Tutorial: Interpretability Methods for Graph Neural Networks

Arijit Khan and Ehsan B. Mobaraki

AALBORG UNIVERSITY

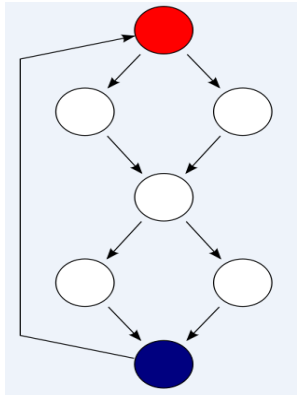
IEEE | Cosponsers ACM/ASA/CCF

DSAA 2023

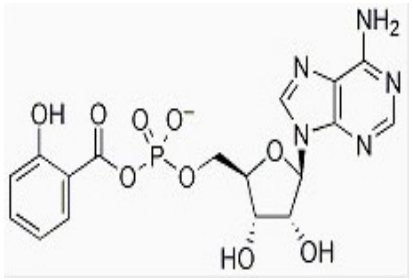
Graph data is everywhere

Graph database with many smaller graphs

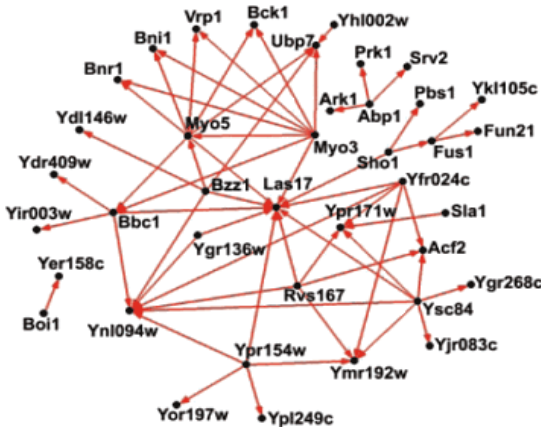
One large graph



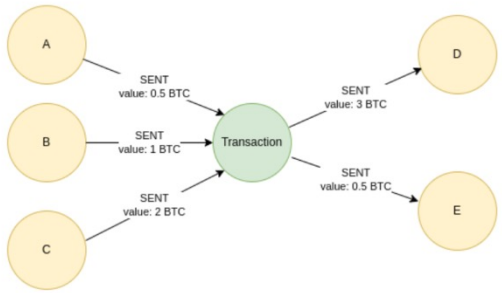
Program flow/ call graph



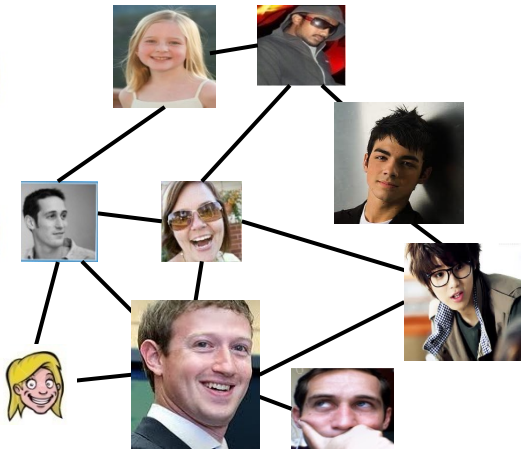
Chemical compound structure



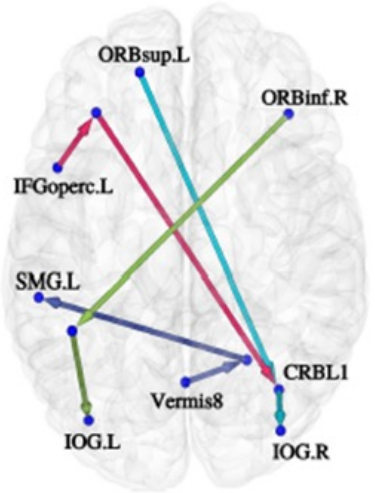
Biological network

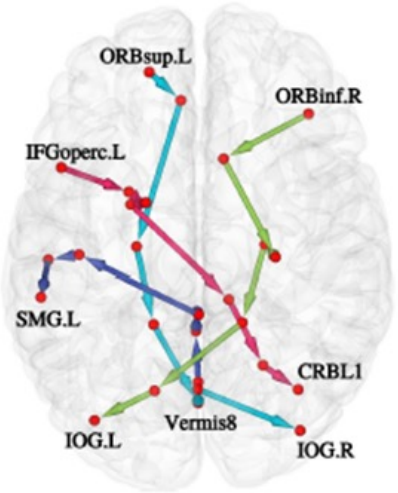


Financial transaction

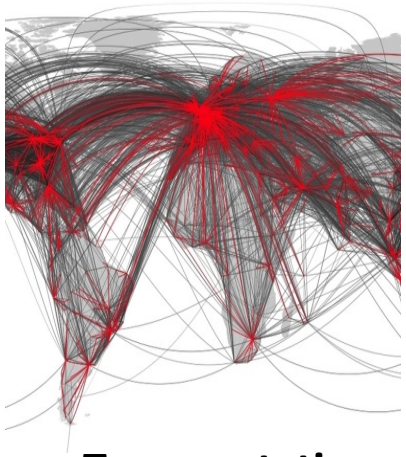


Social network

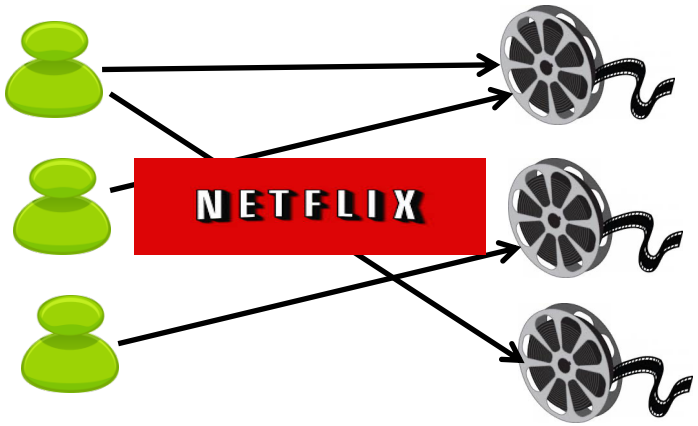




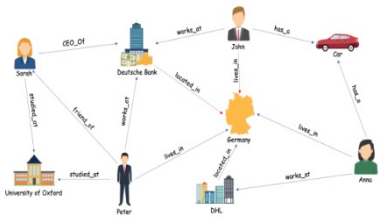
Human brain network



Transportation



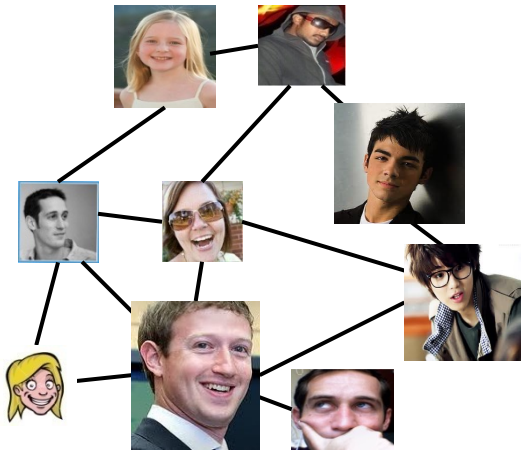
Recommendation



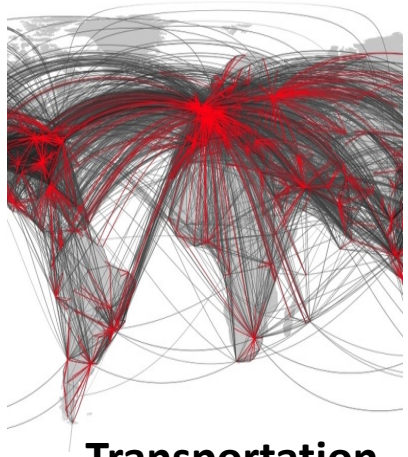
Knowledge graph

Graph data is everywhere

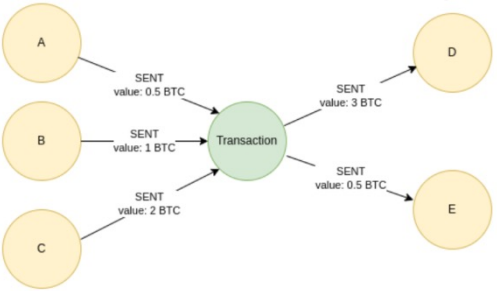
One large graph



Social network

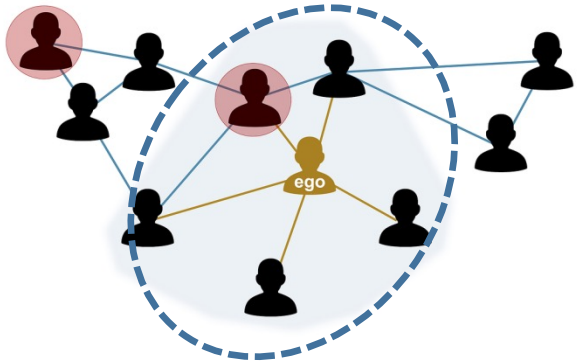


Transportation

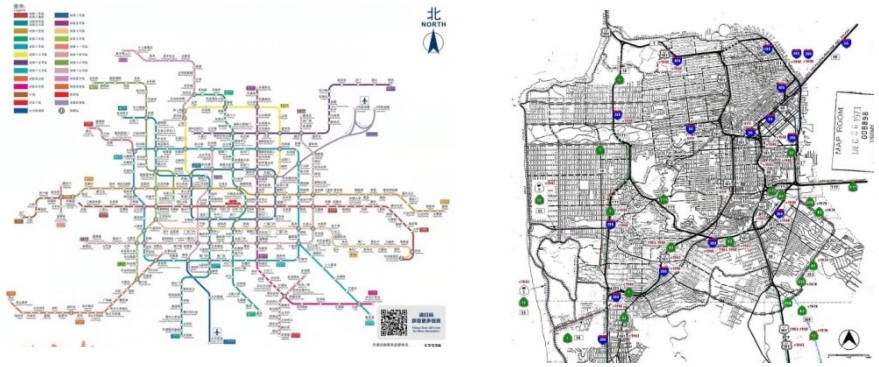


Financial transaction

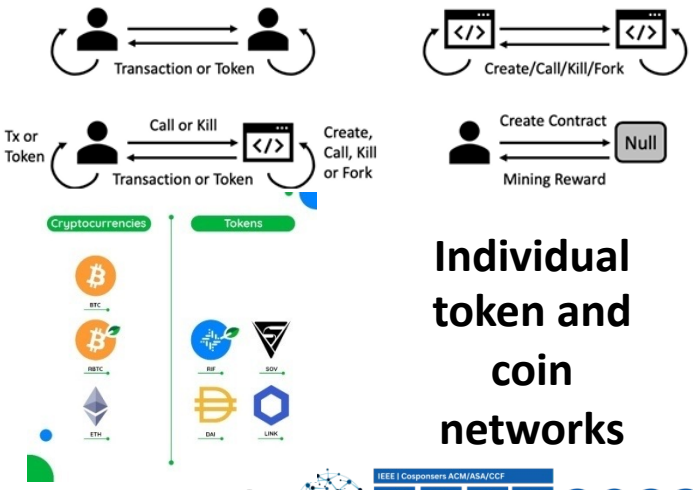
Graph database with many smaller graphs



Ego network of individual users



Road networks of individual cities

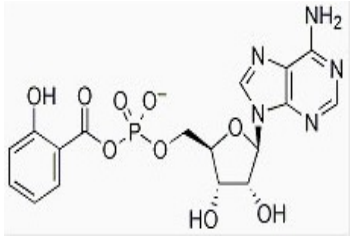


Individual token and coin networks

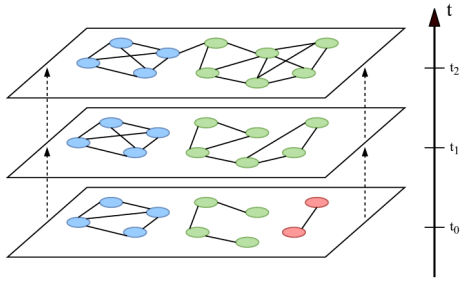
Graph data in many forms

AALBORG UNIVERSITY

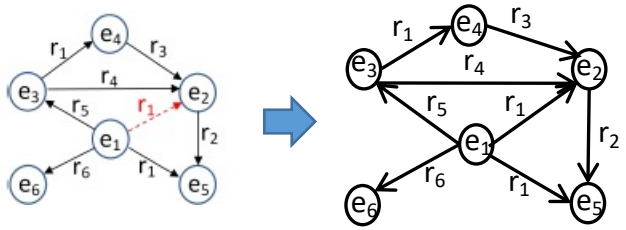
Static graph



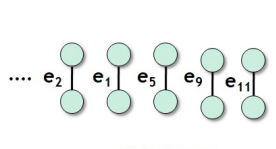
Dynamic / temporal / time-evolving graph



