Explaining Graph Neural Networks

AMLD

March 28,2022

Kenza AMARA

ETH AI CENTER

Application: Financial systems

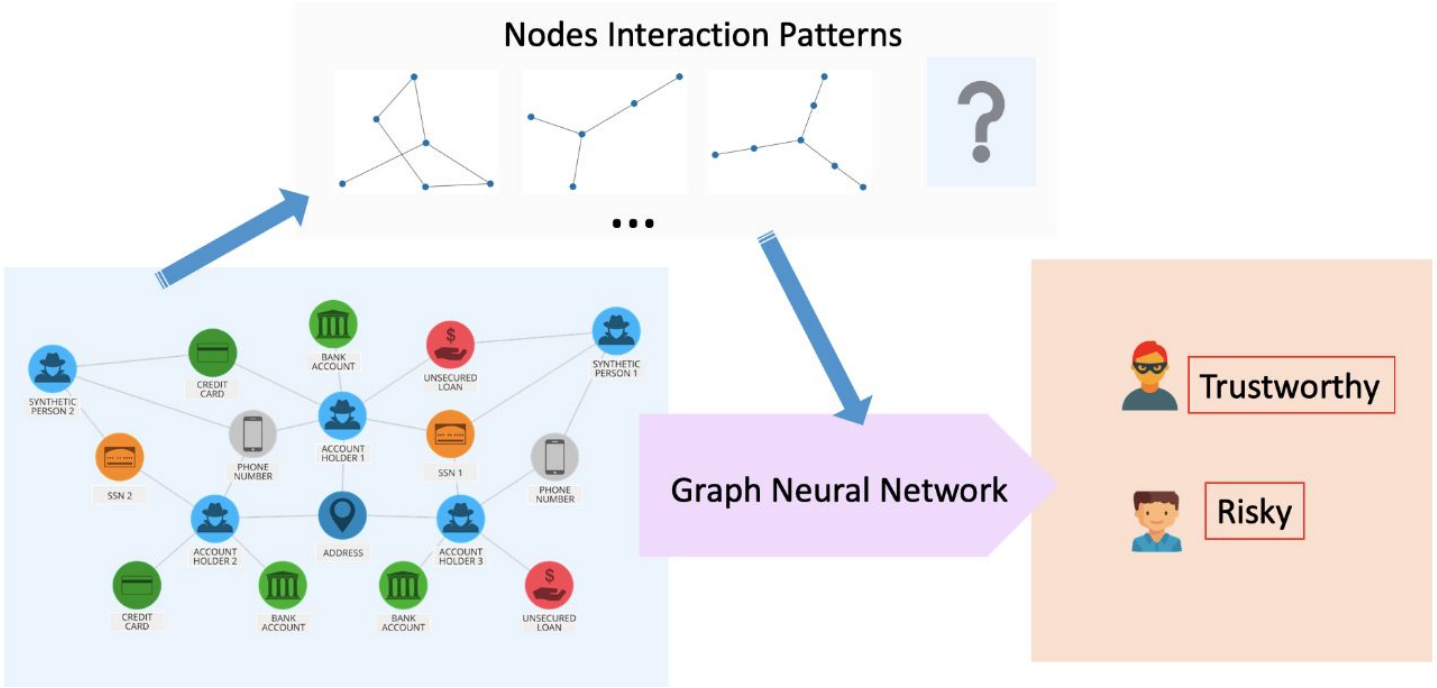


*Image copied from: Li, Xiaoxiao, Joao Saude, Prashant Reddy, and Manuela Veloso. n.d. “Classifying and Understanding Financial Data Using Graph Neural Network.”

Explanation for Graph Neural Networks (GNNs)

= subgraph from the computation graph, with subset of node features.
= mask on nodes/edges/node features

