GStarX: Explaining Graph Neural Networks with Structure-Aware Cooperative Games

Shichang Zhang[1], Yozen Liu[2], Neil Shah[2], Yizhou Sun[1]

[1]University of California, Los Angeles (UCLA) [2]Snap Inc.

Feb 2023

Snap Inc.

Explainable Artificial Intelligence (XAI)

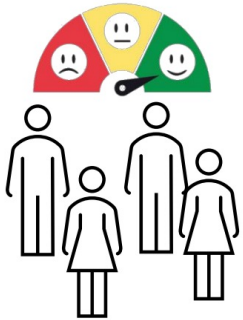
Many of the AI models are neural-network-based black-box

- Explainability is critical for AI models
- Explainability helps to increase user trust and improve model design

Business

Healthcare

Finance



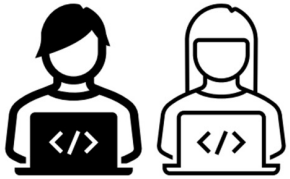
End users

Explain

Model

Explain

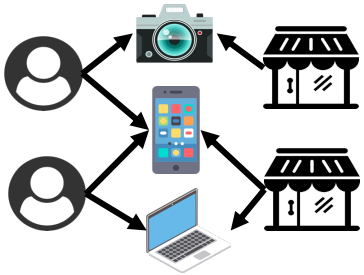
Improve



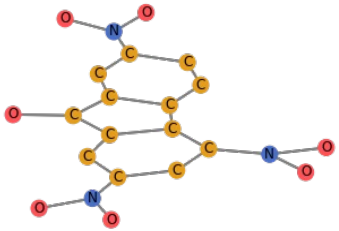
Developers

Learning on Graphs

Graphs are a general language for modeling entities with relations

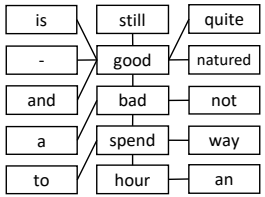


E-commerce graphs



Molecule graphs

"is still quite good – natured and not a bad way to spend an hour"



Text graphs

Transportation graphs, code graphs, and many more …

Graph learning tasks

- Classify molecular properties
- Classify sentence sentiment

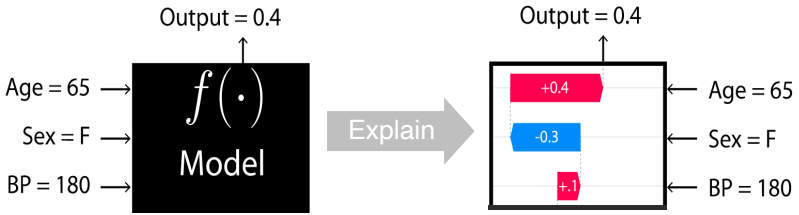
Graph Neural Networks (GNNs) are the SOTA model

Model Explanation

Given a trained black-box model, identify important features for a model prediction

Tabular data

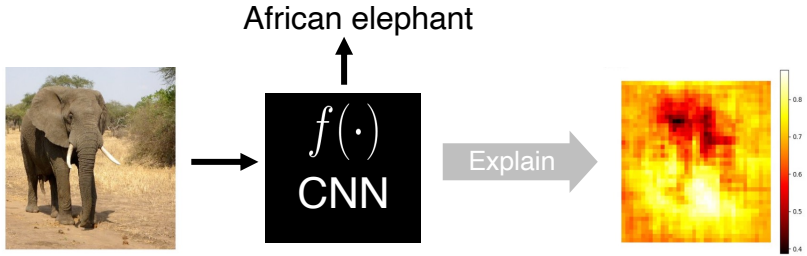
Features are attributes



(Lundberg, S. M., & Lee, S. I. NeurIPS 2017)

