XAI on Graphs
Hands-on Tutorial

Dominik Köhler

Kassel - GAIN, October 11, 2023

Introduction

These slides aim to explain:

- XAI in General
- Explainers of GNNS, which are implemented by GraphXAI [1]

Additionally, everybody is encouraged to work with explainers for synthetic datasets during the Hands-on Tutorial.
All code is and stays available:
`https://github.com/mathematiger/Hands_on_GraphXAI`
The Jupyterlabs will stay open until september ends

Table of Contents

Definitions of XAI

- Interpretability is the degree to which a human can **understand the cause** of a decision. [2]
- Interpretability is the degree to which a human can **consistently predict** the models result. [3]
- The **model itself becomes the source of knowledge** instead of the data. Interpretability makes it possible to extract this additional knowledge captured by the model. [4]

Goals of XAI

DARPA [5]: XAI program aims to create a suite of ML techniques that:

- **Produce more explainable models** while maintaining a high level of prediction accuracy
- Enable human users to **understand, appropriately trust and effectively manage** the emerging generation of AI partners

Who wants to achieve what in XAI? [6] [7]

Data Scientists want to **Improve**

- Did the model use all relevant input?
- Feature Engineering
- Detect flaws: Has the model only learned in the expected way?

Decision Makers want to **Justify & Control**

- Develop trust
- Assess regulatory compliance
- Use models appropriately

Governments want **Fairness & Predictability**

- DSGVO/GDPR: Citizens have a right to an explanation
  e.g. healthcare, police searches
- Detect discriminatory patterns

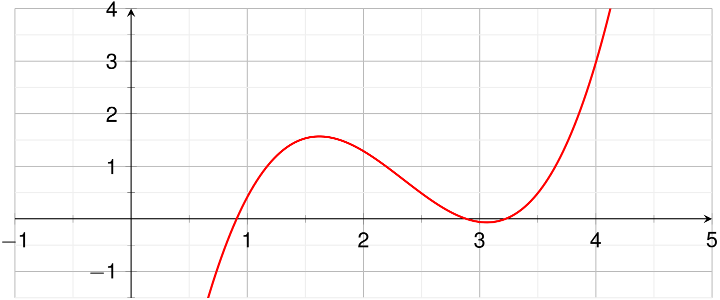
Domain Experts want to **Discover**

- Has the model detected a connection
  that was previously missed out?

What is an explanation?

- Explaining functions in general
- Psychology of explanations
- Why explanations are important

Explaining functions: Discussion of curves (binary classification)



- **Extrema**: What are maxima / minima of the model?
- **Zeros**: Where does the model change labels?
- **Monotonous intervals** (for functions in $\mathbb{R} \to \mathbb{R}$)
- **Approximate**:
  - With a simpler function
  - With the same function on restricted input

Psychology of explanations 1/2 [2] [9]

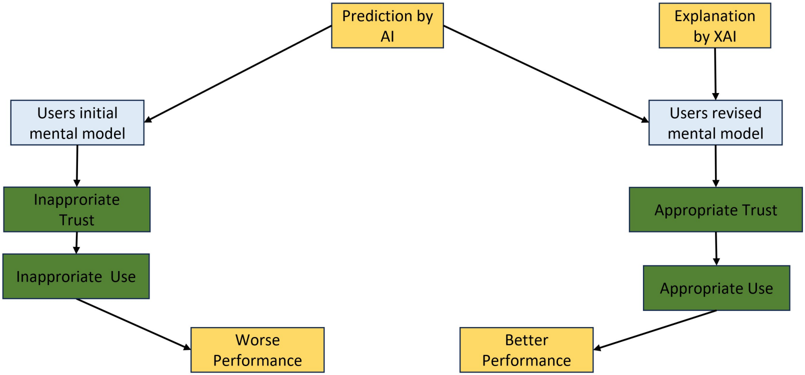
- People often tend to **abductive reasoning**
  - Seeking the simplest & most likely explanation: ”Given our theories T of the world: If A were true, the observation O would be a matter of course”
  - Explanations are incomplete and could be falsified empirically
  - Selecting explanations for correct abductive reasoning is crucial
- **Statistics don’t matter**
  - People cannot think well about statistics (Kahneman [8])
  - People always take the most probable explanation as the best explanation