Historic / temporal snapshot graph

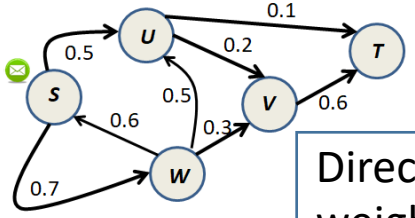


Dynamic / incremental update



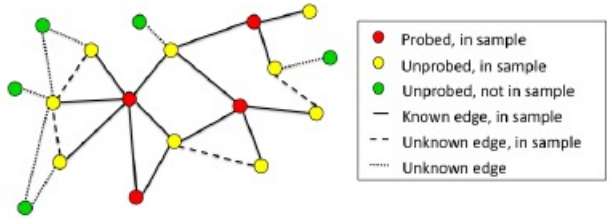
Graph stream

Uncertain graph

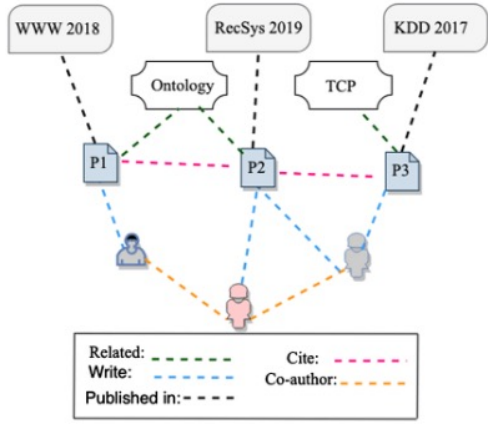


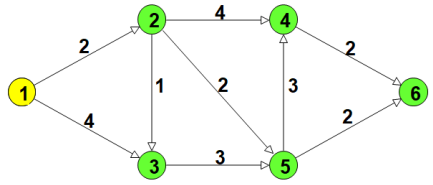
Directed/ weighted/ attributed graph

Incomplete / partially observed graph



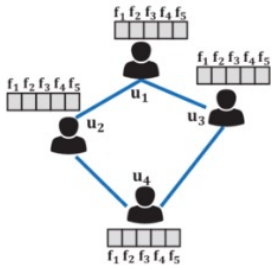
Heterogeneous graph

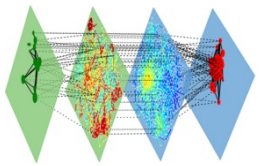




Higher-order graph

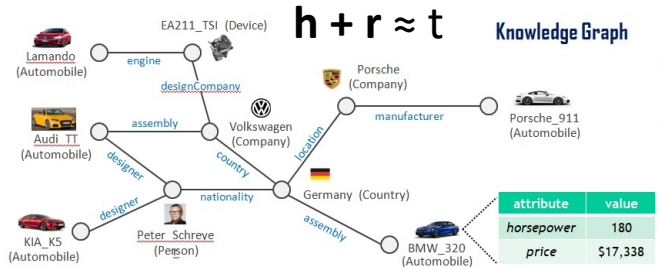
Higher-order interaction

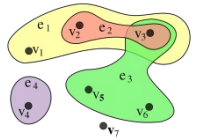




Multilayer network

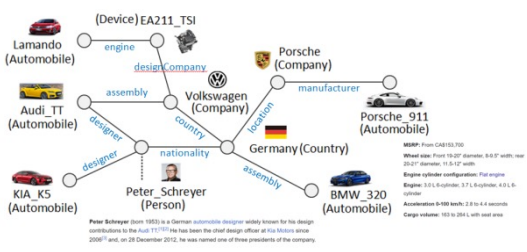
Knowledge graph

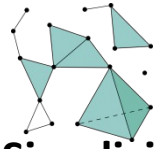




Hypergraph

Multi-modal graph





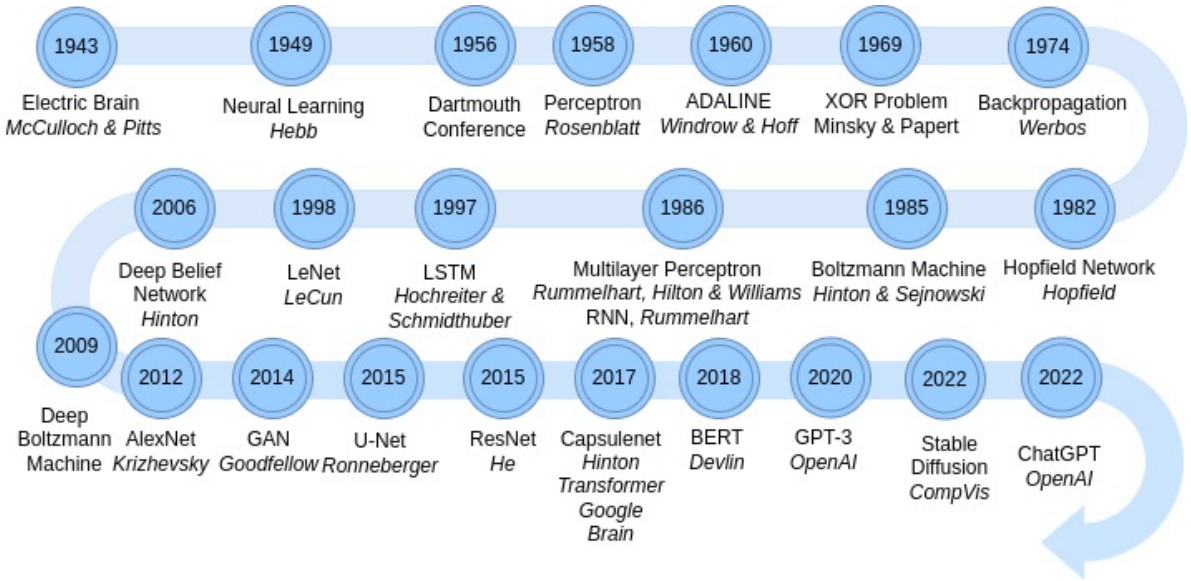
Simplicial complex

Graph data and neural networks

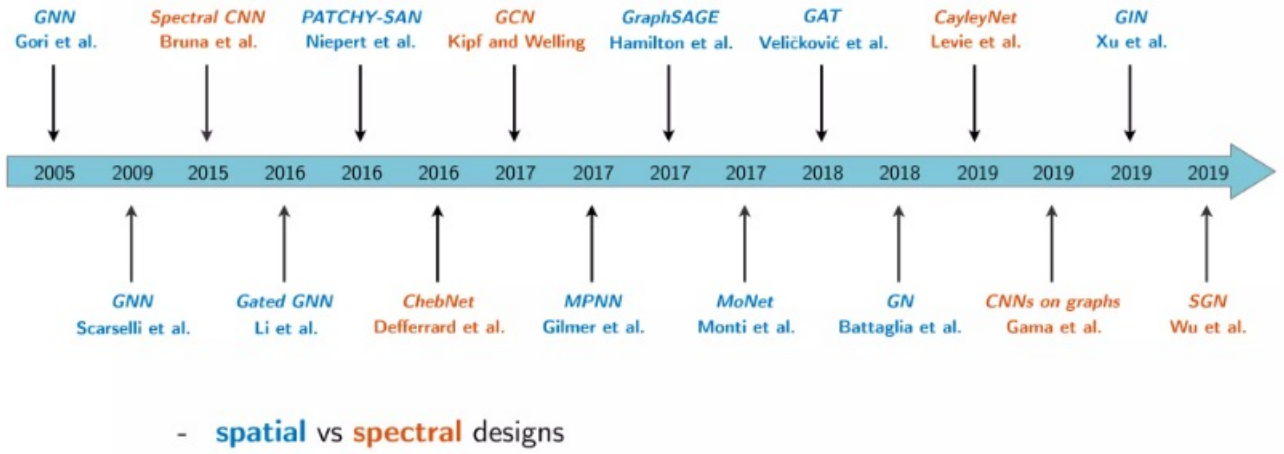
• Based on the increasing usage, popularity and maturity of graph, Gartner estimates that the market for graph technologies, including graph database management systems (DBMSs), will grow to $3.2 billion by 2025 with a compound annual growth rate (CAGR) of 28.1%.

• Vendors in the graph DBMS market are expanding their stacks into platforms for enterprise knowledge graphs or graph artificial intelligence (AI), with associated product ecosystems.

- **Market Guide for Graph Database Management Solutions, Gartner, August 2022**

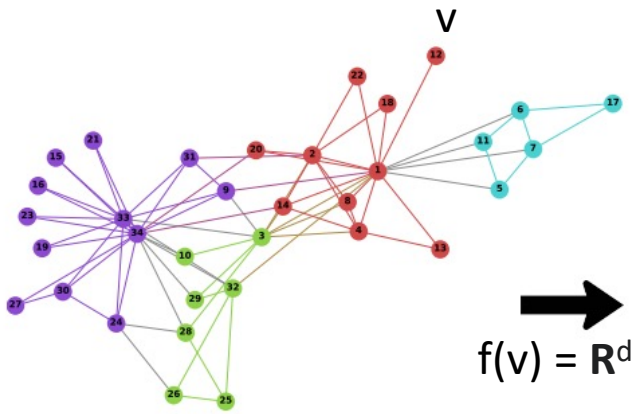


https://pub.towardsai.net/a-brief-history-of-neural-nets-472107bc2c9c

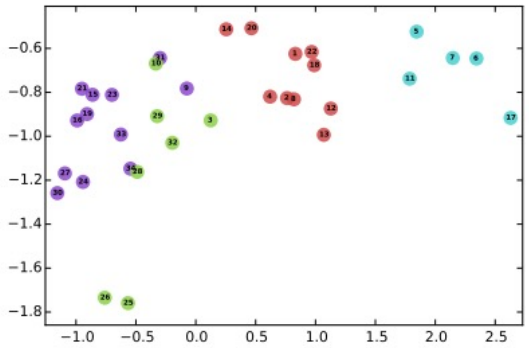


Graph Signal Processing for Machine Learning: A Review and New Perspectives (ICASSP Tutorial, 2021)

Graph neural network (GNN): Key idea



$f(v) = \mathbf{R}^d$



Representation learning or embedding

Downstream tasks (e.g., graph and node classification, recommendation, link prediction, question answering)

Learning could be end-to-end

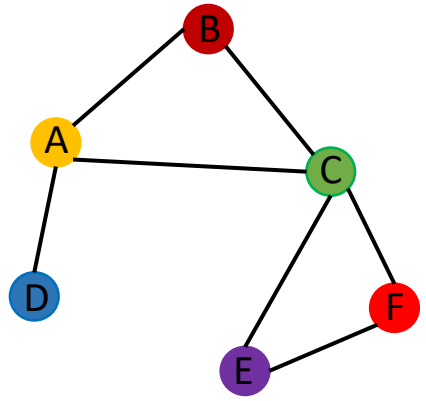
- End-to-end learning → ~~Feature Engineering~~
- Task-independent / task-dependent learning.
- Can capture graph structure and node, edge features.

Graph Convolutional Neural Network (GCN)

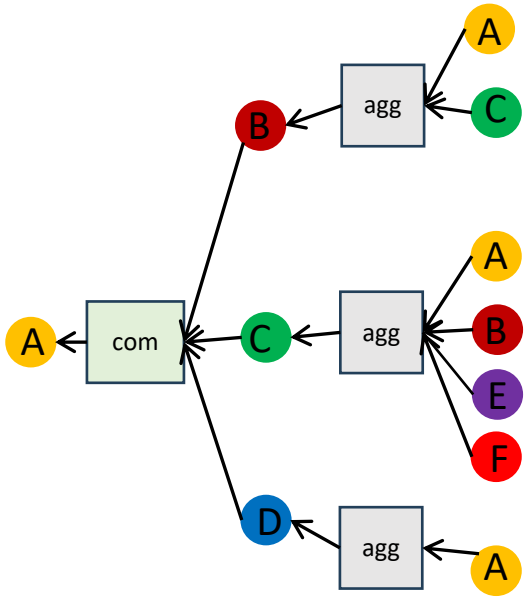
- Massage passing to use aggregation and combine functions repeated several times.

$$H^{(t+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} . \tilde{A} . \tilde{D}^{-\frac{1}{2}} . H^{(t)} . W^{(t)})$$

$\tilde{A} = A + I_N$ $\quad H^{(0)} = X$ Input Node Features



Input graph



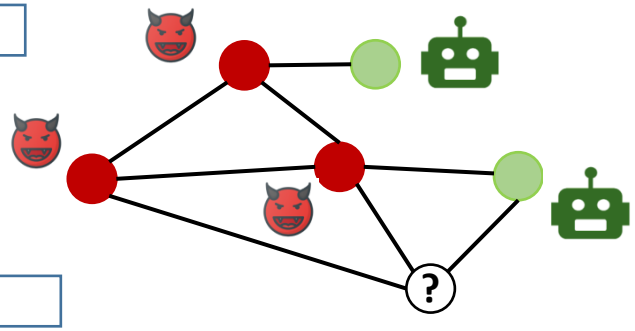
Technique in GCN

Representation Learning on Networks (WWW Tutorial, 2018)

Graph neural network (GNN): Downstream tasks

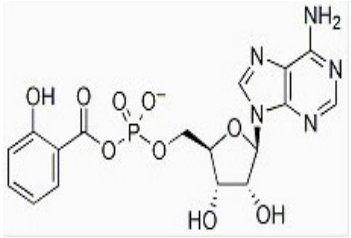
Node classification

Node-level task



Graph classification

Graph-level task

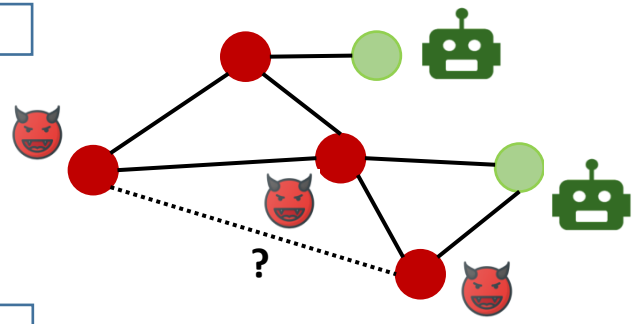


? Mutagen

? Non-mutagen

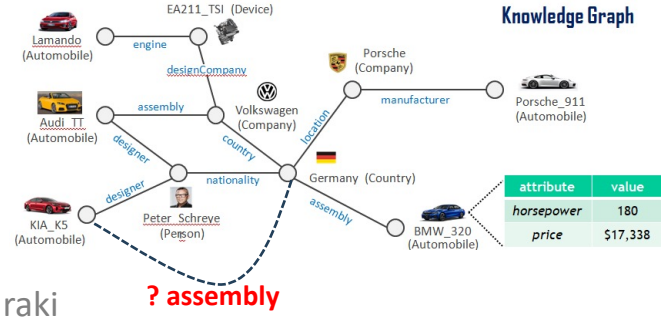
Link prediction

Edge-level task



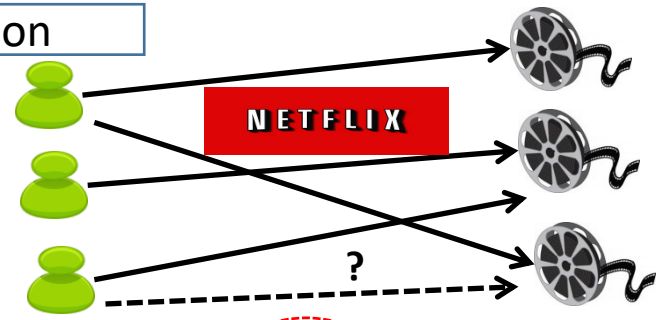
Triple classification

Edge-level task

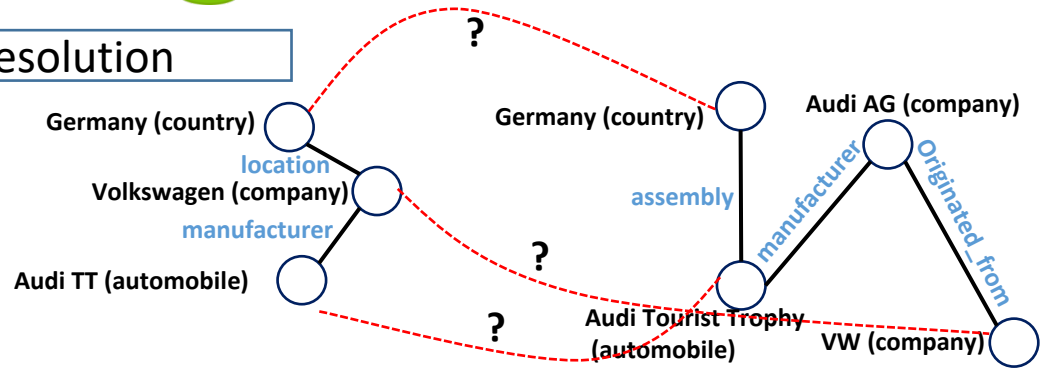


A Comprehensive Survey on Graph Neural Networks. IEEE Trans. Neural Networks Learn. Syst. 2021.

Recommendation



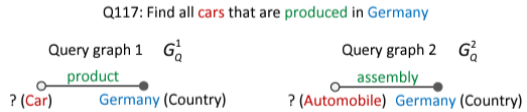
Entity resolution

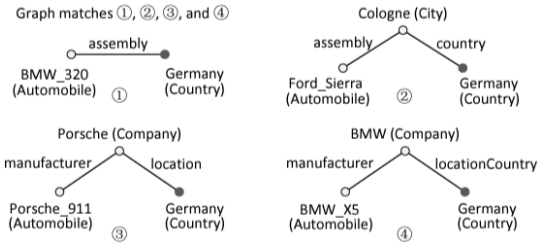


Knowledge graph 1

Knowledge graph 2

Question answering





Drug design

- predicting missing links between drug and disease.

Graph neural network (GNN): Interpretability

AALBORG UNIVERSITY

- Explain the results of high-quality GNNs.
- **[Instance-level]** Understand which aspects of the input data drive the decisions of the GNN – discover critical nodes, edges, subgraphs, and their features that are responsible for GNN outcomes.
- **[Model-level]** Insight on how GNNs work – discover what input subgraph patterns lead to a certain prediction.

Explainability in graph neural networks: A taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell., 2022.

Importance

- Desirable to understand and explain the workings and results of black-box GNNs – bridge domain knowledge with GNN predictions, human-AI collaboration.
- Safety and well-being (e.g., autonomous car, AI in healthcare) – trust in deep learning models.
- Understand bias in machine learning (ML) algorithms – ML algorithms can amplify bias, model debugging.
- Robustness against adversarial examples – improve quality of GNN outputs.
- Legal requirements, e.g., GDPR – algorithms to explain their outputs.

Stakeholders

End users, domain experts, decision makers, policy makers, regulatory agencies, researchers, data scientists, and engineers

Tutorial outline

1 Introduction

1.1 Graph neural networks (GNNs) and applications
1.2 Interpretability of GNNs
- Definitions, importance, and challenges

2 Taxonomy of interpretability methods for GNNs

2.1 Post-hoc vs. intrinsic / self-explainable
2.2 Global/ class-specific vs. local/ instance-specific
2.3 Model-specific vs. model-agnostic
2.4 Forward vs. backward
2.5 Node-level vs. edge-level vs. subgraph-level
2.6 Perturbation vs. gradient vs. decomposition vs. surrogate models vs. counterfactuals

3 Recent interpretability methods for GNNs

GNNExplainer, PGExplainer , GraphMask, SubgraphX , PGMExplainer, CF2, SA, GuidedBP, CAM, Grad-CAM, LRP, ExcitationBP, and XGNN

4 Benchmark & ground truth for GNN interpretability methods

4.1 Interpretability evaluation metrics
4.2 Ground truth datasets, software
4.3 Benchmarking results

5 Future directions

More on interpretability

AALBORG UNIVERSITY

• There is no standard definition – no unique notion of interpretability in the literature.

• Different motivations and requirements for interpretability:

– trust, causality, transferability, informativeness, fair and ethical decision making, model debugging, recourse, mental model comparison, context-dependent, low-level mechanistic understanding of models, high-level human understanding, what makes users confident about the model.

The mythos of model interpretability. Commun. ACM, vol. 61, no. 10, 2018.

"Ability to explain or to present a model in understandable terms to humans"
- Doshi-Velez and Kim 2017

Global concept-based interpretability for graph neural networks via neuron an alysis. AAAI 2023.

• **Outputs of Interpretability**

– Heat map visualization, explanation by example, explanation by text, local explanation, explanation based on higher-level patterns/ rules/ global concepts/ counterfactuals.

Explainability in graph neural networks: A taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell., 2022.

• **Interpretability vs. Explainability**

• Often used interchangeably.

https://www.xcally.com/news/interpretability-vs-explainability-understanding-the-importance-in-artificial-intelligence

• Interpretability concerns the understanding (of inner workings) of the model by AI experts and researchers, while explainability focuses on explaining the decisions made to end users.

Challenges with GNN interpretability

- Many definitions, motivations, and requirements for interpretability.

- Comparing explanations is hard!

- Several quantitative and qualitative evaluation metrics or methods.

– **Quantitative:** faithfulness (fidelity+, fidelity-), sparsity, contrastivity, accuracy, stability.

– **Qualitative:** application-grounded evaluation, human-grounded evaluation, functionally-grounded evaluation.

To be discussed later

HCI, visualization domains; more difficult for GNNs

- Difficult to obtain ground-truth.

– Synthetically created ground-truth: BA-shapes, BA-2Motifs, BA-Community, Tree-Cycle, Tree-Grids, etc.

To be discussed later

- **Other issues:** Evaluation via occlusion creates data outside training distribution, bias terms, redundant evidence, trivial correct explanations, weak GNN model, misaligned GNN architecture, problems due to graph data vs. grid data.

- Capture interplay of graph structure and features in GNN's decision making.

When comparing to ground truth is wrong: on evaluating GNN explanation methods. KDD 2021.

Explainability in graph neural networks: A taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell., 2022.

GNN interpretability: Survey and benchmarking

AALBORG UNIVERSITY

• H. Yuan, H. Yu, S. Gui, and S. Ji. *Explainability in graph neural networks: A taxonomic survey*. IEEE Trans. Pattern Anal. Mach. Intell., 2022.

• P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann. *Explainability methods for graph convolutional neural networks*. in CVPR, 2019.

• C. Agarwal, O. Queen, H. Lakkaraju, and M. Zitnik. *Evaluating explainability for graph neural networks*. Sci Data, vol. 10, no. 1, 2023.

• C. Agarwal, M. Zitnik, and H. Lakkaraju. *Probing GNNexplainers: a rigorous theoretical and empirical analysis of GNN explanation methods*. In AISTATS, 2022.

• K. T. T. Shun, E. E. Limanta, and A. Khan. *An evaluation of backpropagation interpretability for graph classification with deep learning*. In IEEE BigData, 2020.

