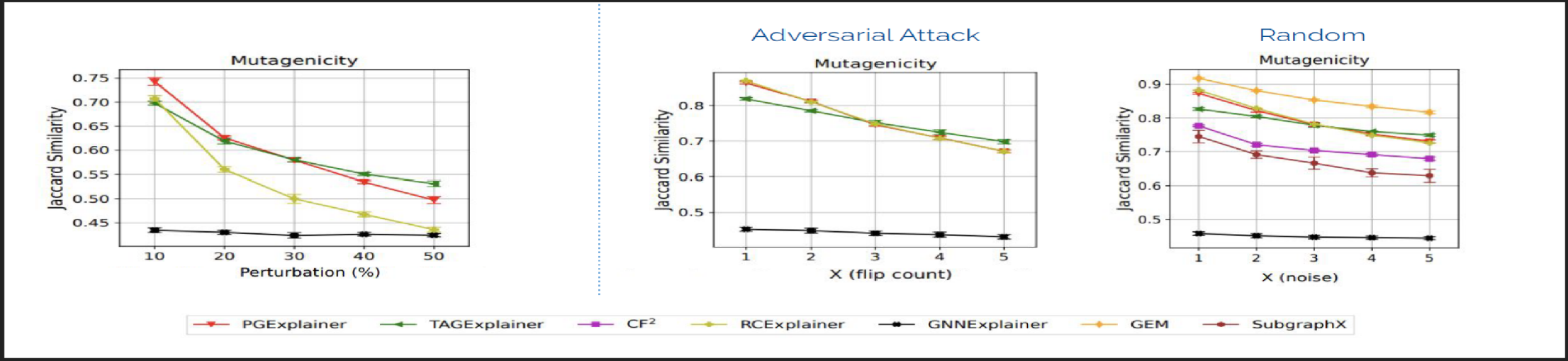
- RCExplainer is the most stable

B. GNN Architecture:

- PGExplainer and RCExplainer are the most stable

	PGExplainer			TAGExplainer			CF^2			RCExplainer			GNNExplainer		
Dataset / Architecture	GAT	GIN	SAGE	GAT	GIN	SAGE	GAT	GIN	SAGE	GAT	GIN	SAGE	GAT	GIN	SAGE
Mutagenicity	0.63	0.65	0.60	0.24	0.25	0.32	0.52	0.47	0.54	0.56	0.52	0.46	0.43	0.42	0.43
Proteins	0.22	0.47	0.38	0.45	0.41	0.18	0.28	0.28	0.28	0.37	0.41	0.42	0.28	0.28	0.28
Mutag	0.57	0.58	0.69	0.60	0.65	0.64	0.58	0.56	0.62	0.47	0.76	0.54	0.55	0.57	0.55
IMDB-B	0.48	0.45	0.56	0.44	0.35	0.47	0.17	0.23	0.17	0.30	0.33	0.26	0.17	0.17	0.17
AIDS	0.81	0.85	0.87	0.83	0.83	0.84	0.80	0.80	0.80	0.81	0.85	0.81	0.8	0.8	0.8
NCI1	0.39	0.41	0.37	0.45	0.17	0.58	0.37	0.38	0.38	0.49	0.53	0.52	0.37	0.38	0.39