Psychology of explanations - 2/2

- Explanations should be **contrastive**
  - Why was P predicted and not Q?
- **Social explanations**
  - Only comment, what is truly believed, what is necessary and what is relevant
  - Handling already known information as unnecessary

How explanations improve performance

Figure inspired by [9]



End-user explanation issues with AI

- How do we know, if the user has achieved a pragmatic understanding of the AI? [10]
- What is the mental model of end-users of XAI? Is it in line with how XAI works? [11]
- Do users learn to appropriately use the model, or do they blindly follow / neglect the model? [12]

Explanator as a function [4] [13]

- Input
- Output
- Method of finding the optimal explanation

Input

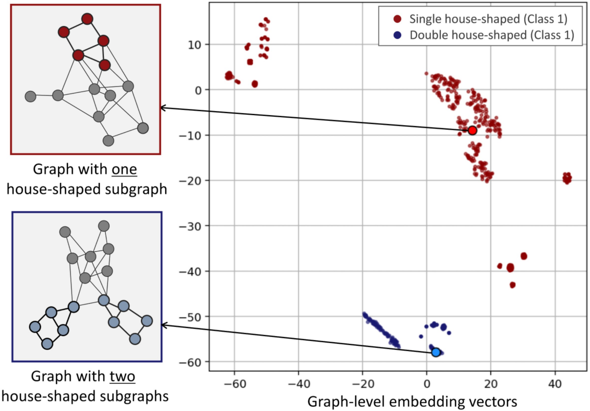
- Post-hoc vs. Intrinsic
  - the model output or the models' input as input
- Locality: Global vs. Instance explanation
  - whole predictions or just one prediction as input
- Portability: Model-specific vs. Model-agnostic
  - The model as input, or just the model output

Output

- Explanatory subgraph
  - As important input for an instance explanation
  - As a global explanation or prototype explanations
  - As counterfactuals
- Surrogate model
- Logical rules/explanations

Prototype Explanations

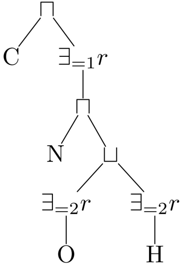
We train a GNN to classify graphs into the number of house-motifs they contain.



From Figure 3 in [14]

Logical Expressions

Example: MUTAG for detecting mutagenic molecules [15] [16]



Class Expression for C with a NH_2 group or a NO_2 group.

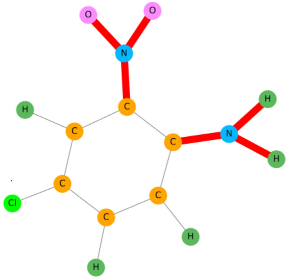
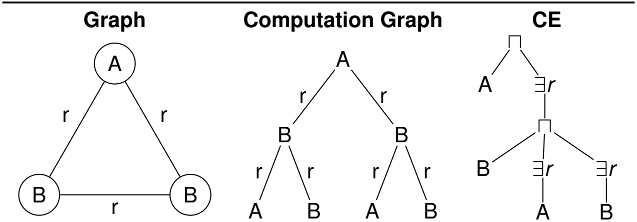


Figure from Figure 5 in [17]

Motivation for using description logic

A GNN computation graph mainly follows a tree structure
$\rightarrow$ we should consider also explaining it with a tree



Finding the best explanation

- Optimization framework
  - Mutual information
  - GNN-Output
  - Backpropagation / Feature Importance
- Regularization
  - Length of explanation
  - Connected subgraphs
  - . . .