• B. S´anchez-Lengeling, J. Wei, B. Lee, E. Reif, P. Wang, W. Qian, K. McCloskey, L. Colwell, and A. Wiltschko. *Evaluating attribution for graph neural networks*. In NeurIPS, 2020.

• F. Baldassarre and H. Azizpour. *Explainability techniques for graph convolutional networks*. In ICML Workshop on Learning and Reasoning with Graph-Structured Representations, 2019.

• J. Kakkad, J. Jannu, K. Sharma, C. C. Aggarwal, and S. Medya. *A survey on explainability of graph neural networks*. CoRR, vol. abs/2306.01958, 2023.

• A. Longa, S. Azzolin, G. Santin, G. Cencetti, P. Liò, B. Lepri, and A. Passerini. *Explaining the explainers in graph neural networks: a comparative study*. CoRR, vol. abs/2210.15304 , 2023.

• P. Li, Y. Yang, M. Pagnucco, and Y. Song. *Explainability in graph neural networks: an experimental survey*. CoRR, vol.abs/2203.09258, 2022.

• M. Khosla and L. Galárraga. *Explainable graph machine learning: techniques to explain black-box models on graphs* (**Tutorial**). ECML 2023.

• 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. LoG 2022.

• …

Motivation of our tutorial

AALBORG UNIVERSITY

- Categorization of GNN interpretability methods in many verticals; advantages and disadvantages.

- Description of 13 representative, recent GNN interpretability methods; advantages and disadvantages.

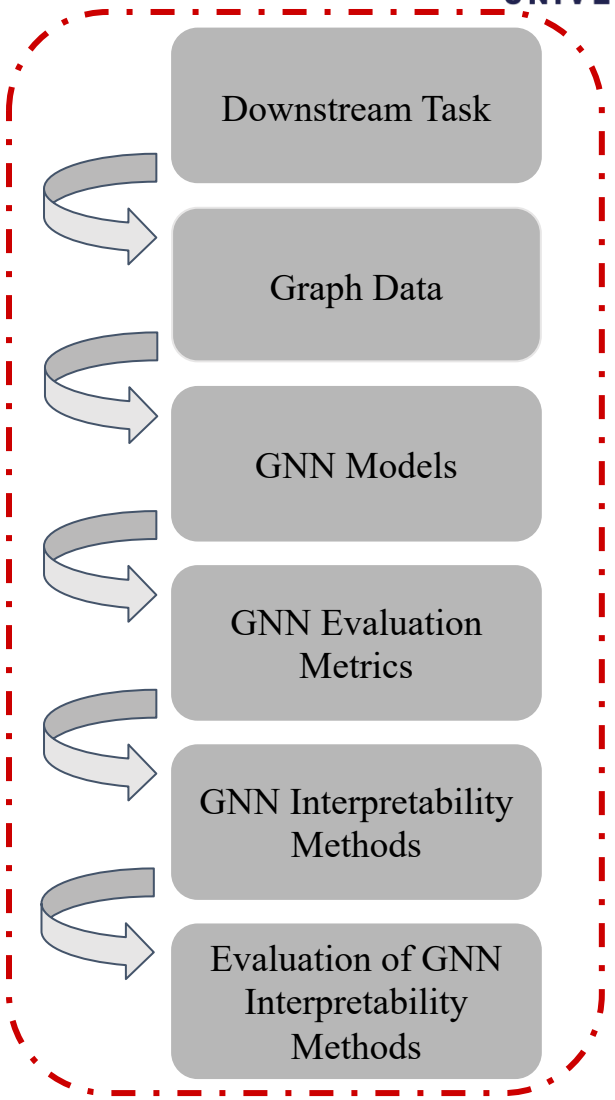
- Challenges with GNN interpretability.

- Evaluation metrics, ground truths, software for GNN interpretability.

- Preliminary benchmarking results and case study on GNN interpretability; beyond graphs classification and nodes classification.

- Interplay of GNN model, graph data, interpretability methods, evaluation metrics, and downstream tasks.

- Future directions.



This tutorial is not about …

- What is interpretability? Interpretability vs. Explainability.

  The mythos of model interpretability. Commun. ACM, vol. 61, no. 10, 2018.

- Qualitative evaluation of interpretability methods.

  Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. CHI 2020.

- Graph counterfactual explanations.

  A survey on graph counterfactual explanations: Definitions, methods, evaluation. CoRR abs/2210.12089 (2022).

- Self-explainable GNNs
  - *Graph attention networks*. In ICLR 2018. – limited by specific GNN architecture.
  - *Towards self-explainable graph neural network*. CIKM 2021.
  - *ProtGNN: Towards self-explaining graph neural networks*. AAAI 2022.
  - *Towards prototype-based self-explainable graph neural network*. In ArXiv 2022.

- Explainability of knowledge graph (KG) embedding, KG link prediction, graph embedding
  - *Explainable graph machine learning: Techniques to explain black-box models on graphs* (**Tutorial**). ECML 2023.
  - *Explaining link prediction systems based on knowledge graph embeddings*. SIGMOD 2022.
  - *On the interpretability and evaluation of graph representation learning*. CoRR abs/1910.03081 (2019).

- Explainability of (only) graph classification
  - *Explainable classification of brain networks via contrast subgraphs. KDD 2020.*
  - *Counterfactual graphs for explainable classification of brain networks*. KDD 2021.

- We shall NOT cover *all* GNN explainability methods from the literature.

13 representative, recent GNN interpretability methods from diverse categories

Tutorial outline

1 Introduction

1.1 Graph neural networks (GNNs) and applications
1.2 Interpretability of GNNs
- Definitions, importance, and challenges

2 Taxonomy of interpretability methods for GNNs

2.1 Post-hoc vs. intrinsic / self-explainable
2.2 Global/ class-specific vs. local/ instance-specific
2.3 Model-specific vs. model-agnostic
2.4 Forward vs. backward
2.5 Node-level vs. edge-level vs. subgraph-level
2.6 Perturbation vs. gradient vs. decomposition vs. surrogate models vs. counterfactuals

3 Recent interpretability methods for GNNs

GNNExplainer, PGExplainer , GraphMask, SubgraphX , PGMExplainer, CF2, SA, GuidedBP, CAM, Grad-CAM, LRP, ExcitationBP, and XGNN

4 Benchmark & ground truth for GNN interpretability methods

4.1 Interpretability evaluation metrics
4.2 Ground truth datasets, software
4.3 Benchmarking results

5 Future directions

Post-hoc vs. intrinsic / self-explainable interpretability methods for GNNs

AALBORG UNIVERSITY

We'll focus on them

• **Post-hoc.** Creating a second model to provide explanations for an existing GNN model.

- e.g., perturbation-based approaches (GNNExplainer, NeurIPS2019).

- could be limited in interpretability performance (e.g., reporting spuriously-correlated features with the task), while keeping the underlying GNN accuracy intact.

• **Intrinsic / self-explainable.** Constructing self-explanatory models which incorporate interpretability directly to their structures.

- e.g., use structural constraints to derive an informative subgraph which is used for both prediction and explanation.

- e.g., graph attention networks (ICLR 2018), SEGNN (CIKM 2021), ProtGNN (AAAI 2022).

- trade-off between good interpretability vs. prediction accuracy.

Global/ class-specific vs. local/ instance-specific interpretability methods for GNNs

AALBORG UNIVERSITY

• **Global / class-specific.** Users can understand how the model works globally by inspecting the structures and parameters of a complex model, thereby explaining the “essence” of a class.

- e.g., explore high-level explanations of GNNs by generating graph patterns to maximize a specific prediction.

- e.g., XGNN, KDD 2020.

• **Local / instance-specific.** Locally examines an individual prediction of a model, trying to figure out why the model makes the decision that it makes for that test instance.

- e.g., GNNExplainer, NeurIPS 2019.

Model-specific vs. model-agnostic interpretability methods for GNNs

AALBORG UNIVERSITY

• **Model-specific / white-box.** Requires access to internal model parameters or embeddings to provide explanations.

- e.g., gradient-based methods calculate the gradient of an output w.r.t. the input using backpropagation to derive the contribution of features (*Explainability methods for graph convolutional neural networks.* in CVPR, 2019).
- Gradient-based methods are more efficient, since they usually need one forward and another backward pass.
- **Issues:** gradient saturation, ..

• **Model-agnostic / black-box.** Does not require internals of the GNNs to generate explanations.

- e.g., perturbation-based methods (GNNExplainer, NeurIPS 2019) determine the contribution of a feature by measuring how prediction score changes when the feature is altered.
- can be computationally inefficient as each perturbation requires a separate forward propagation through the network.

Forward vs. backward interpretability methods for GNNs

AALBORG UNIVERSITY

• **Forward interpretability methods**. GNN model-agnostic, learn evidence about graphs or nodes passed through the GNN.

- e.g., perturbation-based, that is, masking some node features and/ or edge features and analyzing the changes when the modified graphs are passed through the GNN model.
- e.g., employ a simple, interpretable surrogate model to approximate the predictions of a complex GNN.
- e.g., counterfactuals-based, i.e., finding a subgraph whose information is necessary which if removed will result in different predictions.

• **Backward interpretability methods.** GNN model-specific.

- e.g., gradient-based – backpropagating importance signal from the output neuron of the model to the individual nodes of the input graph.

- e.g., decomposition-based – distributing the prediction score in a backpropagation manner until the input layer.

Node-level vs. edge-level vs. subgraph-level interpretability methods for GNNs

AALBORG UNIVERSITY

- **Output of interpretability methods**. Node / node feature, edge, subgraph.

- Node / node feature. E.g., ZORRO, IEEE Transactions on Knowledge & Data Engineering. 35(8), 2023.

- Edge. E.g., PGExplainer, NeurIPS 2020.

- Subgraph. E.g., SubgraphX, ICML 2021.

Counterfactual and Factual Reasoning

AALBORG UNIVERSITY

- **Factual** reasoning**.** Finding a subgraph whose information is sufficient to lead to the same prediction for the input sample. E.g., GNNExplainer, Advances in neural information processing systems 2019.

- **Counterfactual** reasoning**.** Finding a subgraph whose information is necessary hence its removal will result in different predictions (i.e., necessary subgraph for the targeted class). E.g., GCFExplainer, WSDM 2023.

- **Factual and Counterfactual** reasoning**.** Finding a subgraph that follows both the factual and counterfactual reasoning; finding a subgraph that outputs the same prediction and its absence will cause changes on the output of the model. E.g., CF^2, WWW 2022.

Different interpretability methods for GNNs (instance-based)

AALBORG UNIVERSITY

• **Perturbation-based.** Masking some node features and/ or edge features and analyzing the changes when the modified graphs are passed through the GNN model. E.g., GNNExplainer, NeurIPS 2019.

• **Surrogate model.** Employs a simple, interpretable surrogate model to approximate the predictions of a complex GNN. E.g., Pgm-explainer, NeurIPS 2020.

• **Gradient-based.** Backpropagating importance signal from the output neuron of the model to the individual nodes of the input graph. E.g., Grad-CAM (*Explainability methods for graph convolutional neural networks*. in CVPR, 2019).

• **Decomposition-based.** Distributing the prediction score in a backpropagation manner until the input layer. E.g., LRP, TextGraphs 2019.

Different interpretability methods for GNNs (model-based)

AALBORG UNIVERSITY

• **Generation-based.** Graph generator generates example graph patterns that maximize the prediction probability of each class. E.g., XGNN, KDD 2020.

• **Global counterfactuals-based.** Find a small set of representative counterfactual graphs that explains all input graphs. E.g., GCFExplainer, WSDM 2023.

• **Global concept-based.** GNN neurons as global concept detectors. E.g., *Global concept-based interpretability for graph neural networks via neuron analysis*, AAAI 2023.

Tutorial Outline

AALBORG UNIVERSITY

1 Introduction

1.1 Graph neural networks (GNNs) and applications
1.2 Interpretability of GNNs
- Definitions, importance, and challenges

2 Taxonomy of interpretability methods for GNNs

2.1 Post-hoc vs. intrinsic / self-explainable
2.2 Global/ class-specific vs. local/ instance-specific
2.3 Model-specific vs. model-agnostic
2.4 Forward vs. backward
2.5 Node-level vs. edge-level vs. subgraph-level
2.6 Perturbation vs. gradient vs. decomposition vs. surrogate models vs. counterfactuals

3 Recent interpretability methods for GNNs

GNNExplainer, PGExplainer , GraphMask, SubgraphX , PGMExplainer, CF2, SA, GuidedBP, CAM, Grad-CAM, LRP, ExcitationBP, and XGNN

4 Benchmark & ground truth for GNN interpretability methods

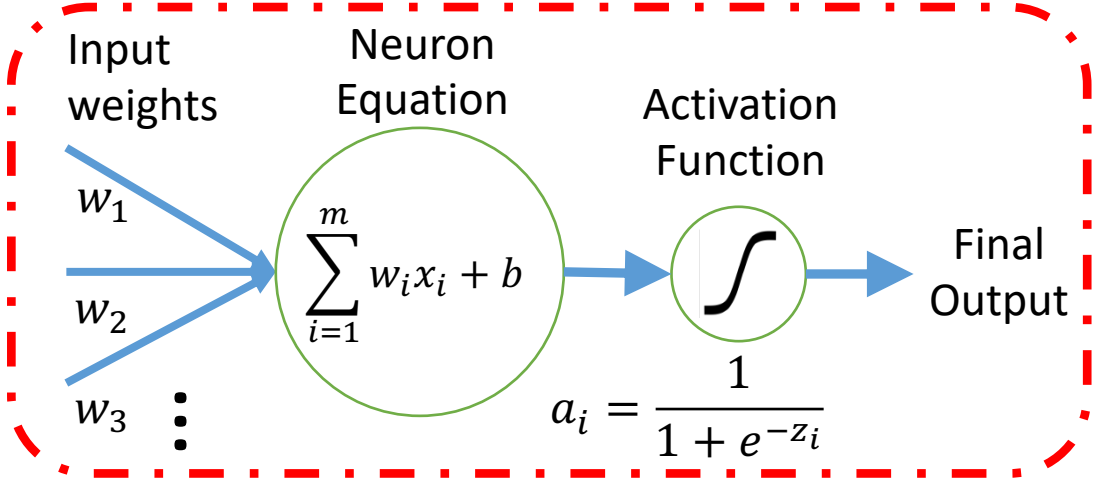
4.1 Interpretability evaluation metrics
4.2 Ground truth datasets, software
4.3 Benchmarking results

5 Future directions

AALBORG UNIVERSITY

Background: Neural Network and Graph Neural Network

Neural Networks

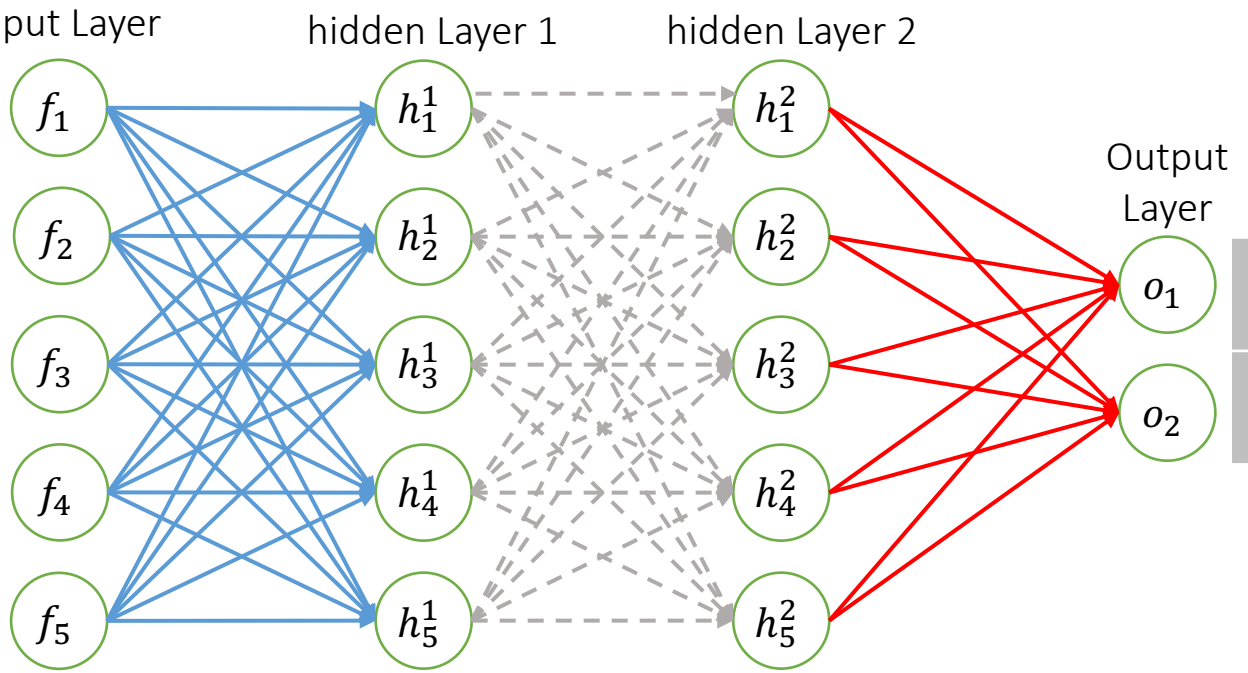


Neuron Equation	$z_i = \sum_{i=1}^{m} w_i x_i + b$
Activation Function	$a_i = \frac{1}{1 + e^{-z_i}}$
Parameter Learning	$w_i^{t+1} = w_i^t - \eta * \frac{\partial L}{\partial w_i}$