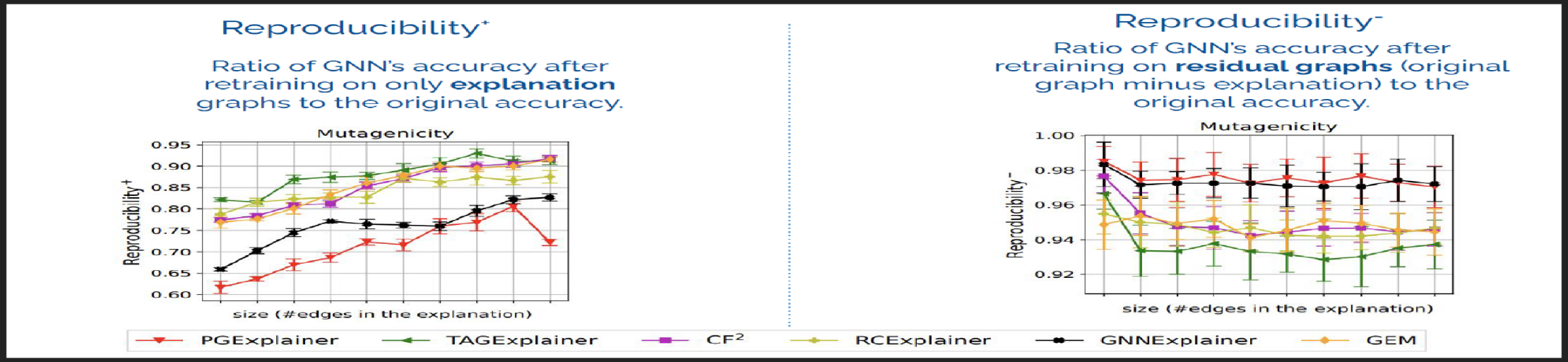
C.(i) Feature Perturbance (ii) Topological Perturbance



- GEM, PGExplainer and RCExplainer are the most stable.
- But, significantly stability issue exists in all explainer.

Reproducibility

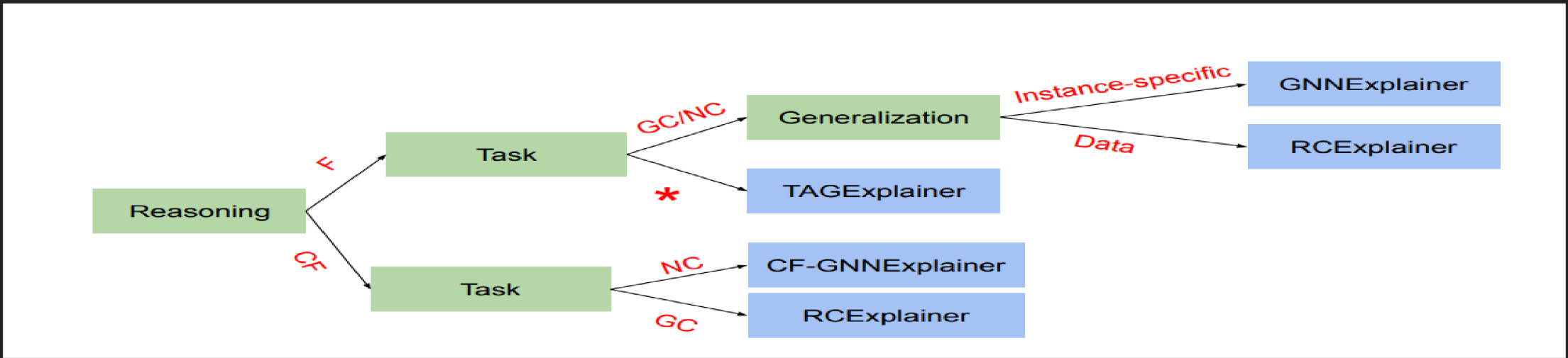
- How well does the explainer explain the model vs the underlying data?



- High reproducibilty demonstrates that the explainers hardly capture the real cause of the GNN predictions.

Key Findings and Recommendation

- RCExplainer shows superior performance in most cases.
- RCEXplainer is consistently the most stable explainer.
- Most Explainer suffer from significant deviations.
- Explainers only captures specific signals learned by the GNNs.
- They do not compasses all underlying data signals.



References

- Link to GitHub repo for the code
- Main papers:
  1. Mert Kosan et al; GNNX-BENCH: UNRAVELLING THE UTILITY OF PERTURBATION-BASED GNN EXPLAINERS THROUGH IN-DEPTH BENCHMARKING

Questions?

Subhankar Mishra Lab Weekly Talks