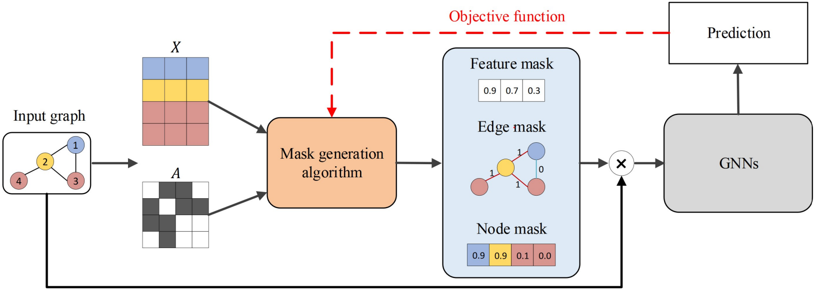
Explaining a GNN

Most common goal: Find a subgraph, which explains a label

- Instance explanations: Masking important input
  - Node masks [16]
  - Edge masks [15]
  - Feature masks [16]
- Global explanations: One explanatory graph for one class/label
  - Select a prototype instance explanation [14]
  - Create synthetic graphs [18]

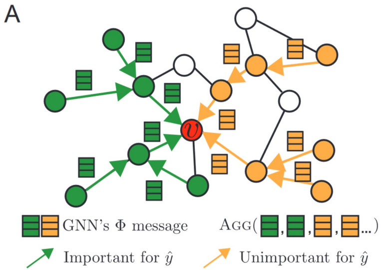
Masking the input data

From Figure 2 in [19]:



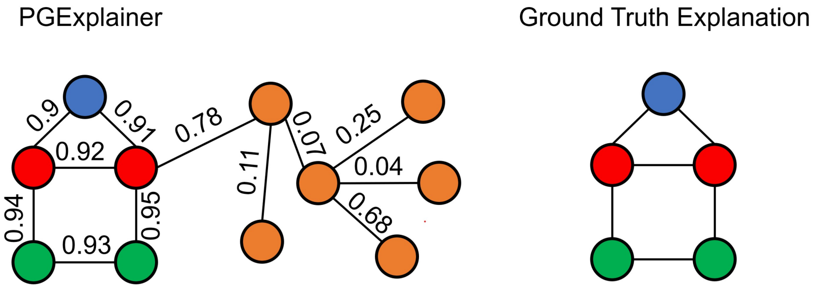
Node masks: GNNExplainer [16]

From Figure 2 in [16]:



Edge masks: PGExplainer [15]

From Figure 2 in [16]:



Obtaining an explanation by removing nodes: SubgraphX [20]

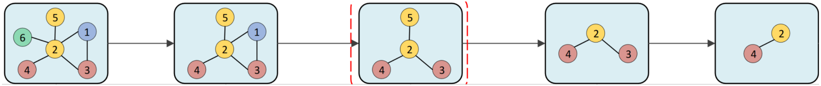
Notation

G_i is subgraph after removing one node, N is the set of nodes not in G_i, f is the GNN

In each step the least contributive node is removed (measured by shapley values)

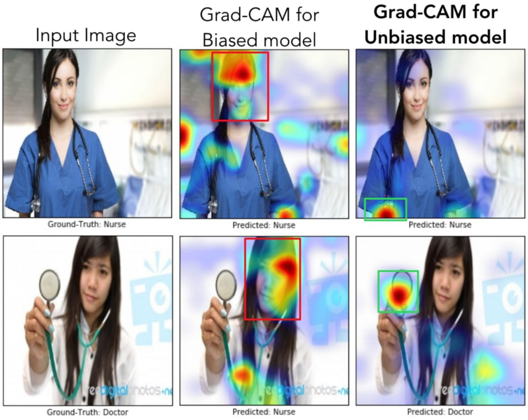
$$\phi(G_i) = \sum_{S \subset N} \frac{|S|!\,(|N| - |S|)\,!}{|N|!}\,(f(S \cup G_i) - f(S))$$

This is how the subgraph contributing the most to the GNN prediction can be found.
From Figure 1 in [20]:



Obtaining a subgraph by gradients: [21] [22] [23]

- Algorithms aggregate gradients in last CNN(GCN)-layer to find relevant input data
- Adopted from GNNs and not specifically tailored to GNNs

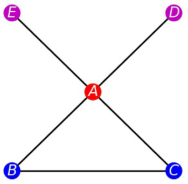


From Figure 8 in [21]