f: feature

	f_1	f_2	f_3	f_4	f_5
Data 1	1	0	0	0	0
Data 2	0	0	0	1	0
Data 3	0	0	0	0	1

Data 1
1
0
0
0
0



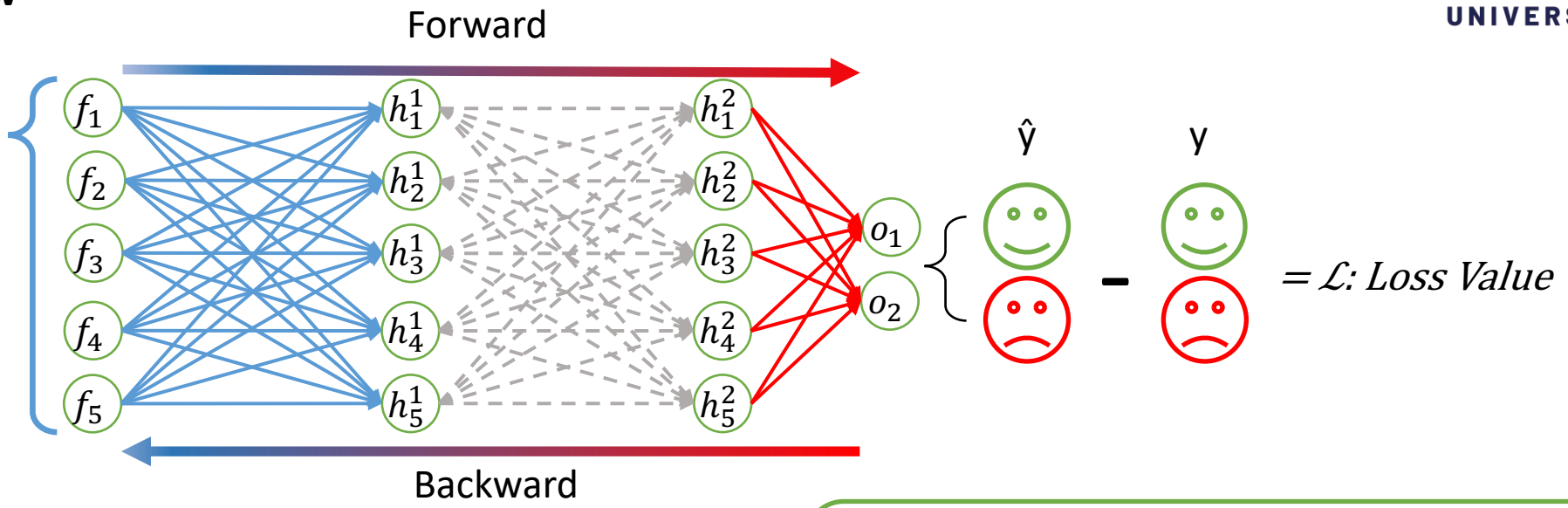
Probability	
0.2	Class 1
0.8	Class 2

Back Propagation on Neural Networks

AALBORG UNIVERSITY

Back Propagation Overview

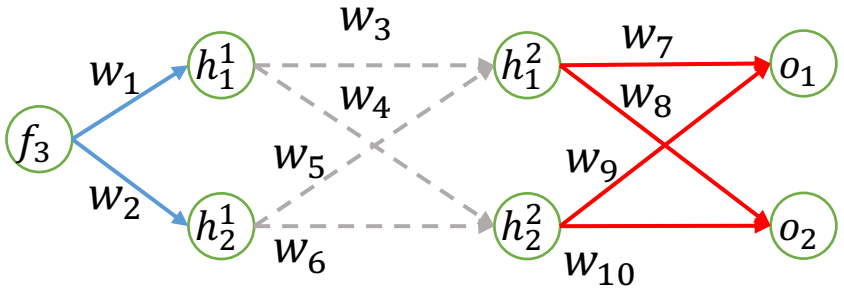
	f_1	f_2	f_3	f_4	f_5
Data 1	1	0	0	0	0
Data 2	0	0	0	1	0
Data 3	0	0	0	0	1



N Datapoints

$$z_{o_1} = \sum_{i=1}^{m=2} w_i x_i + b \quad \textbf{and} \quad a_{o_1} = \frac{1}{1 + e^{-z_{o_1}}}$$

Back Propagation Details



$$\frac{\partial \mathcal{L}}{\partial w_7} = \frac{\partial \mathcal{L}}{\partial o_1} * \frac{\partial o_1}{\partial a_{o_1}} * \frac{\partial a_{o_1}}{\partial z_{o_1}} * \frac{\partial z_{o_1}}{\partial w_7}$$

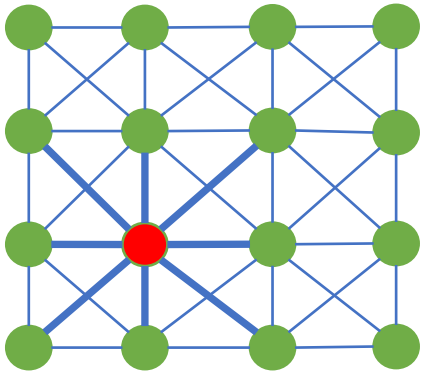
$$w_7^{t+1} = w_7^t - \eta * \frac{\partial \mathcal{L}}{\partial w_7}$$

Gradient production for parameters

$$w_7^{t+1} = w_7^t - \eta * \frac{\partial \mathcal{L}}{\partial o_1} * \frac{\partial o_1}{\partial a_{o_1}} * \frac{\partial a_{o_1}}{\partial z_{o_1}} * \frac{\partial z_{o_1}}{\partial w_7}$$

Grid-Data vs. Graph-Data

Grid {e.g., pixels in images, words in text}



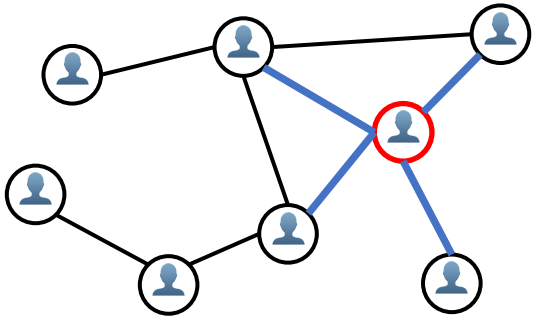
Feature Matrix

1	0	1	0
0	1	0	0
0	1	0	1
1	0	0	1

VS.

!?

Graph



Feature Matrix

	f_1	f_2	f_3	f_4
	1	0	1	0
	0	0	1	0
	0	0	0	1
	1	1	0	0

- Fixed number of neighbors.
- Order of pixels is fixed; thus, no need for adjacency matrix.
- By changing order of pixels, semantic meaning disturbs.

- Variable number of neighbors.
- Nodes are order invariant; thus, need both feature matrix and adjacency matrix.

Adjacency Matrix

	1	1	0	0
	1	0	1	0
	0	1	0	1

We need a different neural network for graphs.

Graph Convolutional Neural Network (GCN)

- Graph convolutional layers.
- Node features update by accumulation of neighborhood features.

$$H^{(t+1)} = \boldsymbol{\sigma}(\widetilde{D}^{-\frac{1}{2}} . \widetilde{A} . \widetilde{D}^{-\frac{1}{2}} . H^{(t)} . W^{(t)}) \qquad \widetilde{A} = A + I_N \qquad H^{(0)} = X$$ Input Node Features

SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS. ICLR 2017.

$$H^{(t+1)} = \underbrace{\boldsymbol{\sigma}\left(\widetilde{D}^{-\frac{1}{2}} . A . \widetilde{D}^{-\frac{1}{2}} . H^{(t)} . W^{(t)}\right)}_{\text{AGGREGATE}} + \underbrace{\boldsymbol{\sigma}\left(\widetilde{D}^{-\frac{1}{2}} . I_N . \widetilde{D}^{-\frac{1}{2}} . H^{(t)} . W^{(t)}\right)}_{\text{COMBINE}}$$

- GCN combines both feature matrix and adjacency matrix.
- More GCN layers imply accumulation of neighborhood information from higher depths.

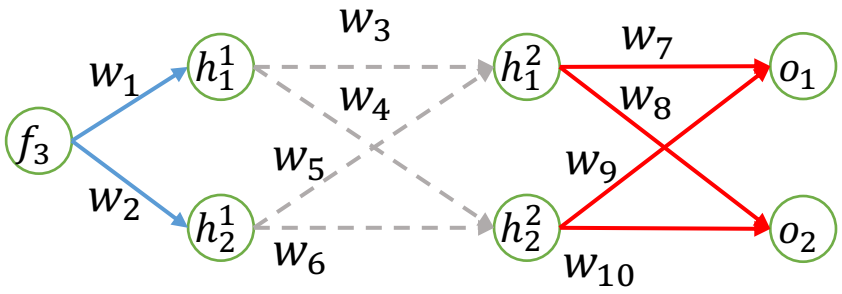
Gradient-Based GNN Interpretability Methods:

- Sensitivity Analysis (SA) [CVPR 2019]
- Guided BackPropagation (GuidedBP) [CVPR 2019]
- Class Activation Mapping (CAM) [CVPR 2019]
- Gradient-weighted Class Activation Mapping (Grad-CAM) [CVPR 2019]

Inspired by neural network interpretation of grid-structured data.
Instance-level.

Sensitivity Analysis

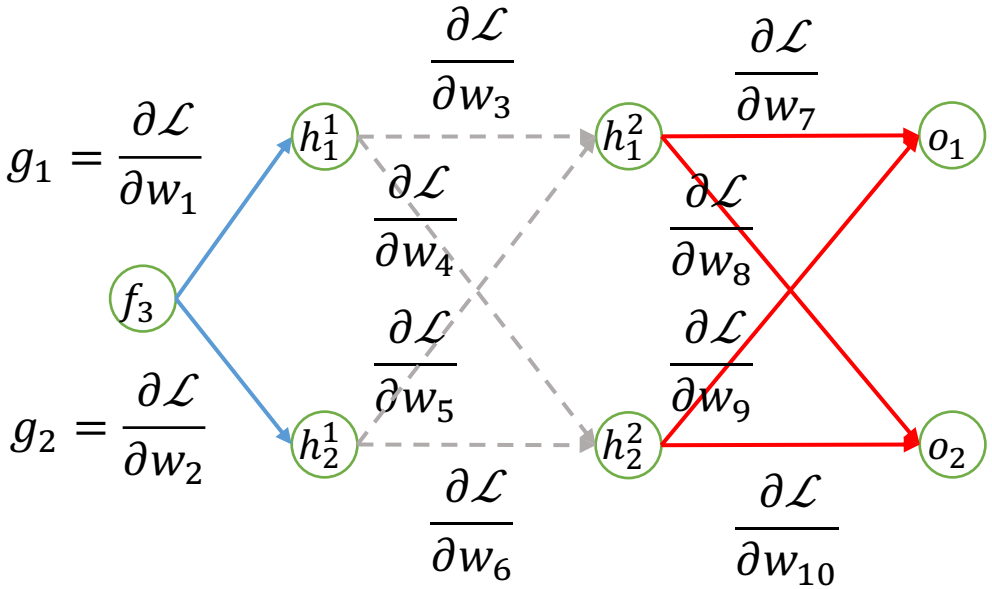
- Summing squared of gradients in the input layer.
- N: number of connected neurons in next layer.
- Attribution_Score(f_i) = $\sum_{j=1}^{N}(g_j)^2$
- E.g., Attribution_Score(f_3) = $\sum_{j=1}^{2}(g_j)^2$



Explainability Methods for Graph Convolutional Neural Networks

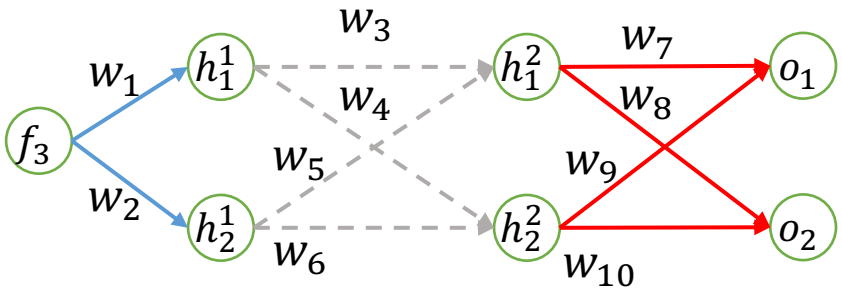
2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Martin E. Charles, Heiko Hoffmann



GuidedBP

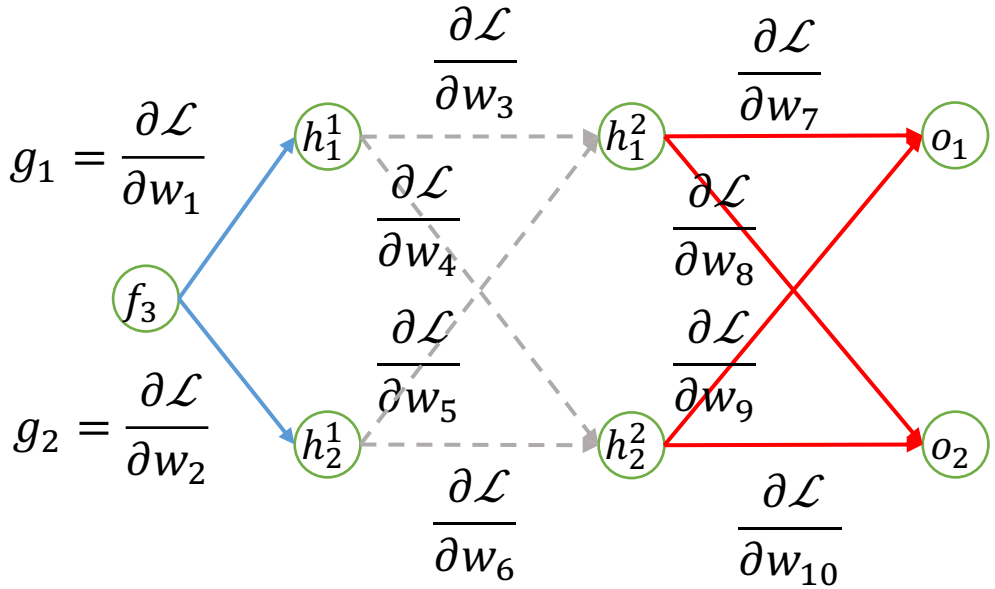
- Summing positive values of gradients in the input layer as the attribution scores.
- N: number of connected neurons in next layer.
- Attribution_Score(f_i) = $\sum_{j=1}^{N} \max(g_j, 0)$
- E.g., Attribution_Score(f_3) = $\sum_{j=1}^{2} \max(g_j, 0)$



Explainability Methods for Graph Convolutional Neural Networks

2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Martin E. Charles, Heiko Hoffmann



Class Activation Mapping (CAM)

- Weights of the predicted class are multiplied by the outputs of the last graph convolutional layer (GCN).
- For node n_i in the input graph:
- Attribution_Score(n_i) =$\sum_{k=1}^{m} w_k^c F_k(i)$

E.g., Attribution_Score(n_1) = $\sum_{k=1}^{5} w_k^{o_1} F_k(1)$

GAP: $Global\ Average\ Pooling\ layer$.

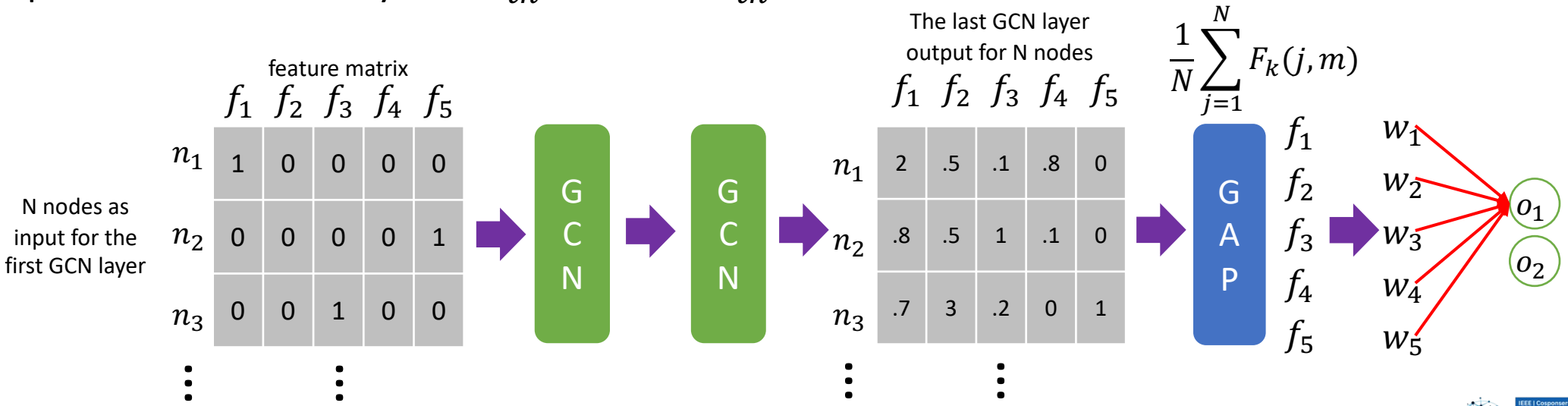
m: # of features per node, f_i: i_{th} node feature.

$F_k(i)$: output of the last GCN layer for i_{th} node and k_{th} feature.