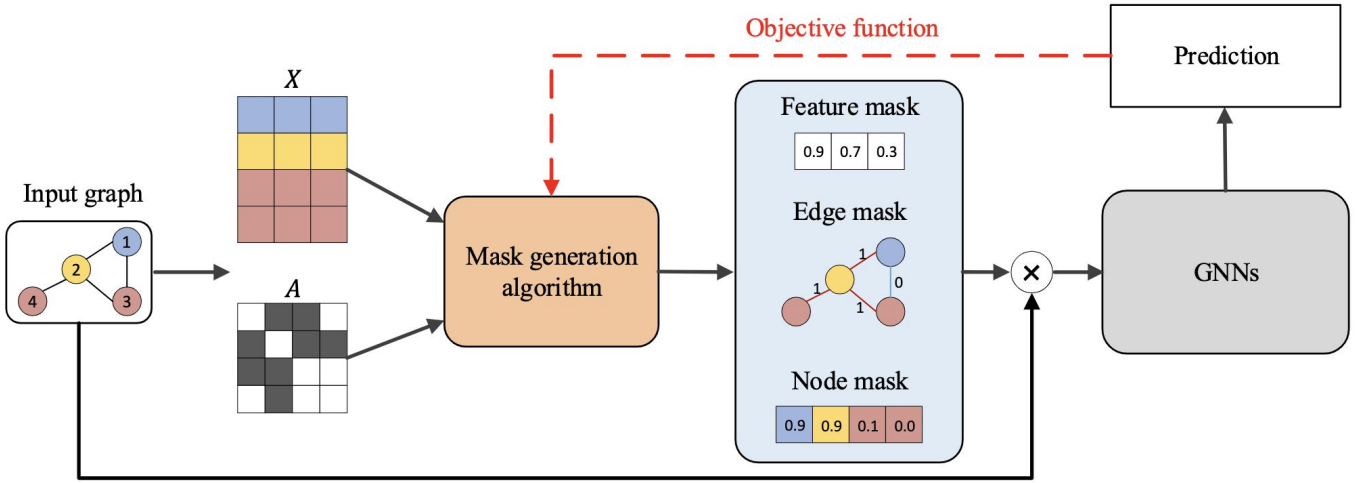


*Image copied from: Yuan, Hao, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2020. “Explainability in Graph Neural Networks: A Taxonomic Survey.”

Taxonomy of explanations

Intrinsic explanation: model/algo structure already understandable, analyse inner workings of the model
For: *linear regression, GLMs, decision trees*

Post-hoc explanation: does not presume any knowledge of the model structure
For: *neural networks*

Model-aware: look inside the model, to analyse where the model puts its attention

gradient/feature-based methods, decomposition methods

Model-agnostic: model is a black-box, only study changes in the output when perturbing the input

perturbation-based methods, counterfactual explanations, surrogate models

Categories of explainers (23)