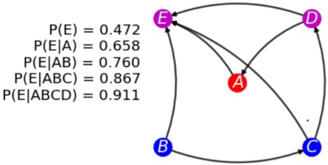
Probabilistic graph model (PGM): [24]

- Surrogate model for node classification
- Trains a bayesian network to capture statistical dependencies, observed by perturbated input

From Figure 1 in [24]:



to-be-explained motif

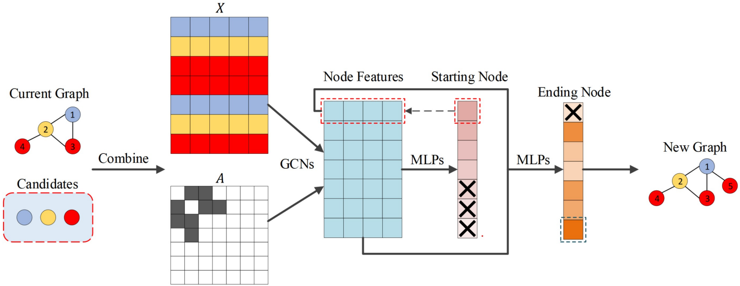


Bayesian network for dependencies

Generating explanations

- Finding an explanatory subgraph via reinforcement learning [25]
- Synthetically creating a new graph as explanation
  - Reinforcement learning [26]
  - One-shot learning [18]

From Figure 2 in [26]:



What does HoT 1 do?

Bemerkung

Motif-Datasets are datasets, where "motifs" (a random, but fixed graph of small size) are attached to a random Barabási-Albert (BA) graph.

- We have 10 different motif-datasets.
- For each iteration, you can choose explainers to visualize
- In each visualization, the ground-truth motif is visualized additionally

How to get started with HoT 1

- Dataset: 10 different Motif Datasets and the explainers from before
- Explanations are visualized with ground truth

Setup:

- Link: `https://github.com/mathematiger/Hands_on_GraphXAI`
- You can install it yourself or use an available VM
- Time: 15min

- Create your own folder, where you copy the folder `GraphXAI/Hands_on_GraphXAI`
- Run Explainers in terminal via: `./run_explainers.sh`
- Find the visualizations in subfolder `content/plots_explainers`

Mutual Information

Question

- *How many questions are needed to identify x, if you know y?*
- *How predictable is x, after observing y?*

Definition

The entropy H measures the surprise of a random variable x:

$$H(x) = E[-\log p(x)] = \sum_x p(x) \log(\frac{1}{p(x)})$$

One can identify:

- $p(x)$ as the probability
- $\log(\frac{1}{p(x)})$ as the number of needed y/n questions to identify this x

Definition

Mutual Information (MI) is measured as:

$$I(x, y) = H(x) - H(x \mid y)$$

MI in XAI

Reminder

$$H(x) = \sum_x p(x) \log\left(\frac{1}{p(x)}\right)$$

$$I(x, y) = H(x) - H(x \mid y)$$

If:

- x is the original output of the GNN
- y is the output of the GNN on an explanatory subgraph

we can ask:

Question

How predictable is x after observing y?

In XAI we are interested in maximizing this predictability.

Usage of MI in [16], [15]

Notation

Y is the GNN-out on the original input data, (G_S, F) is the GNN-out on the masked input data, where G_S is a node-mask and F is a Feature mask.

Aim in GNNExplainer [16]:

$$\max_{G_S, F} MI(Y, (G_S, F)) = H(Y) - H(Y|(G_S, F))$$

Here, for all explainers $H(Y)$ is fixed, hence:

$$\max_{G_S, F} MI(Y, (G_S, F)) = \min_{G_S, F} H(Y|(G_S, F))$$

Two problems arise:

- Which general optimization Framework should be used?
- How to handle the discrete node-masks?

Handling discrete Masks: The reparametrization trick

We add randomness to take the existence of an edge from a probability distribution.