Image data

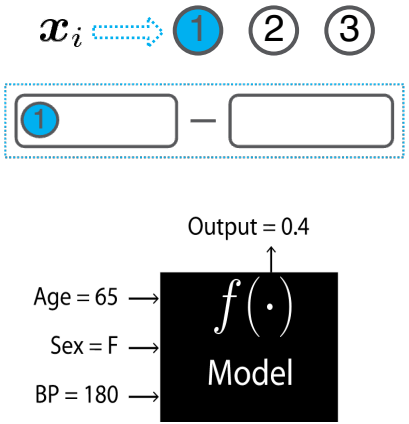
Features are pixels



(Zeiler, M. D., & Fergus, R. ECCV 2014)

Model Explanation as Feature Importance Scoring

	Model explanation (multi-class classification)
$\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$	Features
$f(\cdot) : \boldsymbol{x_S} \to \mathbb{R}$ for $\boldsymbol{x_S} \subseteq \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$	A model. Outputs the predicted probability (of the most likely class) for a set of features
$\textsc{Score}(f(\cdot), i)$	What is a proper importance score for each feature?



- Contribution of the target feature as its importance score

  $\textsc{Score}(f(\cdot), i) := f(\{\boldsymbol{x_i}\}) - f(\emptyset)$

A simple idea ignores interactions between features

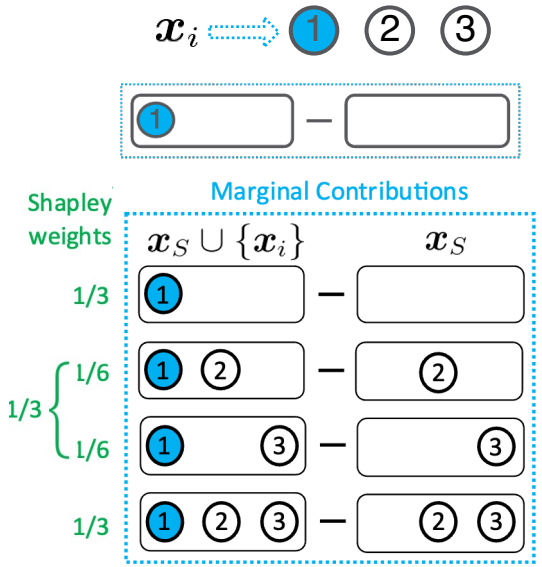
Model Explanation as Feature Importance Scoring

	Model explanation (multi-class classification)	**Cooperative game**
$\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$	Features	Players
$f(\cdot) : \boldsymbol{x_S} \to \mathbb{R}$ for $\boldsymbol{x_S} \subseteq \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$	A model. Outputs the predicted probability (of the most likely class) for a set of features	A payoff function. Outputs the payoff for a set of players
$\textsc{Score}(f(\cdot), i)$	What is a proper importance score for each feature?	What is a fair payoff to each player?

- Cooperative game theory, e.g., Shapley value
  - A weighted aggregation of *marginal contributions* $m(i, S)$

$$m(i, S) := f(\boldsymbol{x_S} \cup \{\boldsymbol{x_i}\}) - f(\boldsymbol{x_S})$$

$$\textsc{Score}(f(\cdot), i) := \overbrace{\frac{1}{n}\sum_{k=0}^{n-1}}^{\text{Average over } k} \overbrace{\frac{1}{\binom{n-1}{k}} \sum_{\substack{S \subseteq N \setminus \{i\} \\ |S|=k}}}^{\text{Average over } S \text{ s.t. } |S|=k} m(i, S)$$