Model-aware methods

Gradient/features-based methods

- **SA**, **IG,** Guided BP
- CAM, Grad-CAM

Decomposition-based methods

- EB, Contrastive EB
- LRP, GNN-LRP

Model-agnostic methods

Direct-masking procedures

- **GNNExplainer**
- PGExplainer, GraphMask, Refine
- Gem

Search-based heuristics

- Causal Screening, ZORRO
- **SubgraphX**

Surrogate methods

- GraphLIME
- RelEx
- **PGM-Explainer**

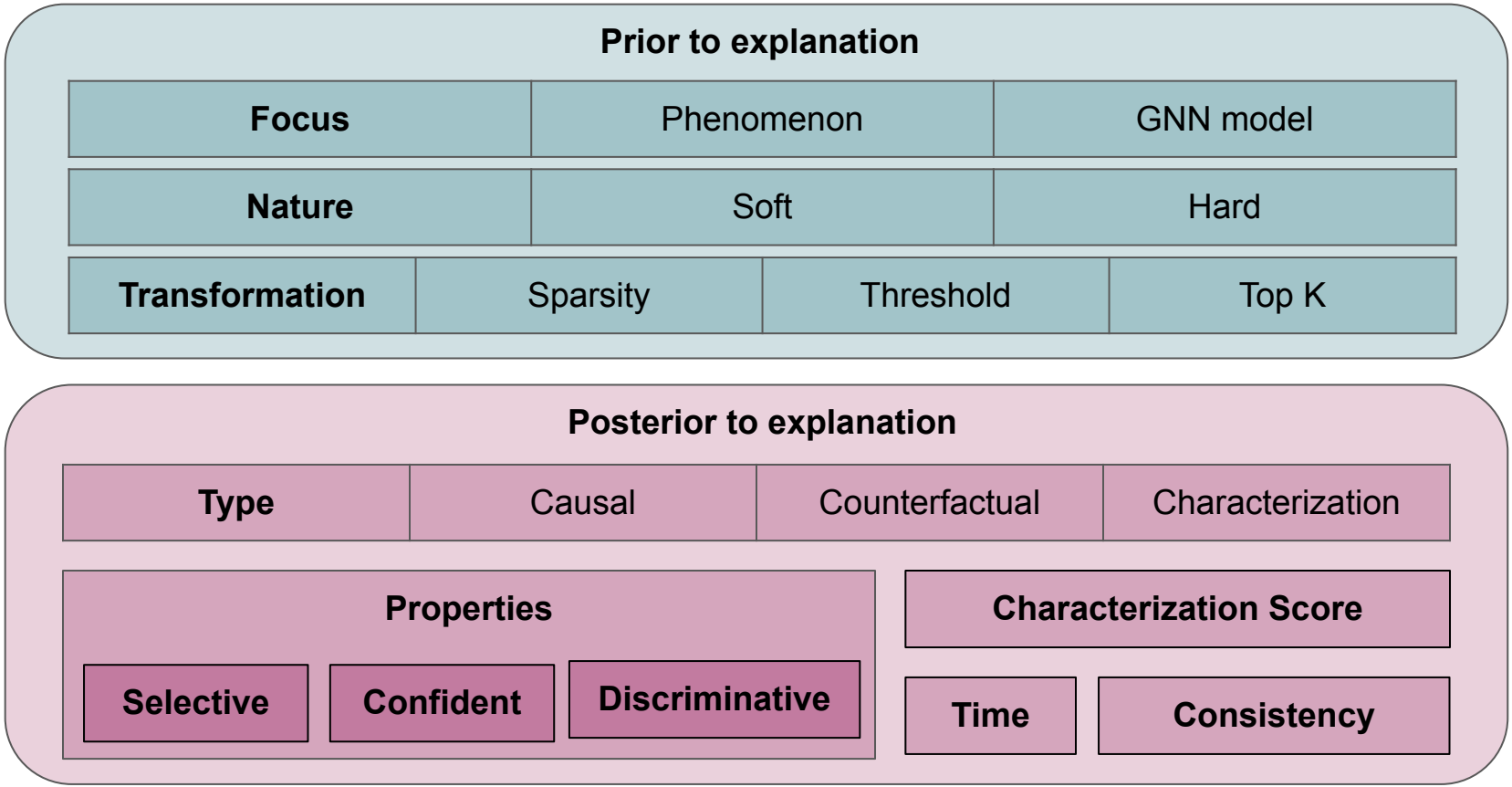
Counterfactual methods