- Choose uniformly random $\varepsilon \in [0, 1]$
- Replace the edge $e_{i,j}$ by the equation (1) below. This approximates the Bernoulli distribution.
- Learn the parameters $\omega_{i,j} \in \mathbb{R}$ instead of $e_{i,j} \in 0, 1$

An edge $e_{i,j}$ gets replaced by:

$$e_{i,j} \leftarrow \sigma\left(\frac{\log \varepsilon - \log 1 - \varepsilon + \omega_{i,j}}{\tau}\right) \quad (1)$$

Here, τ is a parameter to control the approximation.

Optimization Framework in GNNExplainer

We assume convexity and use Jensen's inequality. Additionally, we do not optimize over all possible graphs, but over all possible sugraphs, i.e. $A_S[j,k] \leq A_C[j,k] \forall j,k$ and A_S, A_C being the adjacency matrices of the explanatory and computation subgraph, respectively.

$$\begin{aligned}
\max_{G_S,F} MI(Y,(G_S,F)) &= \min_{G_S,F} H(Y|(G_S,F)) \\
&= \min_{G_S,F} E_{G_S}[H(Y|(G_S,F))] \\
&\overset{\text{Jensen's ineq}}{=} min_{G,F} H(Y|E_G[G_S],F)
\end{aligned}$$

Optimization Framework in PGExplainer

- PGExplainer uses link-prediction techniques to obtain edge masks.
- No usage of features.
- Reparametrization trick: View the adjacency matrix coming from a Bernoulli distribution and optimize:

$$\max_{G_S} MI(Y, G_S) \Leftrightarrow \min_{G_S} H(Y|G_S)$$
$$\approx \min_{\Omega} E_{\varepsilon \tilde{U}(0,1)} H(Y|G_S)$$

- Conditional entropy can be replaced by cross entropy to find "why a certain class label was predicted" rather than "which input improves predictability".
- This can be approximated by Monte-Carlo.

Difference of MI and GNN-output

- MI:
  - Finds input, which very clearly explains one label (not necessary the to-be-explained label)
  - Assumption: Masking the input only purifies the label, instead of changing the prediction
- GNN-output
  - Finds input, which is most relevant for the to-be-explained label, but possible also for other labels

Regularization [25] [18] [15]

- Length of the explanation
  - Number of Nodes
  - Upper sum on probabilities of selected nodes
  - Maximal distance of 2 nodes
- Connected subgraphs
  - Directly from optimization framework
  - Higher likelihood for adjacent edges of selected explanatory output
- Similarity
  - Similarity of node representations
  - Explanation graph being ”realistic” compared to the input data, e.g. molecule data

RG Explainer [25]

RGExplainer uses reinforcement learning to learn build the explanatory subgraph, for node and graph classification tasks. It utilizes:

- Starting Point Selection (for graph classification): Finding the most influential node by Cross-entropy of original prediction and prediction, only based on this node.
- Iterative subgraph search: Reinforcement learning is used to select a neighbor node of the current graph, or the stopping criteria
  - Objective: MLP with loss function Cross-entropy between original prediction and prediction on generated graph
  - Regularization: Size, maximal distance between two nodes, similarity
- Stopping criteria is a threshold for not adding to many nodes

GStarX - making Shapley Values structural dependant [27]

- Uses the GNN-Score as method to optimize
- Modifies Shapley Values: Subgraphs are only allowed to cooperate with their neighbors.
  - Surplus is defined as: $p(j, S) := v(S \cup \{j\}) - v(S) - v(\{j\}$
  - The HN-Values are then defined for any $0 \leq \tau \leq 1$ as:

$$v_\tau^* := \begin{cases} v(S) + \tau \sum_{j \in \bar{S} \setminus S} p(j, S) & \text{if } |S/G| = 1 \\ \sum_{T \in S/G} v_\tau^*(T) & \text{otherwise} \end{cases}$$

- The values are:
  - v is the scoring function (the GNN-out)
  - S is a coalition (i.e. a subgrpah)
  - $\bar{S} = \bigcup_{i \in S} \{N(i)\} \cup S$
  - $j \in \bar{S}$
- Benefit: Using HN Values instead of Shapley Values makes SubgraphX faster by a factor of 2.

XGNN Explainer [26] (global method)

Aim: Generate explanatory graphs using RL.