Explainability Methods for Graph Convolutional Neural Networks

2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Martin E. Charles, Heiko Hoffmann



Gradient-Weighted Class Activation Mapping (Grad-CAM)

AALBORG UNIVERSITY

Explainability Methods for Graph Convolutional Neural Networks

2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

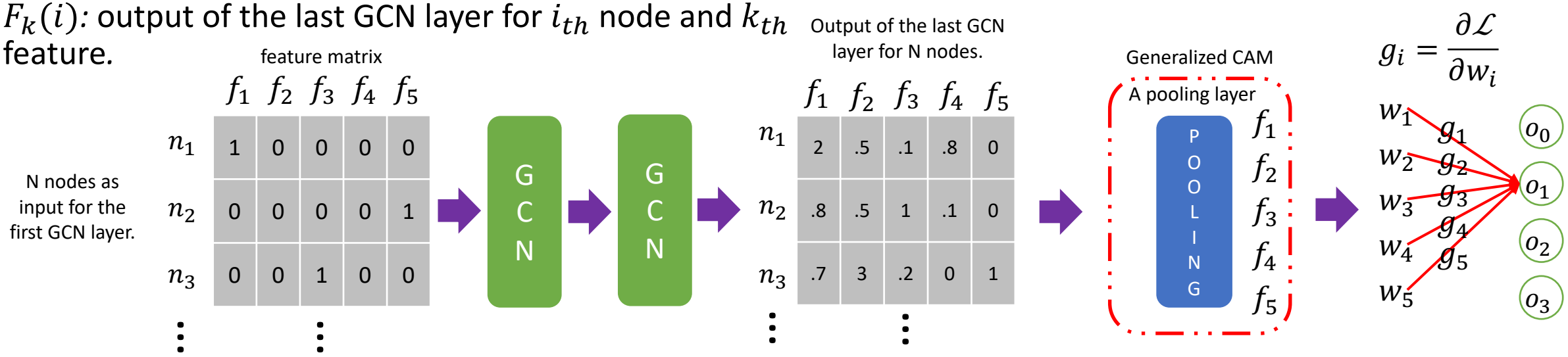
Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Martin E. Charles, Heiko Hoffmann

- Gradients of the predicted class are multiplied by the outputs of the last graph convolutional layer (GCN).
- For node n_i in the input graph:
- Attribution_Score(n_i) = $\sum_{k=1}^{m} \frac{\partial \mathcal{L}}{\partial w_k^c} F_k(i)$

E.g., Attribution_Score(n_1) = $\sum_{k=1}^{5} \frac{\partial \mathcal{L}}{\partial w_k^{o_1}} F_k(1)$

m: # of features per node, f_i: i_{th} node feature.

$F_k(i)$: output of the last GCN layer for i_{th} node and k_{th} feature.



Objectives

AALBORG UNIVERSITY

Method Specific

SA:
Squared gradients emphasize the rate of change (only magnitude, disregard direction) in the model output.

CAM:
Explaining by last GCN layer should be more semantically meaningful compared to input space.

GuidedBP:
Positive gradients point out to the direction of maximum positive rate of change in the model output.

Grad-CAM:
Generalize CAM on neural network architecture.

Pros and Cons

Category Specific
(Gradient-based)

Pros:

- Explanations are quite time-efficient.
- Applicable on different downstream tasks (compatible with GNNs).
- Simple explanations.

Cons:

- Dependent on topology of model (white-box).
- Sensitivity between input and output does not necessarily imply importance.
- Misleading explanations due to the gradient saturation.
- Generate importance scores for nodes, but not for subgraphs.
- Complicated details for non-experts.

Explainability in graph neural networks: A taxonomic survey. IEEE transactions on pattern analysis and machine intelligence

Decomposition-based GNN Interpretability Methods:

- Layer-wise Relevance Propagation (LRP) [PLOS ONE 2015]
- Excitation BackPropagation (EBP) [IJCV 2018]
- GNN-LRP [IEEE TPAMI 2021]

Inspired by neural network interpretation of grid-structured data.
Instance-level.

Difference w.r.t. gradient-based methods:

1. Decomposition methods decompose final output of the model by their own formulations, not by backpropagation rules.
2. Consider output directly, not gradient w.r.t. output.

Layer-wise Relevance Propagation

- Decomposes the final prediction score of a GCN model w.r.t. a class back to the input nodes by using weights and output value of each neuron.
- Attribution_Score(R_i):

$$R_i = \sum_j \frac{a_i w_{ij}}{\epsilon + \sum_i a_i w_{ij}} R_j, \quad \epsilon = 10^{-16}$$

On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation

PloS one 2015

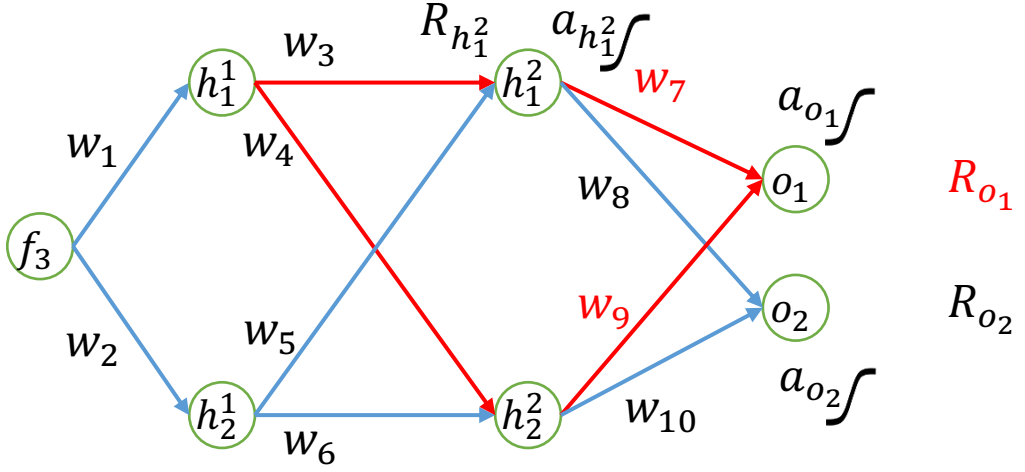
Sebastian Bach, Alexander Binder, Gregoire Montavon, Frederick Klauschen, Klaus-Robert Müller, Wojciech Samek

E.g.,

$$R_{h_1^2} = \frac{a_{h_1^2} w_7}{\epsilon + a_{h_1^2} w_7 + a_{h_2^2} w_9} R_{o_1}$$

$$R_{h_2^2} = \frac{a_{h_2^2} w_9}{\epsilon + a_{h_1^2} w_7 + a_{h_2^2} w_9} R_{o_1}$$

$$R_{h_1^1} = \frac{a_{h_1^1} w_3}{\epsilon + a_{h_1^1} w_3 + a_{h_2^1} w_5} R_{h_1^2} + \frac{a_{h_1^1} w_4}{\epsilon + a_{h_1^1} w_4 + a_{h_2^1} w_6} R_{h_2^2}$$



Excitation Back Propagation (ExcitationBP)

- Decomposes target probability w.r.t. a class into several conditional probability terms by using weights and output values of each neuron.
- Follows the probabilistic Winner-Take-All process and is quite similar approach to LRP
- Attribution_Score(P_i):

$$P_i = \sum_j \frac{a_i w_{ij}}{\epsilon + \sum_i a_i w_{ij}} P_j, \quad \epsilon = 10^{-16}$$

E.g., $$P_{h_1^2} = \frac{a_{h_1^2} w_7}{\epsilon + a_{h_1^2} w_7 + a_{h_2^2} w_9} P_{o_1}$$

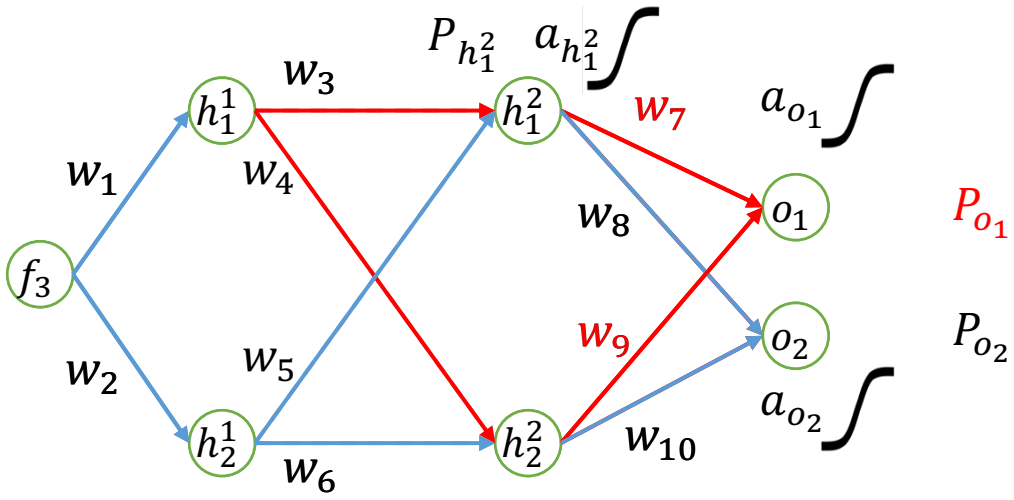
$$P_{h_2^2} = \frac{a_{h_2^2} w_9}{\epsilon + a_{h_1^2} w_7 + a_{h_2^2} w_9} P_{o_1}$$

$$P_{h_1^1} = \frac{a_{h_1^1} w_3}{\epsilon + a_{h_1^1} w_3 + a_{h_2^1} w_5} P_{h_1^2} + \frac{a_{h_1^1} w_4}{\epsilon + a_{h_1^1} w_4 + a_{h_2^1} w_6} P_{h_2^2}$$

Top-down neural attention by excitation backprop

International Journal of Computer Vision 2017

Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, Stan Sclaroff



Pros and Cons

Method Specific Objectives

Layer-wise Relevance Propagation:

- Backpropagating the final output to each neuron by using weights and neuron outputs, could determine contribution of each node.

Excitation BackPropagation:

- Backpropagating the final probability to each neuron by using weights and neuron outputs, could determine contribution of each node.

Category Specific (Decomposition-based)

Pros:

- Explanations are time-efficient.
- Applicable on different downstream tasks (compatible with GNNs).
- Simple explanations.
- Time efficient compared to perturbation-based methods, but inefficient compared to gradient-based methods.

Cons:

- Generate importance scores for nodes, but not for subgraphs.
- Depend on the topology of GNN models (white-box).
- Complicated details for non-experts.

Explainability in graph neural networks: A taxonomic survey. IEEE transactions on pattern analysis and machine intelligence

Perturbation-Based GNN Interpretability Methods:

- GNNExplainer [NeurIPS 2019]
- PGExplainer [NeurIPS 2020]
- GraphMask [ICLR 2021]
- SubgraphX [ICML 2021]
- ZORRO [TKDE 2023]
- Causal Screening [IEEE Trans. Pattern Anal. Mach. Intell. 2023]

Provide instance-level interpretations, by masking input features and graph structures.
(except PGExplainer and GraphMask that provide global interpretability).

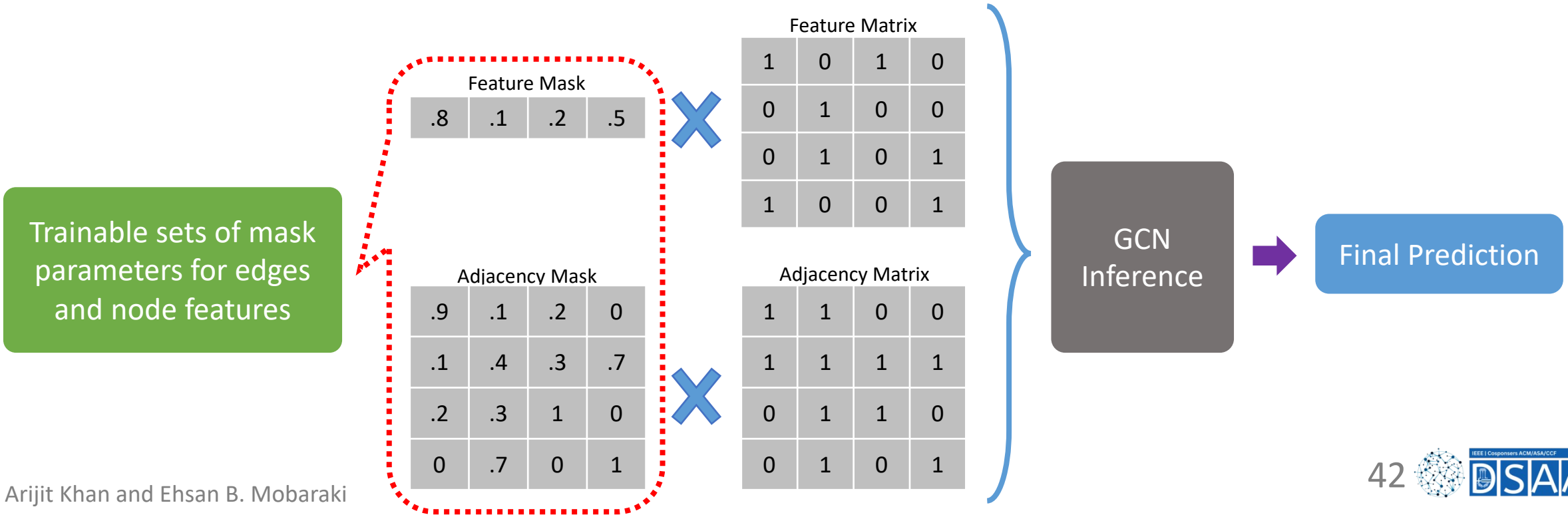
GNNExplainer

- Generates soft masks for edges and node features; and multiplies those masks with node features and edges.
- $\max_{G_s} MI\,(Y, (G_s, X_s)) = H(Y) - H(Y|G = G_s,\ X = X_s).$
- The interpreter gets trained separately on one input sample and generates customized interpretations (i.e., local interpretations).

GNNExplainer: Generating explanations for graph neural networks

Advances in neural information processing systems 2019

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec



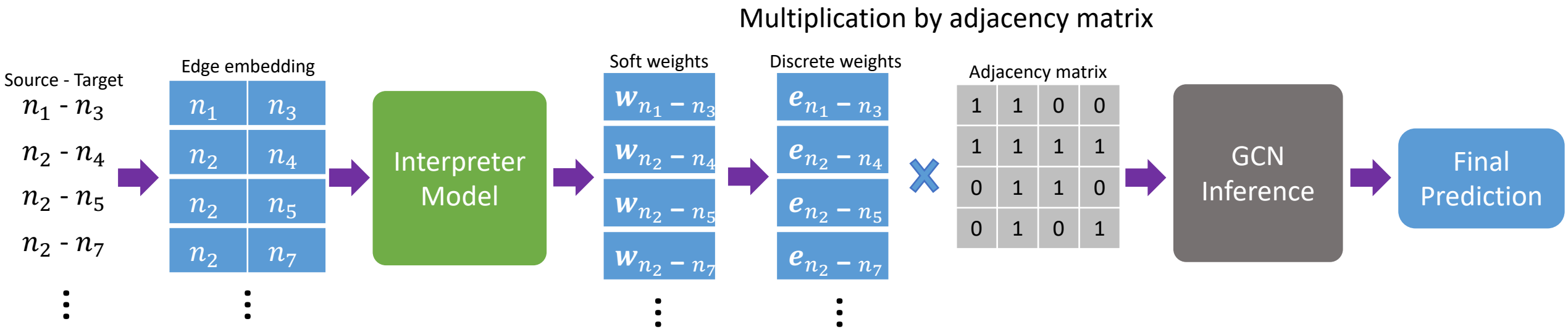
PGExplainer

- Generates discrete masks for edges (only).
- Global interpretations (i.e., not instance-specific).
- Element-wise multiplication of masks by edges.
- $\max_{G_s} MI\,(Y,\ G_s) = H(Y) - H(Y|G = G_s)$
- $e_{n_i - n_j} = \sigma((\log \epsilon - \log(1-\epsilon) + w_{n_i - n_j})/\tau)$

Parameterized explainer for graph neural network

Advances in neural information processing systems 2020

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang



GraphMask

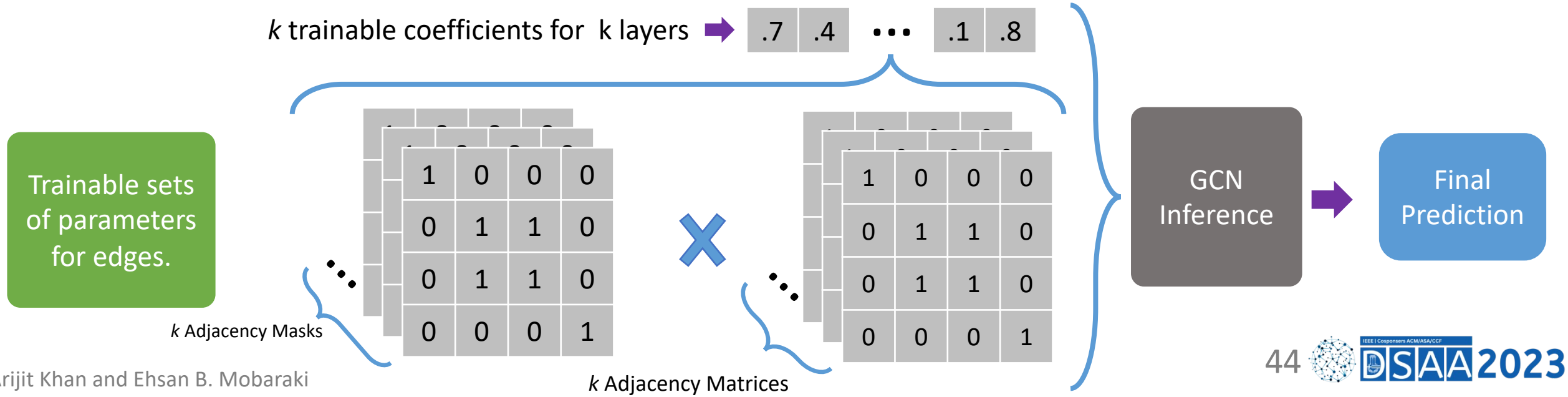
- Generates discrete masks for edges.
- Element-wise multiplication of masks by edges.
- Similar approach to PGExplainer.
- Aggregates edges in every GNN layer by masks and layer coefficients.