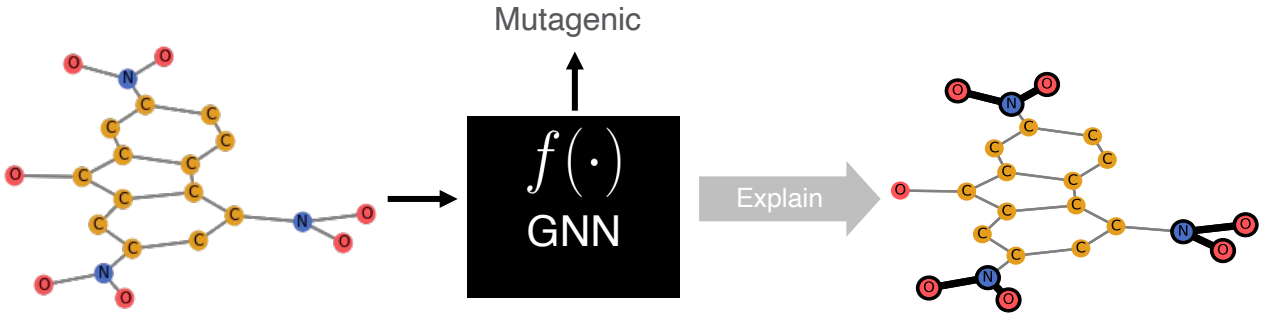


GNN Explanation on Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

Feature importance scoring on graphs

- Nodes as features (like tabular attributes or image pixels as features)
- The graph structure between nodes contains important information

Find an optimal subgraph (set of nodes and the structure between them)

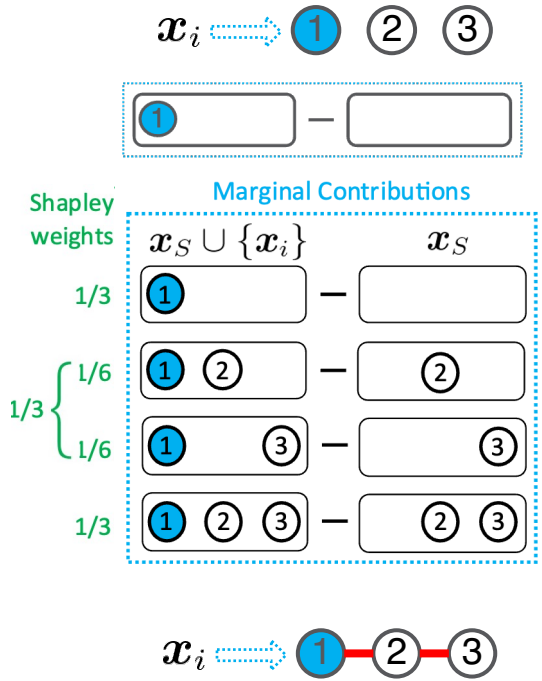


Feature Importance Scoring for GNNs on Graphs

- Feature contribution $\textsc{Score}(f(\cdot), i) := f(\{\boldsymbol{x}_i\}) - f(\emptyset)$
- Shapley value

$$m(i, S) := f(\boldsymbol{x}_S \cup \{\boldsymbol{x}_i\}) - f(\boldsymbol{x}_S)$$

$$\textsc{Score}(f(\cdot), i) := \overbrace{\frac{1}{n}\sum_{k=0}^{n-1}}^{\text{Average over } k} \overbrace{\frac{1}{\binom{n-1}{k}} \sum_{\substack{S \subseteq N \setminus \{i\} \\ |S|=k}}}^{\text{Average over } S \text{ s.t. } |S|=k} m(i, S)$$



Both score functions are not structure aware

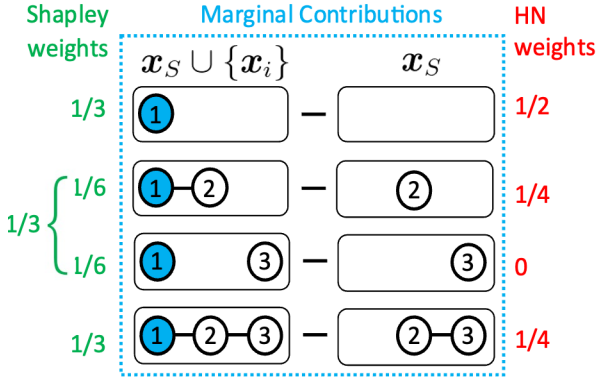
$$\textsc{Score}(f(\cdot), \mathcal{G}, i) := \quad ?$$