- Maximizes GNN Output.
- Regularization for graph rules.

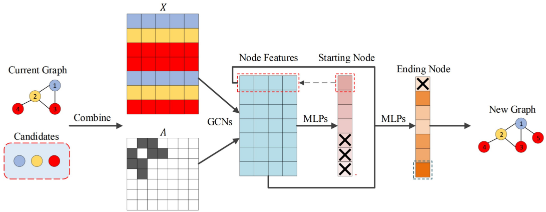
Challenges:

- Requires finite set of graph candidates.
- Graph Rules have to be implemented manually.
- No code provided by authors.

RL Approach:

- **Current:** $C = p(f(G_{t+1}) = c_i)$
- **Rollouts:** $R = \lambda_1 \frac{1}{m} \sum_{i=1}^{m} p(f(\text{Rollout}_i(G_{t+1})) = c_i)$
- **Reg.:** $\lambda_2 \mathcal{L}(t, r)$
- **Loss:** $\mathcal{L} = C + R + \lambda_2 \mathcal{L}(t, r)$
- **Stop:** High loss or after $S_{\max}$ steps.

From Figure 2 in [26]:



Global Explainer: GNNInterpreter [18]

Further interesting Explainers

- FlowX - visualizing the information flow [28]

Different scopes of explanations [29]

- Functionally Grounded: Accuracy, Faithfulness, fidelity, ...
- Human Grounded: Degree of understanding
- Application Grounded: Performance increase

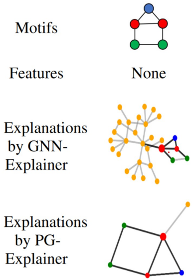
Ground Truth Motifs in GNN Explanations

Ground Truth Motifs assist explainers in identifying the crucial input processed by a GNN.

- **Dataset:** A random (synthetic) graph is augmented with specific motifs. The GNN is trained to differentiate between nodes within and outside the motif.
- **Assumption:** Only nodes within the motif impact predictions. External nodes are noise. It's presumed the GNN identifies this distinction.
- **Explainer's Role:** The explainer should focus solely on the ground-truth motif, sidelining the noise. A subgraph is termed ground-truth if its nodes are exclusively pivotal for predictions while external graph sections act as noise.
- **Efficacy Metric:** The alignment of the explainer's output with the ground truth motive measures its effectiveness. Consistency suggests that the explainer is accurately identifying relevant data.

The use of ground-truth motifs is used to assessing whether the explainer has skilfully recognised the relevant input. If the explanation offered by GNN matches the ground-truth subgraph, it indicates that the explainer has skilfully found salient data.

Accuracy to ground-truth [1]



$$acc = \frac{TP}{TP + FP + FN}$$

Here:

- TN is not counted, as the explanatory motif only gives positives as feedback
- positives and negatives can be counted on nodes or edges

But this method suffers some pitfalls, explained later!

Fidelity: How well can the explanation approximate the model? [30]

$$\mathrm{fid}_+(S) = \frac{1}{n}\sum_{i=1}^{n} |\mathbf{1}_{\hat{y}_{i,G}=\hat{y}_{i,G\setminus S}}|$$

$$\mathrm{fid}_-(S) = \frac{1}{n}\sum_{i=1}^{n} |\mathbf{1}_{\hat{y}_{i,G}=\hat{y}_{i,S}}|$$

Notation

- *$\hat{y}_{ind}$ is the prediction of the GNN evaluated on the input ind*
- *G is the computation graph*
- *S is the explanatory subgraph*
- *n is the number of instances*

Fidelity can be interpreted as follows:

$\mathrm{fid}_+(S)$

- Remove explanatory subgraph
- Measure, how much relevant input is found

$\mathrm{fid}_-(S)$

- Only look at the output of the explanatory subgraph
- Measure, if sufficient input is found for this label

This evaluation can be used for global and local explainers.