Interpreting graph neural networks for NLP with differentiable edge masking

International Conference on Learning Representations, ICLR 2021

Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov

$$\max_{\lambda} \min_{\pi,b} \sum_{g,\, X \in D} \left(\sum_{k=1}^{L} \sum_{(u,v)\in\varepsilon} 1_{R\neq 0}\left(z_{u,v}^{(k)}\right) \right) + \lambda(D_*[f(g,X)\|f(g_s,X)] - \beta)$$



SubGraphX

- Explains by generating important subgraphs for the input graph.
- $g^* = \underset{|g_i| \le N_{min}}{argmax}\, Score(f(.), g, g_i),$

1. Monte Carlo Tree Search for subgraph exploration.
2. Shapley Values on GCN outputs for subgraph selection.
3. Relaxing Subgraphs domain:
   - L-hop neighborhoods
   - Monte Carlo Sampling.

$$P = \{g_i, v_{k+1}, \dots, v_m\}$$

$$a^* = \underset{a_j}{argmax} \frac{W(N_i, a_j)}{C(N_i, a_j)} + \lambda . R(N_i, a_j) \frac{\sqrt{\sum_k C(N_i, a_k)}}{1 + C(N_i, a_j)}$$

$$(f(.), g, g_i) = \sum_{s \subseteq P \setminus \{g_i\}} \frac{P}{\binom{P-1}{s}} (f(s \cup \{g_i\}) - f(s))$$

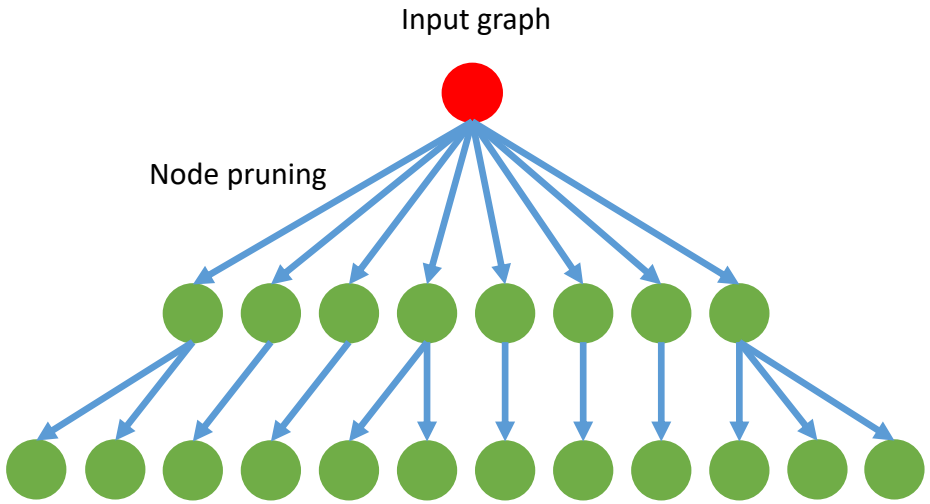
$$C(N_i, a_j) = C(N_i, a_j) + 1$$

$$W(N_i, a_j) = W(N_i, a_j) + Score(f(.), g, g_l)$$

On explainability of graph neural networks via subgraph explorations

International conference on machine learning 2021

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji



Objectives

Method Specific
(Perturbation-based)

GNNExplainer:
The first model-agnostic GNN interpretability method. The method generates a subgraph and small subset of critical node features.

PGExplainer:
Adopts a neural network to parameterize interpretation process, which enables it to interpret multiple instances collectively.

GraphMask:
Adopts a neural network to parameterize interpretation process by doing weighted sum of k GCN layers output.

SubgraphX:
An easy way to explain a graph is to generate subgraphs and select the most important one as an explanation.

Pros and Cons

Method Specific (Perturbation-based)

Pros:

GNNExplainer:
- Interpreting by node features and edges.
- Time efficient compared to SubgraphX, but inefficient compared to gradient-based and decomposition-based methods.

PGExplainer:
- Global interpretation.
- Interpretation by discrete masks.
- Time efficient compared to SubgraphX, but inefficient compared to gradient-based and decomposition-based methods.

GraphMask:
- Global interpretation.
- Interpretation by discrete masks.
- Time efficient compared to SubgraphX, but inefficient compared to gradient-based and decomposition-based methods.

SubgraphX:
- Interpretation by discrete masks.
- Subgraph-based interpretation.

Cons:

GNNExplainer:
- Local interpretation (instance-specific).
- Interpretation by soft masks.
- Introduced evidence.
- Important elements are not guaranteed to be connected.

Explainability in graph neural networks: A taxonomic survey. IEEE transactions on pattern analysis and machine intelligence

PGExplainer:
- Important elements are not guaranteed to be connected (a regularization term is suggested to encourage connectivity).

On explainability of graph neural networks via subgraph explorations. In International conference on machine learning 2021.

GraphMask:
- Important elements are not guaranteed to be connected.

SubgraphX:
- Local interpretation (instance-specific).
- Time-inefficient interpretation.

Surrogate-Based GNN Interpretability Methods:

- PGMExplainer [NeurIPS 2020]
- RelEx [AIES 2021]
- GraphLime [IEEE TKDE 2023]

Provide instance-level interpretations by approximation on the output of a GCN.

PGMExplainer

- Generates interpretation by eliminating unimportant elements.
- Three steps:

1. **Data generation**: add noise to the node features.
2. **Node selection**: Markov Blanket of the node.
3. **Structure learning**: Chi-Squared Test.

$$Score_{BIC} = \ell(D_t[U(t)]) - \frac{\log n}{2} Dim(\beta)$$

$$x_i^2 = \frac{(O_i - E_i)^2}{E_i}$$

Markov Blanket of a node: minimum set of nodes that are conditionally independent of the node.

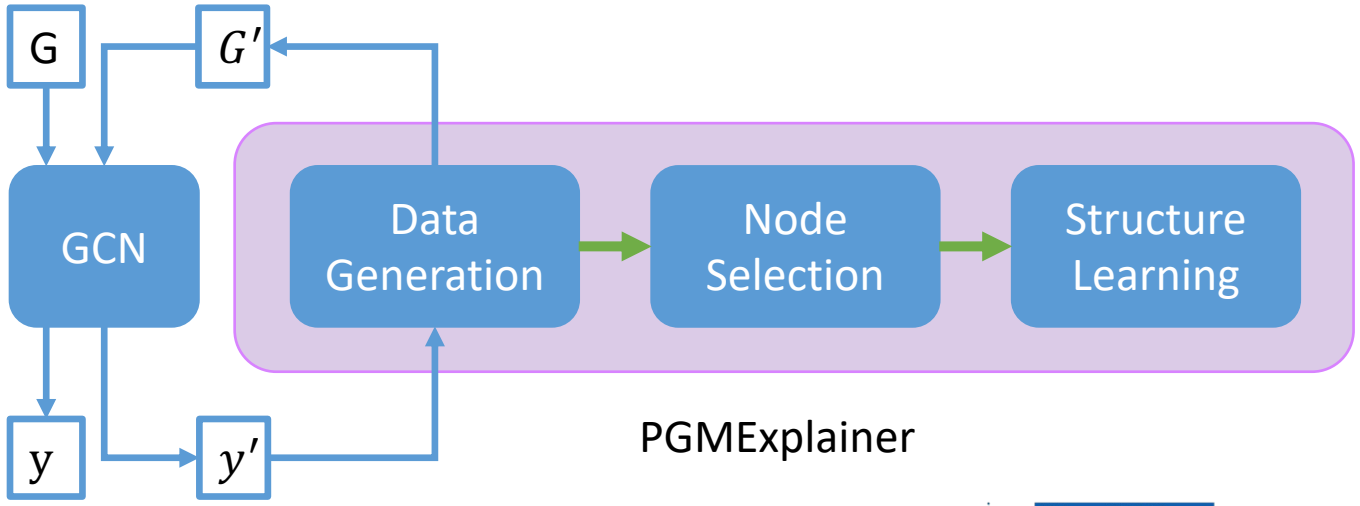
Chi-Squared Test: how much the observed frequencies for a categorical variable match the expected frequencies.

PGM-Explainer: Probabilistic graphical model explanations for graph neural networks

Advances in neural information processing systems 2020

Minh Vu and My T Thai

G = Input graph G' = Perturbed G



PGMExplainer

PGMExplainer

Method Specific
(Surrogate-based)

PGMExplainer:
Pros:

- Time-efficient interpretation compared to SubgraphX, inefficient compared to gradient-based methods.
- Provides interpretation for graph and node classification.

PGMExplainer:
Cons:

- Local interpretation (instance-specific).
- Important elements are not guaranteed to be connected.

Explainability in graph neural networks: A taxonomic survey. IEEE transactions on pattern analysis and machine intelligence

Counterfactual reasoning:

- CounterFactual and Factual (CF2) [ACM Web Conference 2022]
- RCExplainer [NeurIPS 2021]

Provide instance-level interpretations by considering factual and counterfactual reasoning systems.

CF2

CounterFactual and Factual

- ❖ Strength: { 1. Sufficient: Factual; 2. Necessary: CounterFactual } Effectiveness
- ❖ Simpleness: simple interpretation is preferred (Occam Razor's principle).
- ❖ Mask generation follows a perturbation-based technique.
- ❖ Objective is to:
  1. Minimize interpretation complexity.
  2. Interpretation is sufficient and necessary.

- Factual reasoning:

$$\underset{c \in C}{\operatorname{argmax}} P_{\emptyset}(c | A_k \odot M_k, X_k \odot F_k) = \hat{y}_k$$

- Counterfactual reasoning:

$$\underset{c \in C}{\operatorname{argmax}} P_{\emptyset}(c | A_k - A_k \odot M_k, X_k - X_k \odot F_k) \neq \hat{y}_k$$

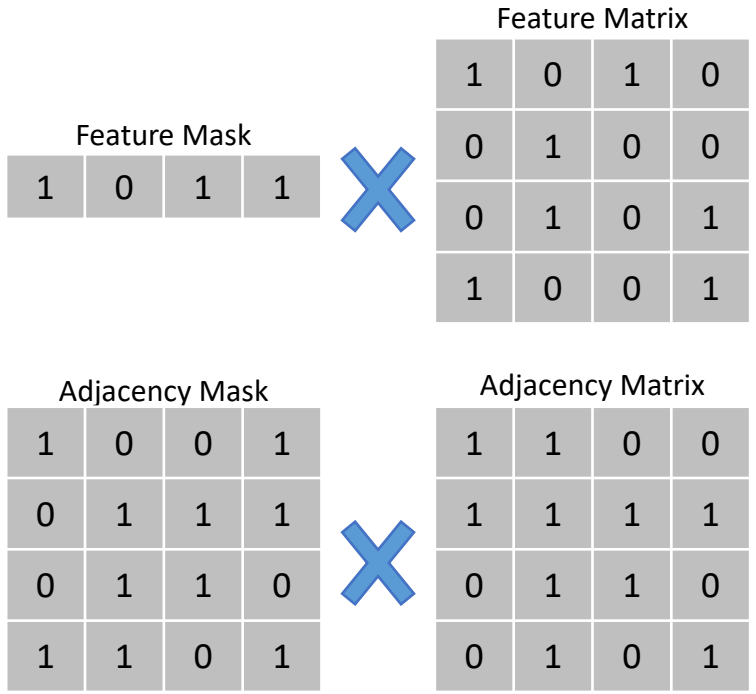
- Simpleness measurement:

$$C(M, F) = \|M\|_1 + \|F\|_1$$

Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning

Proceedings of the ACM Web Conference 2022

Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang



AALBORG UNIVERSITY

Method Specific
(Counterfactual reasoning)

Pros:

- Interpretation by node features and edges.
- Counterfactual and factual reasoning.

Cons:

- Local interpretation (instance-specific).
- Important elements are not guaranteed to be connected.
- Time-efficient compared to SubgraphX, but inefficient compared to gradient-based, surrogate-based, and decomposition-based methods.

Model-Level Interpretations:

- **XGNN [SIGKDD 2020]**

Provide model-level interpretations by using a graph generator.

XGNN

- Find best input patterns for the given input graph and GNN model.

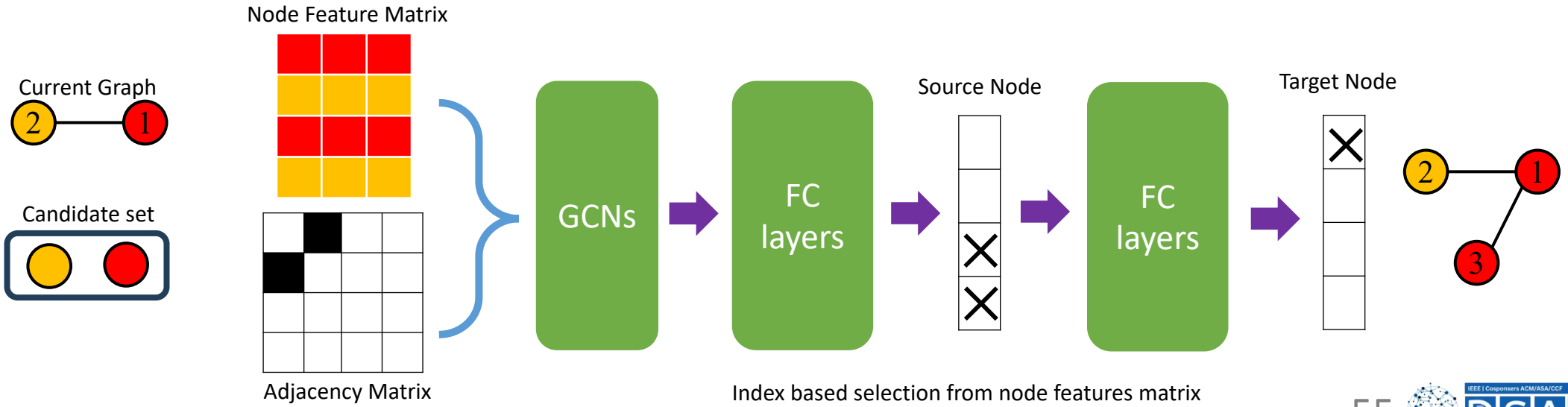
$$G^* = \max_G P(f(G) = c_i)$$

- Generates graph by a graph generator that maximizes Reinforcement Learning reward (i.e., R_t) on the following objective function:
- $R_t = R_{t,f}(G_{t+1}) + \lambda_1 \frac{\sum_{i=1}^{m} R_{t,f}(Rollout(G_{t+1}))}{m} + \lambda_2 R_{t,r}$
- $R_{t,f}(G_{t+1}) = p(f(G_{t+1}) = c_i) - 1/\ell$
- If $R_t < 0$: $G_{t+1} = G_t$

XGNN: Towards Model-Level Explanations of Graph Neural Networks

Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2020

Hao Yuan, Jiliang Tang, Xia Hu, Shuiwang Ji



AALBORG UNIVERSITY

Method Specific
(Model-Level Interpretation)

Pros:

- Global interpretation.
- Class-discriminative interpretation.
- Subgraph-based interpretation.

Cons:

- Only graph classification.
- The generated subgraph still might be unrealistic (however, some graph rules are incorporated to encourage the explanations to be valid and human-intelligible).

Explainability in graph neural networks: A taxonomic survey. IEEE transactions on pattern analysis and machine intelligence

Tutorial outline

AALBORG UNIVERSITY

1 Introduction

1.1 Graph neural networks (GNNs) and applications
1.2 Interpretability of GNNs
- Definitions, importance, and challenges

2 Taxonomy of interpretability methods for GNNs

2.1 Post-hoc vs. intrinsic / self-explainable
2.2 Global/ class-specific vs. local/ instance-specific
2.3 Model-specific vs. model-agnostic
2.4 Forward vs. backward
2.5 Node-level vs. edge-level vs. subgraph-level
2.6 Perturbation vs. gradient vs. decomposition vs. surrogate models vs. counterfactuals

3 Recent interpretability methods for GNNs

GNNExplainer, PGExplainer , GraphMask, SubgraphX , PGMExplainer, CF2, SA, GuidedBP, CAM, Grad-CAM, LRP, ExcitationBP, and XGNN

4 Benchmark & ground truth for GNN interpretability methods

4.1 Interpretability evaluation metrics
4.2 Ground truth datasets, software
4.3 Benchmarking results

5 Future directions

Interpretability Evaluation Metrics:

Provide quantitative measurements on the performance of interpretability methods.

- Faithfulness (Fidelity+ / Comprehensiveness, Fidelity- /Sufficiency/ Validity)
- Contrastivity
- Sparsity
- Accuracy
- Stability/ Robustness (RDT Fidelity)
- Consistency
- Interpretation time
- Plausibility 

Faithfulness

- Do explanation results faithfully explain the behaviors of GNN models?

Fidelity+ /Comprehensiveness:

- Are all nodes/edges/features in the graph needed to make a prediction selected in the explanation?
- Occlusion of explanation elements should decrease prediction accuracy.
- High fidelity+ /comprehensiveness score is better.

$$comprehensiveness = f\left(\mathcal{G}\right)_j - f\left(\mathcal{G}\backslash\mathcal{G}'\right)_j$$

Explainability in graph neural networks: A taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell., 2022.

Fidelity- /Sufficiency / Validity:

- Is explanation sufficient for the original prediction?
- Occlusion of non-explanation elements should not decrease prediction accuracy.
- Low fidelity- /sufficiency score is better.

$$sufficiency = f\left(\mathcal{G}\right)_j - f\left(\mathcal{G}'\right)_j$$

Contrastivity

- Explanations for different classes should be significantly different.

- Ratio of the Hamming distance between binarized saliency maps for positive and negative classes, normalized by the total number of salient nodes identified by either method.

$$\frac{d_H(\hat{m_0},\hat{m_1})}{\hat{m_0} \vee \hat{m_1}}$$

Explainability methods for graph convolutional neural networks. in CVPR, 2019.

Sparsity

- Conciseness of explanation.

- Higher sparsity values tend to be better. It indicates that the explanations capture the most important input information.

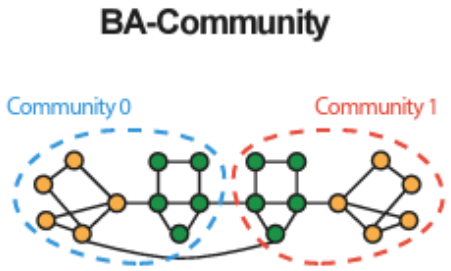
$$(1 - \frac{|m_i|}{|M_i|})$$

Explainability in graph neural networks: A taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell., 2022.