- CF-Explainer
- STEEX

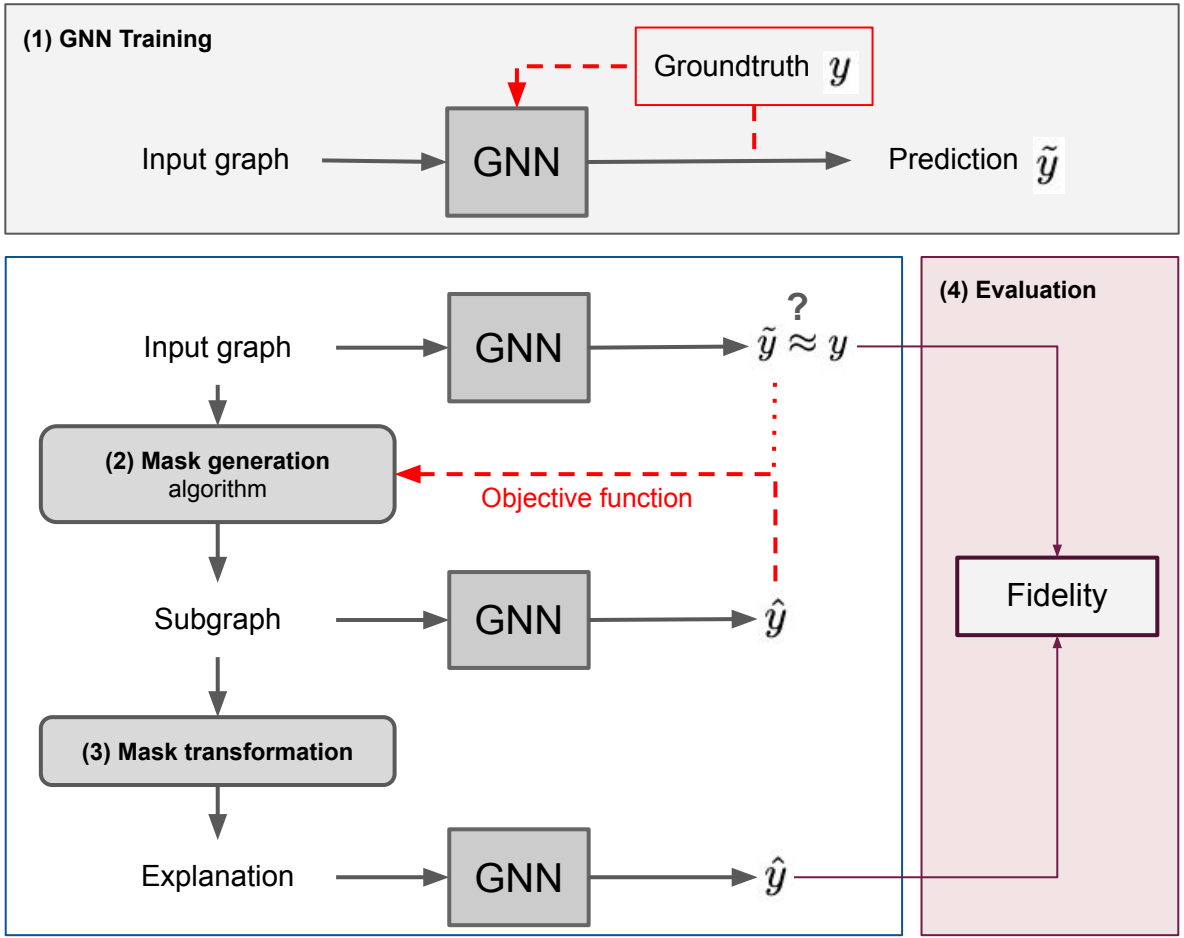
Iterative RL-based methods

- XGNN
- RG-Explainer

How to explain a GNN?



Protocol

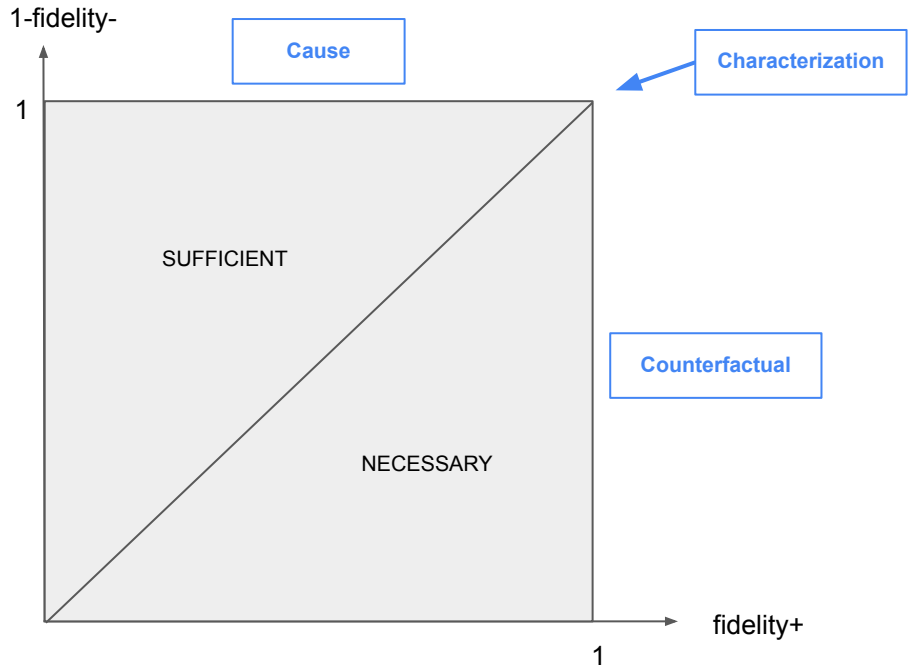

(1) GNN Training
Groundtruth y
Input graph
GNN
Prediction $\tilde{y}$
(4) Evaluation
Input graph
GNN
$\tilde{y} \approx y$?
(2) Mask generation algorithm
Objective function
Subgraph
GNN
$\hat{y}$
Fidelity
(3) Mask transformation
Explanation
GNN
$\hat{y}$

What type of explanation?