Pitfalls [31]

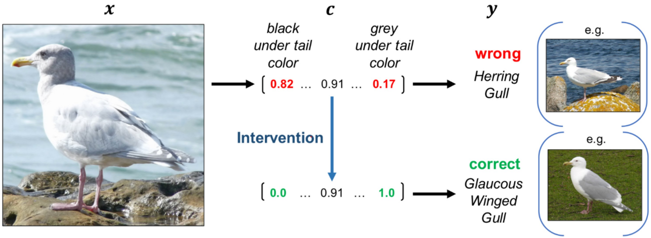
Conclusion

- XAI on graphs is helpful for many stackholders
- Finding the right explanation for the right audience stays challenging
- There exists an available framework for creating and evaluating local and global explanations
- Still a lot of research possible

Outlook 1: Intrinsic Explainers [32] [33]

- Idea: Classify many intermediate concepts and base the prediction on these.
- Advantage: Intermediate predictions are changeable by humans, hence it is easier to find the errors of the model

Figure from Figure 3 in [32]



Explainable Embeddings: ExCut [34]

Outlook 2: Further Research

- Identifying important concepts in graphs
- Global Explanations
  - Generating synthetic graphs
  - Extracting meaningful explanations
- Explainers for tasks beyond node and graph classification
- Explaining heterogeneous graph data (graphs with different node types)
- Using methods from knowledge graphs for XAi of GNNs
- Finding the optimal way for introducing XAI to end-users

Newer Research on Graph ExplainabilityQuestions

Some amazing research is going on, see
https://github.com/flyingdoog/awesome-graph-explainability-papers
I am happy for any references to insert here!

HoT2

- You are given a motif-dataset like for HoT1 and a trained GNN ontop with 100% accuracy.
- Task: Find the optimal graph with features in $\{0, 1, 2, 3\}$, which has the highest summed up prediction for all nodes
- Maximal 6 nodes
- input:
  - List of edges [(0,1), (1,2)] between node types
  - List of features [0,1,2], i-th space for the i-th node
- Run: `./find_motif.sh`

- Time: until 12.00
- No cheating! If any questions arise, ask me.
- Fun-Fact: Your number will probably be a lot higher, than the motif used to generate the dataset.

HoT3

- Task: Explore Class Expressions as explanations: Find the Class Expression with the highest fidelity.
- Input: A tree as list-form: [class, [subtree 1], [subtree 2]]. The subtrees will be added with an intersection.
- Allowed Classes are '0','1', '2','3'. Please don't forget to write the numbers as strings.
- Run: `./run_ce_expl.sh`

HoT4 [31]

- Task: Explore a dataset and explainers
- The dataset has Class 1, if 1 node in a 1-hop neighborhood has features $[1, 0]$ (all other nodes have features $[0, 1]$)
  - Sadly, not all explainers work well on this dataset (due to my bad implementation)
- run by `./find_dataset.sh`
- The created explanations are in the folder content_fds/plots_explainers

Setup:

- Link: `https://github.com/mathematiger/Hands_on_GraphXAI`

Sources I

[1] C. Agarwal, O. Queen, H. Lakkaraju, and M. Zitnik, "Evaluating explainability for graph neural networks," *Scientific Data*, vol. 10, no. 144, 2023. [Online]. Available: https://www.nature.com/articles/s41597-023-01974-x

[2] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019. [Online]. Available: https://doi.org/10.1016/j.artint.2018.07.007

[3] B. Kim, O. Koyejo, and R. Khanna, "Examples are not enough, learn to criticize! criticism for interpretability," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 2280–2288. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html

[4] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[5] Defense Advanced Research Projects Agency (DARPA). (Year Unknown) Explainable artificial intelligence program. Accessed on October 11, 2023. [Online]. Available: https://www.darpa.mil/program/explainable-artificial-intelligence

Sources II

[6] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[7] G. Vilone and L. Longo, "Explainable artificial intelligence: a systematic review," *CoRR*, vol. abs/2006.00093, 2020. [Online]. Available: https://arxiv.org/abs/2006.00093

[8] D. Kahneman, *Thinking, fast and slow*. macmillan, 2011.

[9] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance," *Frontiers Comput. Sci.*, vol. 5, 2023. [Online]. Available: https://doi.org/10.3389/fcomp.2023.1096257

[10] ——, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[11] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum, "What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research," *Artificial Intelligence*, vol. 296, p. 103473, 2021.

