On Explainability of Graph Neural Networks via Subgraph Explorations

Malika Sanhinova

GDLMA Seminar,
Technische Universität München, Germany
10.05.2022

Content

- Motivation
- Methodology
- Experiments
- Comparison with other methods
- Take Home Message
- Discussion

Motivation

Explainability in medical applications:

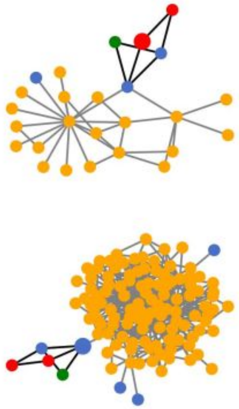
- Prevent misdiagnosis
- Reason relationships behind predictions
- Understand underlying concepts of the data
- Interpret domain-specific results

news.mit.edu

Motivation

Explainability of the GNNs:

- Consider important structural data
- Importance of the nodes does not directly imply importance of a subgraph
- Identify graph substructures directly
- Subgraphs explanations are more human-intelligible



General idea of the SubgraphX

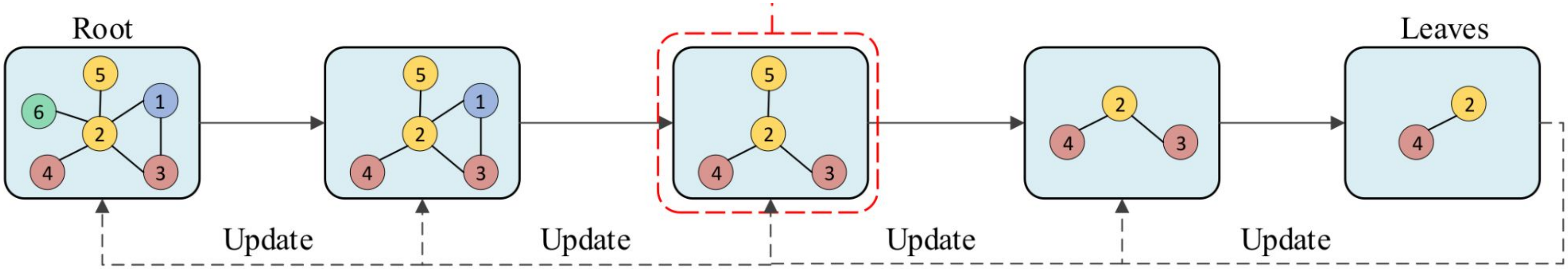
Find the most important subgraph for the prediction *y*

Monte Carlo Tree Search:
Explore different subgraphs

Shapley Value:
evaluate the importance of every subgraph

Methodology: SubgraphX

Monte Carlo Tree Search:



Methodology: SubgraphX

Shapley value: adaptation from game theory

- GNN predictions are the game gain
- Different subgraphs are players
- Each subgraph ‘plays’ against the other individual nodes
- While the nodes form all possible coalitions
- Guarantees correctness and fairness of the explanations