HN value: A Structure-aware Game Theory Value

A structure-aware HN value

$$\textsc{Score}(f(\cdot), \mathcal{G}, i) := \lim_{t\to\infty} f_\tau^t(\{\boldsymbol{x}_i\})$$

HN equals to Shapley for complete graphs



HN value: A Structure-aware Game Theory Value

A structure-aware HN value

$$\textsc{Score}(f(\cdot), \mathcal{G}, i) := \lim_{t\to\infty} f_\tau^t(\{\boldsymbol{x}_i\})$$

$f_\tau^t(\cdot)$ is computed recursively over $\boldsymbol{x}_S$

- Base case

  $$f_\tau^t(\boldsymbol{x}_S) = f(\boldsymbol{x}_S) \text{ when } t = 0$$

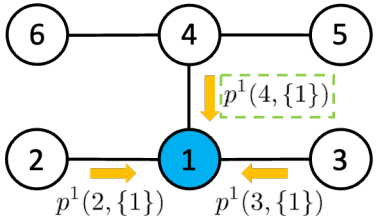
- Recursive case

  $$f_\tau^t(\boldsymbol{x}_S) = f_\tau^{t-1}(\boldsymbol{x}_S) + \tau \sum_{j\in\mathcal{N}(\boldsymbol{x}_S)} p^{t-1}(j, S)$$

  - Cooperation surplus

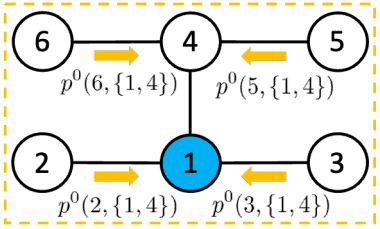
    $$p^t(j, S) := f^t(\boldsymbol{x}_S \cup \{\boldsymbol{x}_j\}) - f^t(\boldsymbol{x}_S) - f^t(\{\boldsymbol{x}_j\})$$

  - Hyperparameter τ *in* [0,1]



$$f_\tau^2(\{1\}) = f_\tau^1(\{1\}) + \tau \sum_{j\in\{2,3,4\}} p^1(j, \{1\})$$

$$p^1(4,\{1\}) = f_\tau^1(\{1,4\}) - f_\tau^1(\{1\}) - f_\tau^1(\{4\})$$



$$f_\tau^1(\{1,4\}) = f_\tau^0(\{1,4\}) + \tau \sum_{j\in\{2,3,5,6\}} p^0(j, \{1,4\})$$

Experiments: Quantitative Evaluation

Task: Graph classification (top) and node classification (bottom)

GStarX outperforms other baselines in terms of Harmonic fidelity

Synthetic → BA2Motifs

Molecule → BACE, BBBP

Text → GraphSST2, MUTAG, Twitter

Dataset	GNNExplainer	PGExplainer	SubgraphX	GraphSVX	OrphicX	GStarX
BA2Motifs	0.4841	0.4879	**0.6050**	0.5017	0.5087	0.5824
BACE	0.5016	0.5127	0.5519	0.5067	0.4960	**0.5934**
BBBP	0.4735	0.4750	**0.5610**	0.5345	0.4893	0.5227
GraphSST2	0.4845	0.5196	0.5487	0.5053	0.4924	**0.5519**
MUTAG	0.4745	0.4714	0.5253	0.5211	0.4925	**0.6171**
Twitter	0.4838	0.4938	0.5494	0.4989	0.4944	**0.5716**
Average	0.4837	0.4934	0.5569	0.5114	0.4952	**0.5732**

Synthetic →

Dataset	GNNExplainer	PGExplainer	SubgraphX	GraphSVX	OrphicX	GStarX
BAShape	0.4772	0.5042	0.6050	0.4916	0.5081	**0.5321**

Summary

- Solving an important model explanation problem on graphs
- Model explanation can be formulated as feature importance scoring
- Existing importance scoring functions are not structure aware
- Our approach:
  - A **structure-aware** importance scoring function based-on the HN value
- Our result:
  - Explanations with better fidelity and more intuitive visualizations

Thank you!
Q & A

Paper link

Contact author