Sources III

[12] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. M. Wallach, “Manipulating and measuring model interpretability,” in *CHI ’21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. M. Drucker, Eds. ACM, 2021, pp. 237:1–237:52. [Online]. Available: https://doi.org/10.1145/3411764.3445315

[13] G. Schwalbe and B. Finzel, “A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts,” *Data Mining and Knowledge Discovery*, 01 2023.

[14] Y. Shin, S. Kim, E. Yoon, and W. Shin, “Prototype-based explanations for graph neural networks (student abstract),” in *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 2022, pp. 13 047–13 048. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/21660

[15] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, “Parameterized explainer for graph neural network,” in *NeurIPS*, 2020.

Sources IV

[16] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," in *NeurIPS*, 2019, pp. 9240–9251.

[17] T. Bui, V. Le, W. Li, and S. Cha, "INGREX: an interactive explanation framework for graph neural networks," *CoRR*, vol. abs/2211.01548, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2211.01548

[18] X. Wang and H. Shen, "Gnninterpreter: A probabilistic generative model-level explanation for graph neural networks," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/pdf?id=rqq6Dh8t4d

[19] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5782–5799, 2023. [Online]. Available: https://doi.org/10.1109/TPAMI.2022.3204236

[20] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji, "On explainability of graph neural networks via subgraph explorations," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 12 241–12 252.

Sources V

[21] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016. [Online]. Available: http://arxiv.org/abs/1610.02391

[22] F. Baldassarre and H. Azizpour, "Explainability techniques for graph convolutional networks," *CoRR*, vol. abs/1905.13686, 2019. [Online]. Available: http://arxiv.org/abs/1905.13686

[23] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 10 772–10 781.

[24] M. N. Vu and M. T. Thai, "Pgm-explainer: Probabilistic graphical model explanations for graph neural networks," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/8fb134f258b1f7865a6ab2d935a897c9-Abstract.html

Sources VI

[25] C. Shan, Y. Shen, Y. Zhang, X. Li, and D. Li, "Reinforcement learning enhanced explainer for graph neural networks," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 22 523–22 533. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/be26abe76fb5c8a4921cf9d3e865b454-Abstract.html

[26] H. Yuan, J. Tang, X. Hu, and S. Ji, "XGNN: towards model-level explanations of graph neural networks," in *KDD*. ACM, 2020, pp. 430–438.

[27] S. Zhang, Y. Liu, N. Shah, and Y. Sun, "Gstarx: Explaining graph neural networks with structure-aware cooperative games," in *NeurIPS*, 2022.

[28] S. Gui, H. Yuan, J. Wang, Q. Lao, K. Li, and S. Ji, "Flowx: Towards explainable graph neural networks via message flows," *CoRR*, vol. abs/2206.12987, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2206.12987

[29] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv: Machine Learning*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:11319376

Sources VII

[30] K. Amara, Z. Ying, Z. Zhang, Z. Han, Y. Zhao, Y. Shan, U. Brandes, S. Schemm, and C. Zhang, “Graphframex: Towards systematic evaluation of explainability methods for graph neural networks,” in *LoG*, ser. Proceedings of Machine Learning Research, vol. 198. PMLR, 2022, p. 44.

[31] L. Faber, A. K. Moghaddam, and R. Wattenhofer, “When comparing to ground truth is wrong: On evaluating GNN explanation methods,” in *KDD*. ACM, 2021, pp. 332–341.

[32] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, “Concept bottleneck models,” in *ICML*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 5338–5348.

[33] S. Azzolin, A. Longa, P. Barbiero, P. Liò, and A. Passerini, “Global explainability of gnns via logic combination of learned concepts,” in *ICLR*. OpenReview.net, 2023.

[34] M. H. Gad-Elrab, D. Stepanova, T. Tran, H. Adel, and G. Weikum, “Excut: Explainable embedding-based clustering over knowledge graphs,” in *ISWC*, ser. Lecture Notes in Computer Science, vol. 12506. Springer, 2020, pp. 218–237.