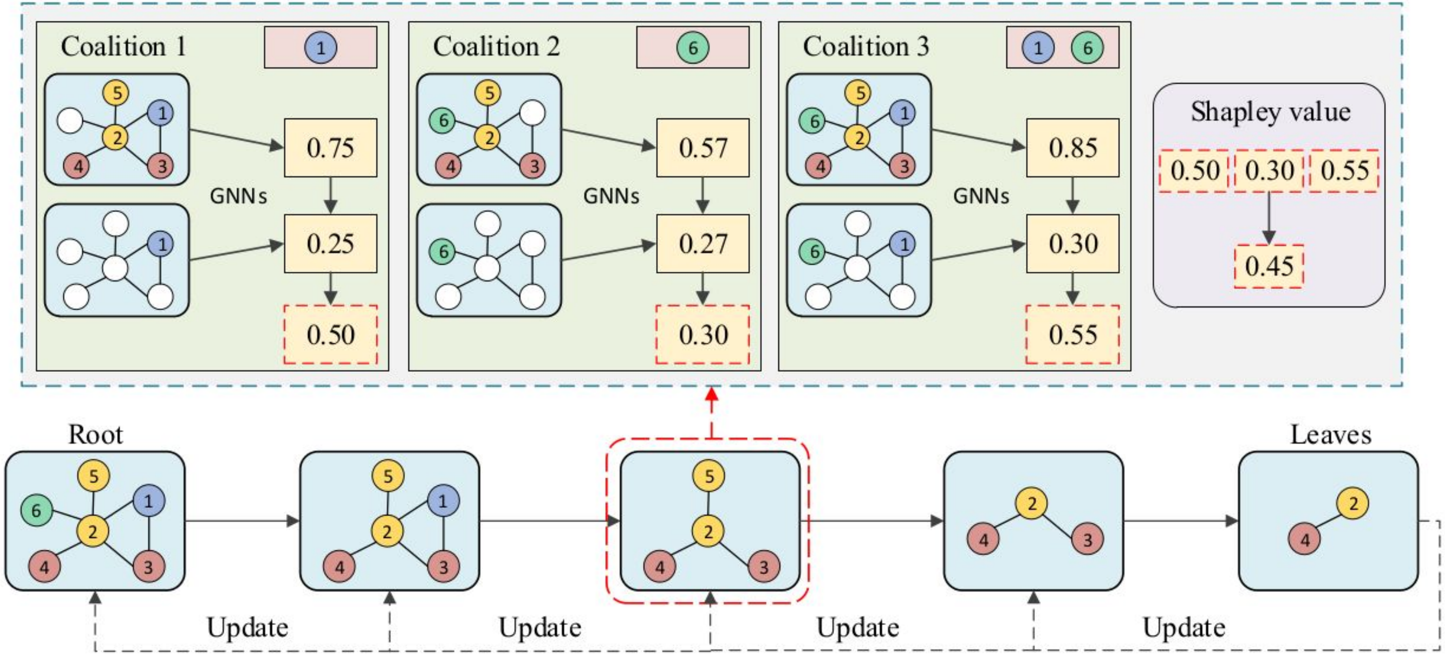
$$\phi(\mathcal{G}_i) = \sum_{S \subseteq P \setminus \{\mathcal{G}_i\}} \frac{|S|!\,(|P| - |S| - 1)!}{|P|!} m(S, G_i),$$

$$m(S, \mathcal{G}_i) = f\left(S \cup \{\mathcal{G}_i\}\right) - f(S),$$

Difference of predictions **with** and **without** the coalition set *S*

Methodology: SubgraphX

Shapley value and coalition formation:



Methodology: SubgraphX

Problem: Shapley value enumerates all possible coalitions -> not efficient

Solution: Only consider the ***neighbouring*** nodes

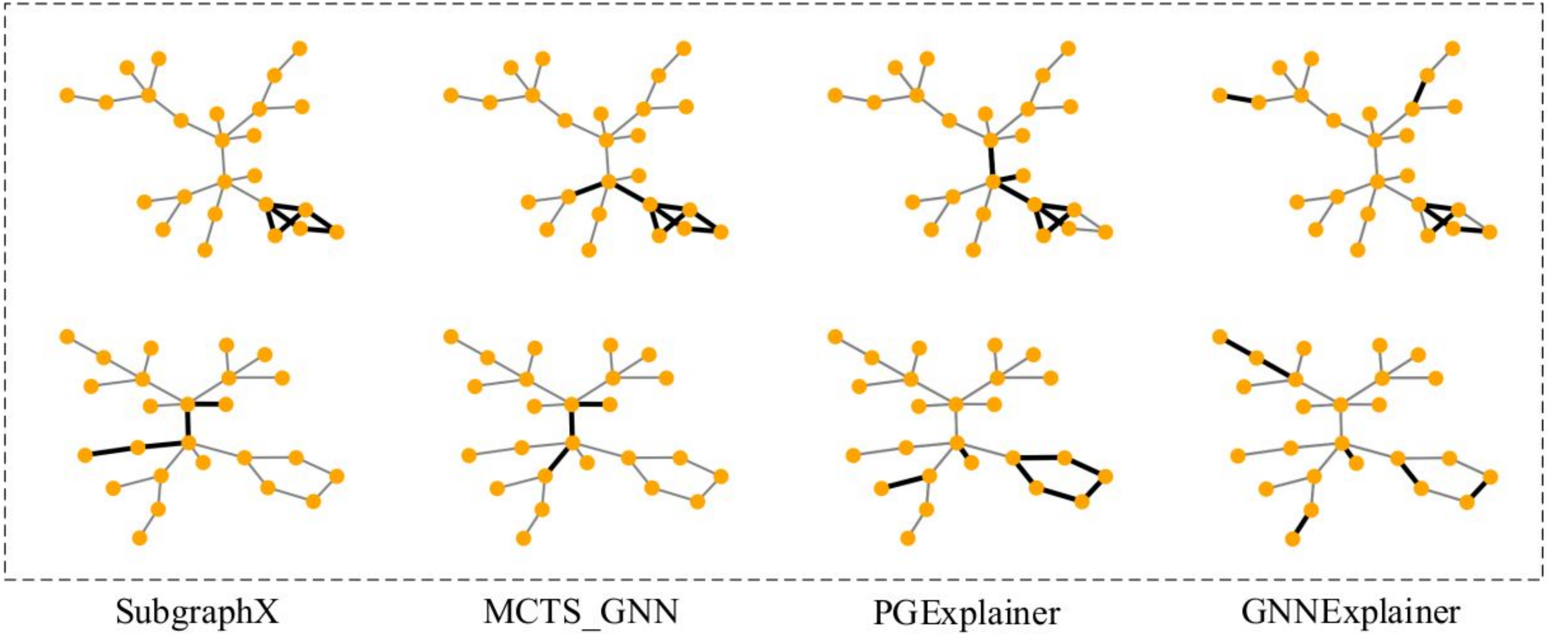
Problem: Different nodes have variable number of neighbours