Accuracy

- Applicable when ground-truth explainability is available (e.g., synthetic datasets).

- Metrics can include general accuracy, F1 score, ROC-AUC score.



GNNExplainer: generating explanations for graph neural networks. NeurIPS 2019.

Stability/ Robustness

- When small changes are applied to the input without affecting the predictions, the explanations should remain similar.

- Stable explanations are better.

RDT Fidelity:

- Principles of rate-distortion theory.

Perturbed input: $Y_{\mathcal{S}} = X \odot M(\mathcal{S}) + Z \odot (\mathbb{1} - M(\mathcal{S})), Z \sim \mathcal{N}$

RDT Fidelity is expected fidelity- of the perturbed input.

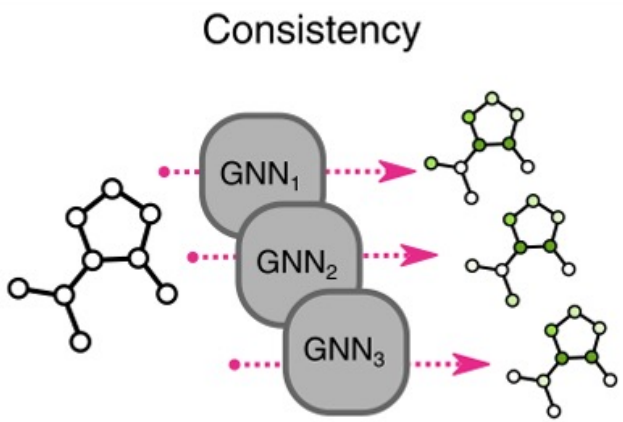
$$\mathcal{F}(\mathcal{S}) = \mathbb{E}_{Y_{\mathcal{S}}|Z\sim\mathcal{N}}\left[\mathbb{1}_{\Phi(X)=\Phi(Y_{\mathcal{S}})}\right]$$

Zorro: Valid, sparse, and stable explanations in graph neural networks. IEEE Transactions on Knowledge & Data Engineering. 35(8), 2023.

Robust counterfactual explanations on graph neural networks. NeurIPS2021

Consistency

- The explanations should be consistent across high-performing GNN models.



Evaluating attribution for graph neural networks. NeurIPS 2020

- **Concern:** Multiple explanations - different high-performing models may capture different relationships.

Interpretation Time

- Time to generate interpretation.
- E.g., perturbation-based methods are generally slower than gradient-based interpretability approaches.
- Useful for human-in-the-loop, qualitative evaluation.

Plausibility

AALBORG UNIVERSITY

- Human-grounded evaluation.
- Agreement of explanation with domain knowledge / expert.
- Metrics can include general accuracy, F1 score, ROC-AUC score.

Ground Truth Datasets (Synthetic)

AALBORG UNIVERSITY

- Motifs added to base graphs as ground truth explanations.
- **Base graphs:** grid, binary tree, Barabasi-Albert (BA) graph, etc.
- **Motifs:**
  - house motif with five nodes, formed by a top, a middle, and a bottom node
  - cycle motif with five or six nodes
  - grid-structured motif, etc.
- **Example synthetic datasets:** BA-Shapes, BA-Community, Tree Cycle, Tree Grids, BA-2Motifs, Spurious Motifs, etc.
- **Synthetic datasets generator:** ShapeGGen https://zitniklab.hms.harvard.edu/projects/GraphXAI/

GNNExplainer: generating explanations for graph neural networks. NeurIPS 2019.

Explainability in graph neural networks: A taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell., 2022.

A survey on explainability of graph neural networks. CoRR, vol. abs/2306.01958, 2023.

Evaluating explainability for graph neural networks. Sci Data, vol. 10, no. 1, 2023

Ground Truth Datasets (Real-world)

AALBORG UNIVERSITY

- Molecular graphs due to domain knowledge (i.e., known chemical properties of the molecules).
- **Examples:** Mutag, NCI1, Tox21, Blood-brain barrier penetration (BBBP), Alkane carbonyl, Fluoride carbonyl, etc.
- **Others:** Graph SST2, Graph SST5, and Graph Twitter, MNIST-75sp, Visual Genome dataset, Recidivism, etc.

GNNExplainer: generating explanations for graph neural networks. NeurIPS 2019.

Explainability in graph neural networks: A taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell., 2022.

A survey on explainability of graph neural networks. CoRR, vol. abs/2306.01958, 2023.

Explainability methods for graph convolutional neural networks. in CVPR, 2019.

Evaluating explainability for graph neural networks. Sci Data, vol. 10, no. 1, 2023.

Software

- **StellarGraph Machine Learning Library** (Data61, CSIRO)

https://stellargraph.readthedocs.io/en/stable/

- **DIG: Dive into Graphs** (DIVE Lab, led by Dr. Shuiwang Ji, Texas A&M University)

https://diveintographs.readthedocs.io/en/latest/

- **DGL** (Deep Graph Library is developed and maintained by NYU, NYU Shanghai, AWS Shanghai AI Lab, and AWS MXNet Science Team)

https://docs.dgl.ai/tutorials/blitz/index.html

- **torch_geometric.explain** (PyG Team)

https://pytorch-geometric.readthedocs.io/en/latest/modules/explain.html

- **GraphXAI** (Zitnik Lab, led by Dr. Marinka Zitnik, Harvard University)

https://zitniklab.hms.harvard.edu/projects/GraphXAI/

AALBORG UNIVERSITY

(Preliminary) Benchmarking Results:

- **Evaluation of Graph Neural Networks**

Evaluation of Graph Neural Networks

AALBORG UNIVERSITY

MUTAG dataset for graph classification.
Optimized hyperparameters w.r.t. AUC-ROC.

GNN Model Name	AUC-ROC	AUC-PR	Acc	Avg Running Time per Epoch [Sec]
GIN	**0.913**	**0.93**	0.835	0.076
GCN+GAP	0.8	0.867	0.714	**0.053**
DIFFPOOL	0.876	0.89	0.824	0.191
DGCNN	0.91	0.872	**0.862**	0.128

HOW POWERFUL ARE GRAPH NEURAL NETWORKS? ICLR 2019.

SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS. ICLR 2017.

Hierarchical Graph Representation Learning with Differentiable Pooling. NeurIPS 2018.

An End-to-End Deep Learning Architecture for Graph Classification. AAAI-2018.

AUC-ROC: Area Under the Curve Receiver Operating Characteristics

AUC-PR: Area Under the Curve Precision Recall

Acc: Accuracy

AALBORG UNIVERSITY

(Preliminary) Benchmarking Results:

- **Evaluation of Interpretability Methods on Graph Neural Networks**

Evaluation of Interpretability Methods

MUTAG dataset for graph classification.
Optimized hyperparameters w.r.t. AUC-ROC.

AALBORG UNIVERSITY

	Fidelity$^+$			
	GCN+GAP	**DGCNN**	**DIFFPOOL**	**GIN**
SA	0.178	0.007	0.044	0.015
GuidedBP	0.104	0.072	0.022	0.016
CAM	0.105	0.097	0.175	0.129
Grad-CAM	**0.283**	0.233	0.194	0.184
GNNExplainer	0.223	**0.276**	0.099	0.096
PGExplaienr	0.126	0.072	0.167	0.196
Graph-Mask	0.154	-0.107	0.061	0.186
SubGraphX	0.0625	-0.053	**0.287**	0.162
LRP	0.035	0.173	0.152	0.138
ExcitationBP	0.112	0.169	0.17	0.101
PGM-Explainer	0.122	0.077	0.242	**0.202**
CF^2	0.101	0.109	0.052	0.126

	Contrastivity			
	GCN+GAP	**DGCNN**	**DIFFPOOL**	**GIN**
SA	0.503	0.146	0.056	0.481
GuidedBP	0.467	0.199	0.176	0.526
CAM	0.548	0.533	0.493	0.507
Grad-CAM	0.476	0.519	0.482	0.489
GNNExplainer	0.618	0.576	0.47	0.442
PGExplaienr	0.509	0.486	0.477	0.512
Graph-Mask	0.558	0.544	0.825	0.721
SubGraphX	0.471	0.601	0.394	0.427
LRP	0.511	0.402	0.519	0.488
ExcitationBP	0.586	0.392	0.433	0.584
PGM-Explainer	**0.747**	**0.744**	**0.746**	**0.754**
CF^2	0.505	0.472	0.479	0.503

Evaluation of Interpretability Methods

MUTAG dataset for graph classification.
Optimized hyperparameters w.r.t. AUC-ROC.

AALBORG UNIVERSITY

	Sparsity			
	GCN+GAP	**DGCNN**	**DIFFPOOL**	**GIN**
SA	0.375	0.418	0.378	0.522
GuidedBP	0.187	0.577	0.561	0.729
CAM	0.492	0.522	0.398	0.471
Grad-CAM	0.453	0.541	0.506	0.513
GNNExplainer	0.495	0.476	0.547	0.472
PGExplaienr	0.516	0.488	0.487	0.498
Graph-Mask	0.721	0.572	0.488	0.613
SubGraphX	**0.794**	**0.705**	**0.727**	**0.794**
LRP	0.631	0.491	0.549	0.583
ExcitationBP	0.688	0.476	0.486	0.547
PGM-Explainer	0.563	0.436	0.436	0.436
CF^2	0.499	0.486	0.49	0.501

	Avg. Explaining Time [Sec]			
	GCN+GAP	**DGCNN**	**DIFFPOOL**	**GIN**
SA	0.003	0.074	0.005	0.006
GuidedBP	0.002	0.075	0.004	**0.006**
CAM	**0.001**	0.019	**0.004**	0.012
Grad-CAM	0.002	**0.004**	0.005	0.009
GNNExplainer	0.657	0.926	1.11	0.78
PGExplaienr	0.066	0.649	0.088	0.079
Graph-Mask	0.036	0.233	0.124	0.129
SubGraphX	4.211	24.051	7.764	12.416
LRP	0.018	0.034	0.078	0.043
ExcitationBP	0.011	0.027	0.069	0.056
PGM-Explainer	1.161	1.287	2.197	0.622
CF^2	0.952	0.636	1.43	2.184

Evaluation of Interpretability Methods

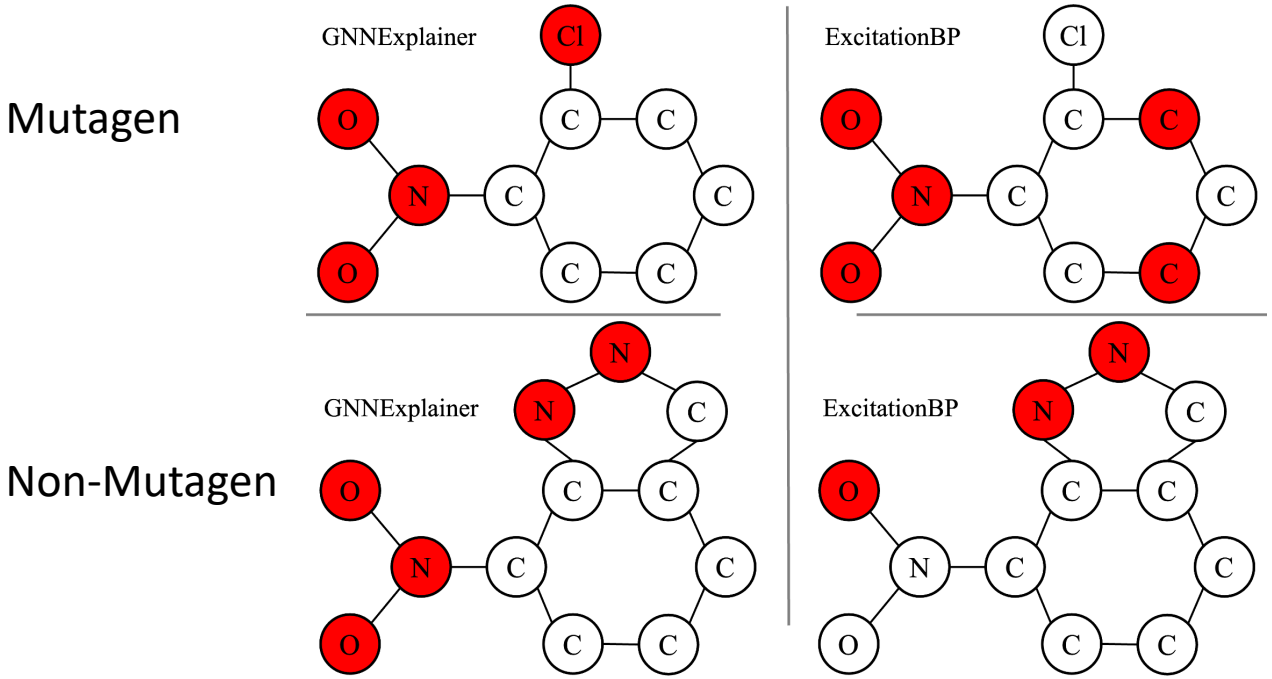
MUTAG dataset for graph classification.
Optimized hyperparameters w.r.t. AUC-ROC.

AALBORG UNIVERSITY

	Fidelity$^+$	Contrastivity	Sparsity	Avg. Explaining Time
GCN+GAP	Grad-CAM	PGM-Explainer	SubgraphX	CAM
DGCNN	GNNExplainer	PGM-Explainer	SubgraphX	Grad-CAM
DIFFPOOL	SubgraphX	PGM-Explainer	SubgraphX	CAM
GIN	PGM-Explainer	PGM-Explainer	SubgraphX	GuidedBP

- The more effective GNN model, the better interpretation result.
- Time Efficiency:
  - Fastest method:
    - CAM, on top of GCN+GAP: **0.001 sec.**
  - Slowest method:
    - SubgraphX, on top of DGCNN: **24.051 sec.**
- GNNExplainer, PGExplainer, SubgraphX, PGMExplainer, and CF^2 are more effective in general, in combination with more effective GNN models.

Case Study for Interpretability Methods on Graph Classification

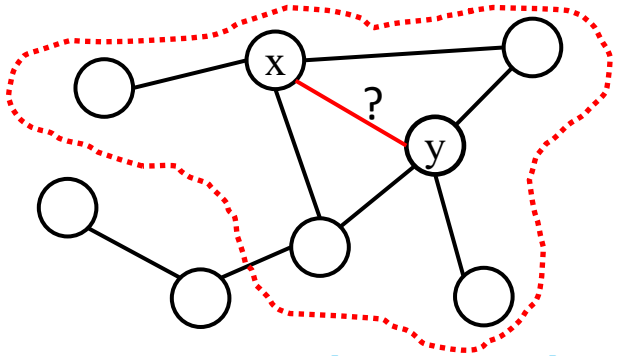


- GNNExplainer is among the top-ranking methods in terms of Fidelity and Contrastivity scores.

Link Prediction Models

SEAL: learning from subgraphs embedding and attributes for link prediction [NeuIPS 2018]

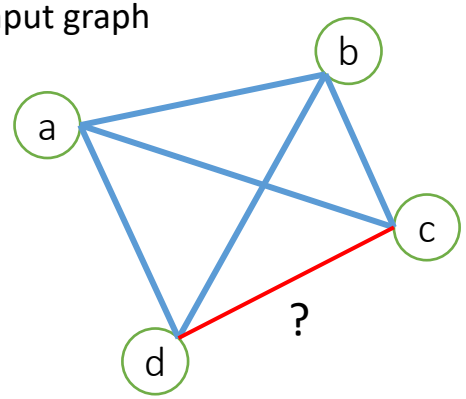
- Takes neighborhood subgraphs of the link and uses GNN to predict the probability of existence for the edges.



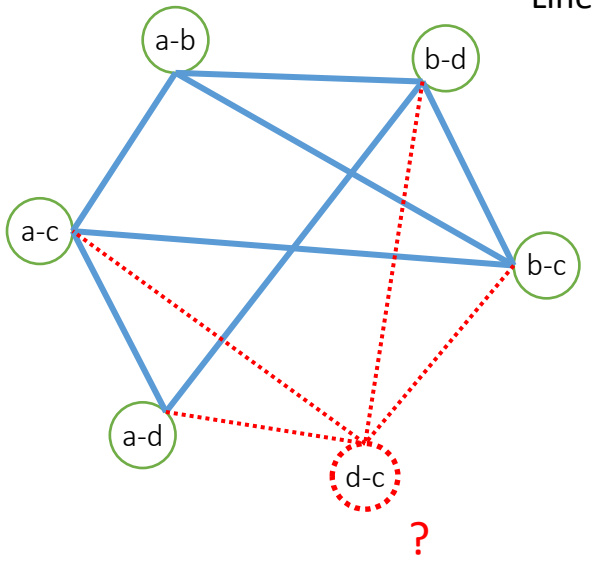
LGLP: Line graph link prediction [TPAMI 2021]

- Converts the input graph to a line graph and performs node labeling.