- **Causal explanation**:
  sufficient but not necessary
  Fidelity- → 0
- **Counterfactual explanation**:
  necessary but not sufficient
  Fidelity+ → 1
- **Characterisation**:
  necessary AND sufficient
  Fidelity+ → 1 & Fidelity- → 0

Characterization power

$$Charact = \frac{2 \times Fidelity+^{prob} \times (1 - Fidelity-^{prob})}{Fidelity+^{prob} + (1 - Fidelity-^{prob})}$$



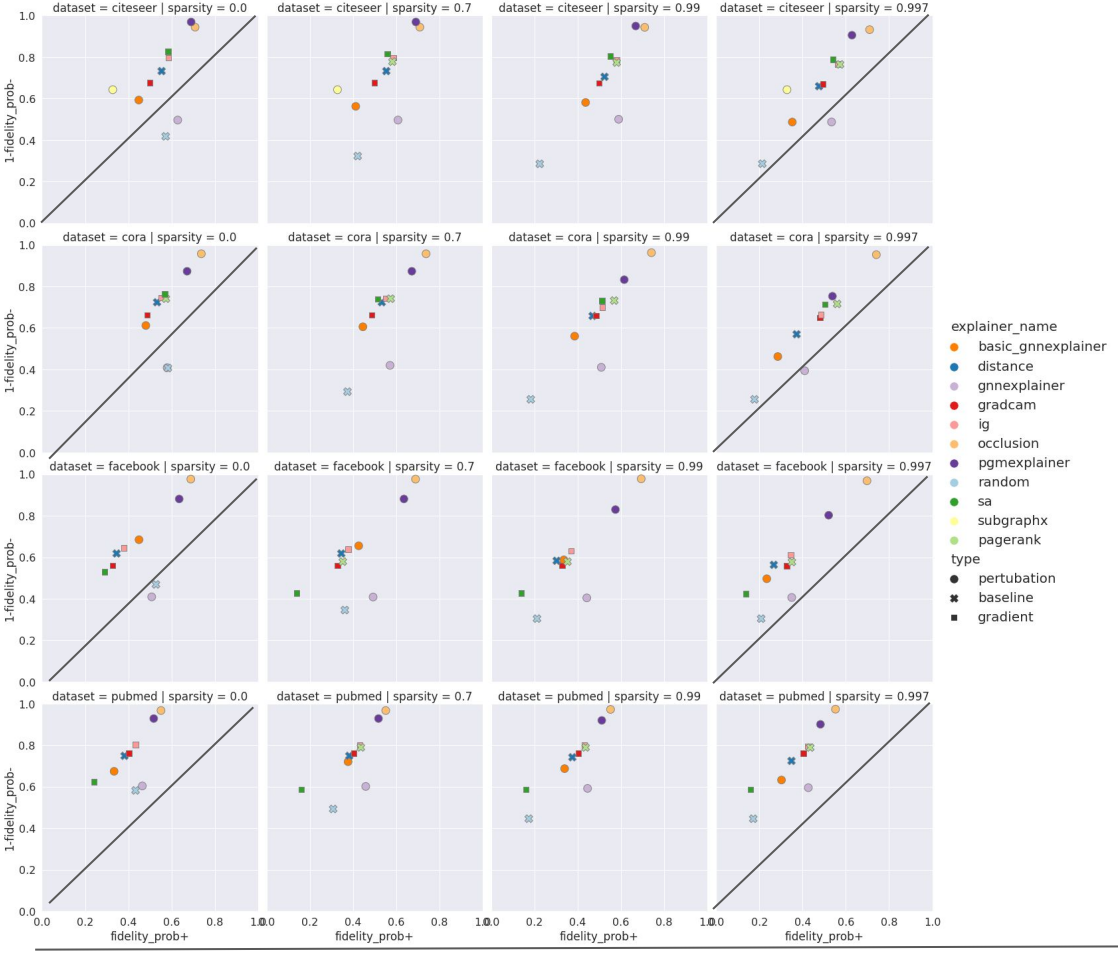
Type of explanations: Causal, counterfactual, characterization

CiteSeer

Cora

Facebook

PubMed



sparsity

Comparative Framework of Explainability Methods

	Time	Characterization power		Selective	Discriminative	Confident	Ranking ability
Methods		Soft	Hard				
Random	+	−	−	−	−	+-	-
Distance	+	-	+	−	−	-	+
PageRank	+	+	++	−	-	−	++
Saliency	+	-	-	+	+	+-	+
Integrated Gradient	+	+	++	−	+	-	+
GradCAM	+	-	++	++	++	-	++
Occlusion	-	++	++	+	+	++	++
GNNExplainer (E)	-	-	+-	+	++	-	-
GNNExplainer (E+NF)	-	-	+	+	++	-	+
PGMExplainer	−	++	+	+	-	+	+

Thank you for your attention

Questions?