Solution: Sampling!

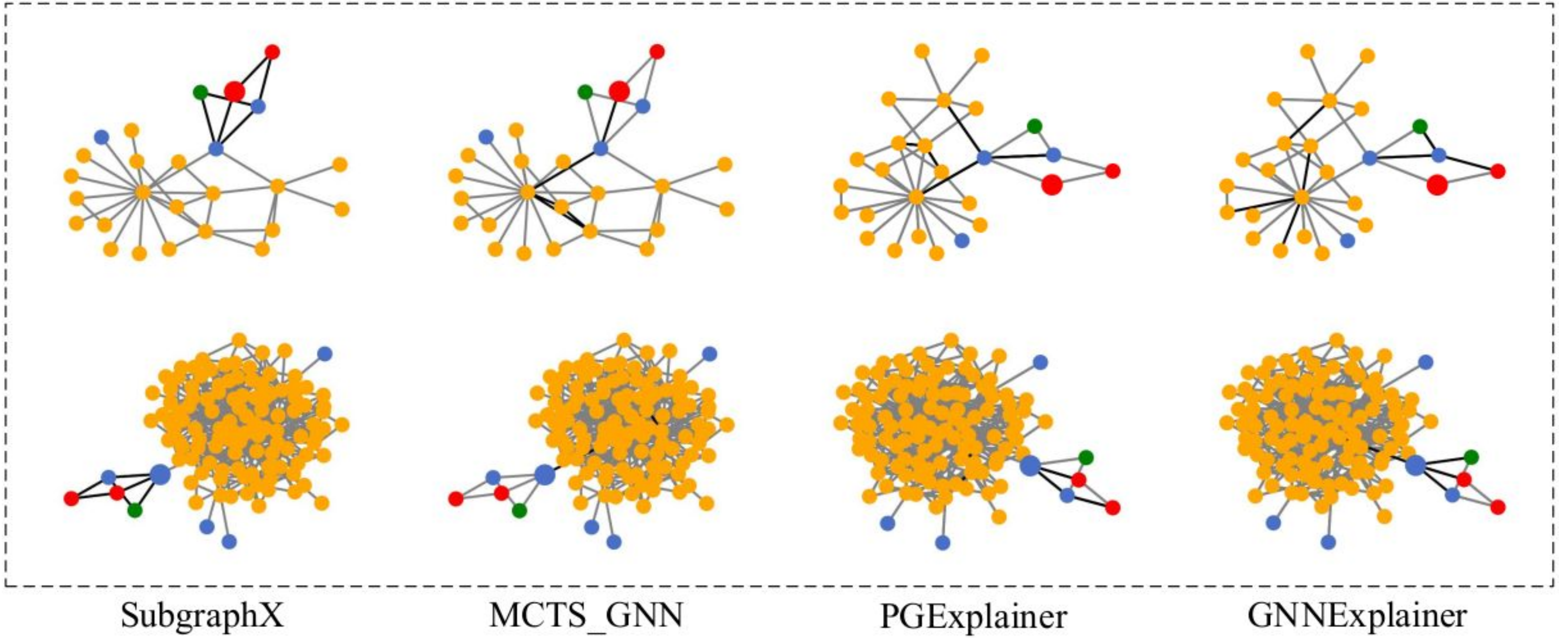
Namely, ***Monte Carlo sampling***

$$\phi(\mathcal{G}_i) = \frac{1}{T}\sum_{t=1}^{T}(f\left(S_i \cup \{\mathcal{G}_i\}\right) - f(S_i))$$

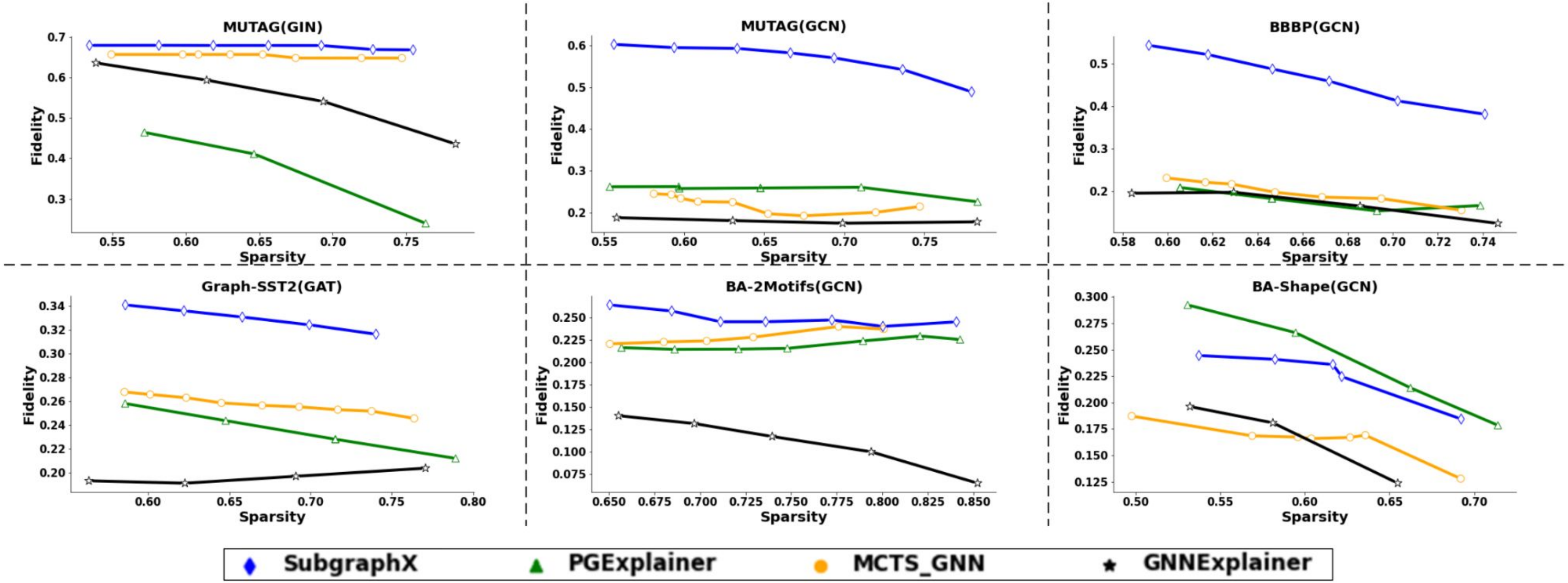
Experimental results: graph classification



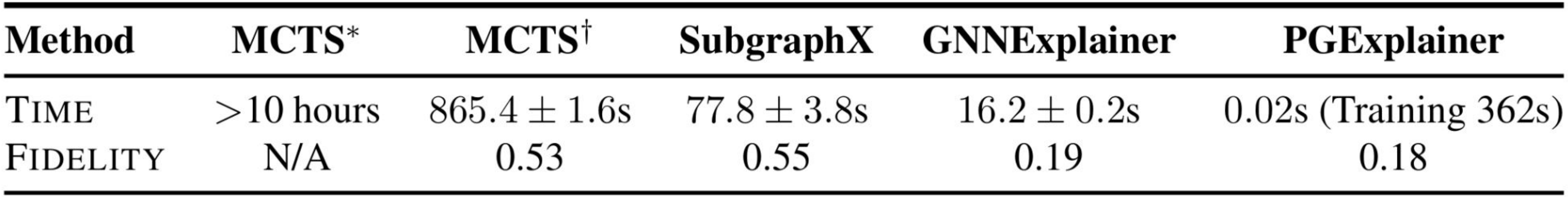
Experimental results: node classification



Experimental results: quantitative studies



Experimental results: computational efficiency



Method	MCTS*	MCTS†	SubgraphX	GNNExplainer	PGExplainer
TIME	>10 hours	865.4 ± 1.6s	77.8 ± 3.8s	16.2 ± 0.2s	0.02s (Training 362s)
FIDELITY	N/A	0.53	0.55	0.19	0.18

Take Home Message

- Subgraph explanation is more intuitive and human-intelligible
- Subgraphs are more informative than individual nodes
- SubgraphX can be used for graph classification, node classification and link prediction
- SubgraphX treats GNN as a black box
- Efficiency is achieved by sampling the node space

Discussion

- **Multiple disconnected subgraphs**
- **Relies only on visualization (e.g. not on features)**
- **Consider GNN to improve the accuracy?**