Input graph

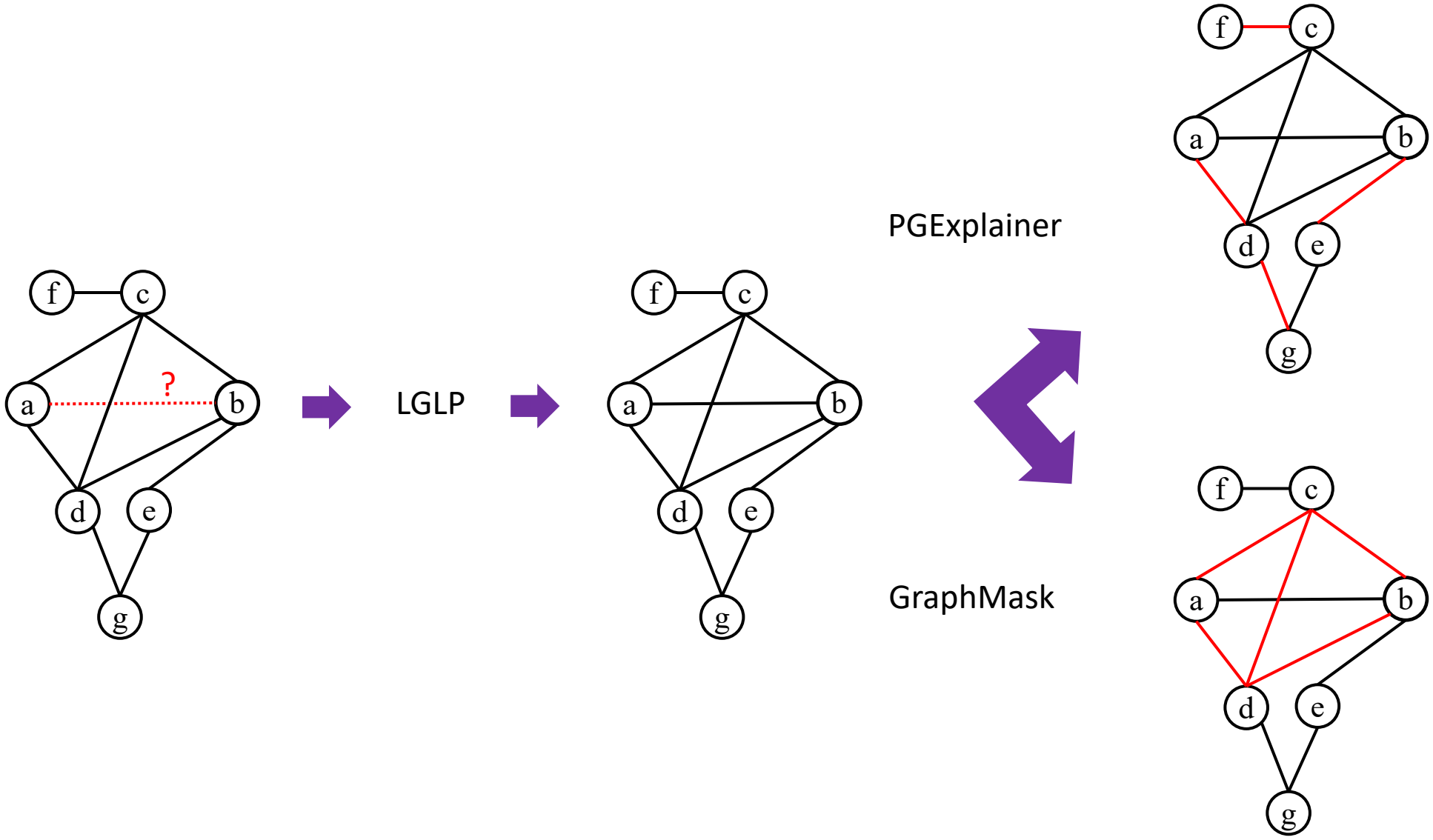


Line graph



Case Study for Interpretability Methods on Link Prediction

Facebook network data



Tutorial outline

1 Introduction

1.1 Graph neural networks (GNNs) and applications
1.2 Interpretability of GNNs
- Definitions, importance, and challenges

2 Taxonomy of interpretability methods for GNNs

2.1 Post-hoc vs. intrinsic / self-explainable
2.2 Global/ class-specific vs. local/ instance-specific
2.3 Model-specific vs. model-agnostic
2.4 Forward vs. backward
2.5 Node-level vs. edge-level vs. subgraph-level
2.6 Perturbation vs. gradient vs. decomposition vs. surrogate models vs. counterfactuals

3 Recent interpretability methods for GNNs

GNNExplainer, PGExplainer , GraphMask, SubgraphX , PGMExplainer, CF2, SA, GuidedBP, CAM, Grad-CAM, LRP, ExcitationBP, and XGNN

4 Benchmark & ground truth for GNN interpretability methods

4.1 Interpretability evaluation metrics
4.2 Ground truth datasets
4.3 Benchmarking results

5 Future directions

Future Directions

AALBORG UNIVERSITY

• Benchmarking interplay of GNN models, graph data, interpretability methods, evaluation metrics, and downstream tasks. Downstream tasks beyond graphs and nodes classification.

• Qualitative evaluation of GNN interpretation – usability, interactive-ness, querying with domain knowledge, trustworthiness, deployment, visualization and HCI tools.

• Obtain real-world ground truth.

• Higher-order explanation, e.g., motif, example, and rule-based explanations.

• Interpretability for more complex graph neural networks, e.g., hypergraph neural networks, temporal graph neural networks, task-agonistic GNN explanation, causal explainer, etc.

• Self-explainable GNNs, explainability to improve GNN prediction, robust and consistent explanations.

• Relation to GNN robustness, fairness, and privacy.

References

• X. Dong, D. Thanou, L. Toni, M. Bronstein, and P. Frossard. *Graph signal processing for machine learning: A review and new perspectives*. ICASSP 2021 Tutorial.

• W. L. Hamilton, R. Ying, J. Leskovec, and R. Sosic. *Representation learning on networks.* WWW 2018 Tutorial.

• Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. *A comprehensive survey on graph neural networks*. IEEE Trans. Neural Networks Learn. Syst. 32(1): 4-24 (2021).

• Z. Zhang, P. Cui, and W. Zhu. *Deep learning on graphs: A Survey*. IEEE Trans. Knowl. Data Eng. 34(1): 249-270 (2022).

• V. P. Dwivedi, C. K. Joshi, A. T. Luu, T. Laurent, Y. Bengio, and X. Bresson. *Benchmarking graph neural networks*. J. Mach. Learn. Res. 24: 43:1-43:48 (2023).

• A. Rossi, D. Barbosa, D. Firmani, A. Matinata, and P. Merialdo. 2021. *Knowledge graph embedding for link prediction: a comparative analysis*. ACM Trans. Knowl. Discov. Data 15, 2 (2021).

• N. Fanourakis, V. Efthymiou, D. Kotzinos, and V. Christophides. 2023. *Knowledge graph embedding methods for entity alignment: experimental review*. Data Mining and Knowledge Discovery 37 (2023), 2070–2137.

• Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, and Q. He. 2022. *A survey on knowledge graph-based recommender systems*. IEEE Trans. Knowl. Data Eng. 34, 8 (2022), 3549–3568.

• S. Bonner, I. P. Barrett, C. Ye, R. Swiers, O. Engkvist, C. T. Hoyt, and W. L. Hamilton. 2022. *Understanding the performance of knowledge graph embeddings in drug discovery*. Artificial Intelligence in the Life Sciences 2 (2022), 100036.

• T. Chowdhury, C. Ling, X. Zhang, X. Zhao, G. Bai, J. Pei, H. Chen, and L. Zhao. 2023. *Knowledge-enhanced neural machine reasoning: a review*. CoRR abs/2302.02093 (2023).

• H. Ren, M. Galkin, M. Cochez, Z. Zhu, and J. Leskovec. 2023. *Neural graph reasoning: complex logical query answering meets graph databases*. CoRR abs/2303.14617 (2023).

• W. Zhang, J. Chen, J. Li, Z. Xu, J. Z. Pan, and H. Chen. 2022. *Knowledge graph reasoning with logics and embeddings: survey and perspective*. CoRR abs/2202.07412 (2022).

• H. Yuan, H. Yu, S. Gui, and S. Ji. *Explainability in graph neural networks: A taxonomic survey*. IEEE Trans. Pattern Anal. Mach. Intell., 2022.

References

• G. Ras, N. Xie, M. v. Gerven, and D. Doran. *Explainable deep learning: a field guide for the uninitiated*. J. Artif. Intell. Res., vol. 73, pp. 329396, 2022.

• Z. C. Lipton. *The mythos of model interpretability*. Commun. ACM, vol. 61, no. 10, 2018.

• M. Du, N. Liu, and X. Hu. *Techniques for interpretable machine learning*. Commun. ACM, vol. 63, no. 1, 2020.

• D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. *Machine learning interpretability: a survey on methods and metrics*. Electronics, vol. 8, no. 8, 2019.

• R. Pradhan, A. Lahiri, S. Galhotra, and B. Salimi. *Explainable AI: foundations, applications, opportunities for data management research*. in ICDE, 2022.

• Z. Yang, N. Liu, X. B. Hu, and F. Jin. *Tutorial on deep learning interpretation: a data perspective*. in CIKM, 2022.

• C. Wang, X. Li, H. Han, S. Wang, L. Wang, C. C. Cao, and L. Chen. *Counterfactual explanations in explainable AI: a tutorial*. in KDD, 2021.

• A. Datta, M. Fredrikson, K. Leino, K. Lu, S. Sen, and Z. Wang. *Machine learning explainability and robustness: connected at the hip*. in KDD, 2021.

• K. Gade, S. C. Geyik, K. Kenthapadi, V. Mithal, and A. Taly. *Explainable AI in industry*. in KDD, 2019.

• C. Agarwal, E. Saxena, S. Krishna, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, and H. Lakkaraju. *OpenXAI: towards a transparent evaluation of model explanations*. CoRR, vol. abs/2206.11104, 2022.

• P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann. *Explainability methods for graph convolutional neural networks*. in CVPR, 2019.

• C. Agarwal, O. Queen, H. Lakkaraju, and M. Zitnik. *Evaluating explainability for graph neural networks*. Sci Data, vol. 10, no. 1, 2023.

•C. Agarwal, M. Zitnik, and H. Lakkaraju. *Probing GNNexplainers: a rigorous theoretical and empirical analysis of GNNexplanation methods*. In AISTATS, 2022.

• K. T. T. Shun, E. E. Limanta, and A. Khan. *An evaluation of backpropagation interpretability for graph classification with deep learning*. In IEEE BigData, 2020.

• B. S´anchez-Lengeling, J. Wei, B. Lee, E. Reif, P. Wang, W. Qian, K. McCloskey, L. Colwell, and A. Wiltschko. *Evaluating attribution for graph neural networks*. In NeurIPS, 2020.

• F. Baldassarre and H. Azizpour. *Explainability techniques for graph convolutional networks*. In ICML Workshop on Learning and Reasoning with Graph-Structured Representations, 2019.

References

• C. J. Cai, J. Jongejan, and J. Holbrook. *The effects of example-based explanations in a machine learning interface*. IUI 2019.

• H. Xuanyuan, P. Barbiero, D. Georgiev, L. C. Magister, and P. Liò. *Global concept-based interpretability for graph neural networks via neuron analysis*. AAAI 2023.

• F. Doshi-Velez and B. Kim. *Towards A rigorous science of interpretable machine learning*. CoRR abs/1702.08608, 2017.

• S. R. Hong, J. Hullman, E. Bertini. *Human Factors in Model Interpretability: industry practices, challenges, and needs*. Proc. ACM Hum. Comput. Interact. 4(CSCW): 068:1-068:26 (2020).

• H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. M. Wallach, and J. Wortman Vaughan. *Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning*. CHI 2020.

• I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, and F. Doshi-Velez. *Human evaluation of models built for interpretability*. HCOMP 2019.

• F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. M. Wallach. *Manipulating and measuring model interpretability*. CHI 2021.

• H. Lakkaraju, J. Adebayo, and S. Singh. *Explaining machine learning predictions: State-of-the-art, challenges, and opportunities.* NeurIPS 2020.

• L. Faber, A. K. Moghaddam, and R. Wattenhofer. *When comparing to ground truth is wrong: on evaluating GNN explanation methods*. KDD 2021.

• Y.-X. Wu, X. Wang, A. Zhang, X. Hu, F. Feng, X. He, and T.-S. Chua. *Deconfounding to explanation evaluation in graph neural networks*. CoRR abs/2201.08802 (2022).

• M. A. Prado-Romero, B. Prenkaj, G. Stilo, and F. Giannotti. *A survey on graph counterfactual explanations: definitions, methods, evaluation*. CoRR abs/2210.12089 (2022).

• J. Kakkad, J. Jannu, K. Sharma, C. C. Aggarwal, and S. Medya. *A survey on explainability of graph neural networks*. CoRR,vol. abs/2306.01958, 2023.

• A. Longa, S. Azzolin, G. Santin, G. Cencetti, P. Liò, B. Lepri, and A. Passerini. *Explaining the explainers in graph neural networks: a comparative study*. CoRR, vol. abs/2210.15304 , 2023.

• P. Li, Y. Yang, M. Pagnucco, and Y. Song. *Explainability in graph neural networks: an experimental survey*. CoRR, vol.abs/2203.09258, 2022.

References

- P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Li`o, and Y. Bengio. *Graph attention networks*. In ICLR 2018.
- E. Dai and S. Wang. *Towards self-explainable graph neural network*. in CIKM 2021.
- Z. Zhang, Q. Liu, H. Wang, C. Lu, and C. Lee. *Protgnn: Towards self-explaining graph neural networks*. in AAAI 2022.
- E. Dai and S. Wang. *Towards prototype-based self-explainable graph neural network*. arXiv preprint arXiv:2210.01974, 2022.
- Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. *GNNExplainer: generating explanations for graph neural networks*. NeurIPS 2019.
- H. Yuan, J. Tang, X. Hu, and S. Ji. *Xgnn: Towards model level explanations of graph neural networks*. In KDD. 2020.
- T. Funke, M. Khosla, and A. Anand. *Zorro: Valid, sparse, and stable explanations in graph neural networks. IEEE Transactions on Knowledge & Data Engineering. 35(8), 2023.*
- D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang. *Parameterized explainer for graph neural network.* NeurIPS 2020.
- H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji. *On explainability of graph neural networks via subgraph explorations*. ICML 2021.
- M. Vu and M. T. Thai. *Pgm-explainer: Probabilistic graphical model explanations for graph neural networks*. NeurIPS 2020.
- R. Schwarzenberg, M. H¨ubner, D. Harbecke, C. Alt, and L. Hennig. *Layerwise relevance visualization in convolutional text graph classifiers*. TextGraphs 2019.
- J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li, and Y. Zhang. *Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning*. WWW 2022.
- M. Khosla and L. Galárraga. *Explainable graph machine learning: techniques to explain black-box models on graphs* (Tutorial). ECML 2023.
- E. Dai and S. Wang. *Towards prototype-based self-explainable graph neural network*. CoRR abs/2210.01974 (2022).
- Z. Zhang, Q. Liu, H. Wang, C. Lu, and C. Lee. *ProtGNN: Towards self-explaining graph neural networks*. AAAI 2022.
- E. Dai and S. Wang. *Towards self-explainable graph neural network*. CIKM 2021.

References

• M. Gevrey, I. Dimopoulos, and S. Lek. *Review and comparison of methods to study the contribution of variables in artificial neural network models.* Ecological modelling 2003.

• J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller (2014). *Striving for simplicity: The all convolutional net.*

• B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba. *Learning deep features for discriminative localization.* In Proceedings of the IEEE conference on computer vision and pattern recognition 2016.

•R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra. *Grad-cam: Visual explanations from deep networks via gradient-based localization*. In Proceedings of the IEEE international conference on computer vision 2017.

• S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Müller, W. Samek. *On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation*. PloS one.

• J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, S. Sclaroff. *Top-down neural attention by excitation backprop*. International Journal of Computer Vision. 2018 Oct.

• D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, X. Zhang. *Parameterized explainer for graph neural network*. Advances in neural information processing systems. 2020.

• M. S. Schlichtkrull, N. De Cao, I. Titov. *Interpreting graph neural networks for nlp with differentiable edge masking*. ICLR 2021.

• M. Vu, M. T. Thai. *Pgm-explainer: Probabilistic graphical model explanations for graph neural networks*. Advances in neural information processing systems. 2020.

• J. Tan, S. Geng, Z. Fu, Y. Ge, S. Xu, Y. Li, Y. Zhang. *Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning*. InProceedings of the ACM Web Conference 2022.

References

• L. Veyrin-Forrer, A. Kamal, S. Duffner, M. Plantevit, and C. Robardet. *On GNN explainability with activation rules*. Data Min Knowl Disc (2022).

• G. Jaume, P. Pati, B. Bozorgtabar, A. Foncubierta, A. M. Anniciello, F. Feroce, T. Rau, J.-P. Thiran, M. Gabrani, and O. Goksel. *Quantifying Explainers of Graph Neural Networks in Computational Pathology*. CVPR 2021.

• T. Zhao, D. Luo, X. Zhang, S. Wang. *Towards Faithful and Consistent Explanations for Graph Neural Networks*. WSDM 2023.

• Y. Gao, T. S. Sun, R. Bhatt, D. Yu, S. R. Hong, and L. Zhao. *GNES: Learning to Explain Graph Neural Networks*. ICDM 2021.

• T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, and G. Montavon. *Higher-Order Explanations of Graph Neural Networks via Relevant Walks*. IEEE Trans. Pattern Anal. Mach. Intell. 44(11): 7581-7596 (2022).

• Y. Xie, S. Katariya, X. Tang, E. W. Huang, N. Rao, K. Subbian, and S. Ji. *Task-Agnostic Graph Explanations*. NeurIPS 2022.

• I. E. Olatunji, M. Rathee, T. Funke, and M. Khosla. *Private Graph Extraction via Feature Explanations*. Proc. Priv. Enhancing Technol. 2023(2): 59-78 (2023).

• X. Wang, Y. Wu, A. Zhang, F. Feng, X. He, and T.-S. Chua. *Reinforced Causal Explainer for Graph Neural Networks*. IEEE Trans. Pattern Anal. Mach. Intell. 45(2): 2297-2309 (2023).

• E. Dai, T. Zhao, H. Zhu, J. Xu, Z. Guo, H. Liu, J. Tang, and S. Wang. *A Comprehensive Survey on Trustworthy Graph Neural Networks: Privacy, Robustness, Fairness, and Explainability*. CoRR abs/2204.08570 (2022).

• A. Rossi, D. Firmani, P. Merialdo, and T. Teofili. *Explaining link prediction systems based on knowledge graph embeddings*. SIGMOD Conference 2022.

• A. Gogoglou, C. B. Bruss, and K. E. Hines. *On the interpretability and evaluation of graph representation learning*. CoRR abs/1910.03081 (2019).

• T. Lanciano, F. Bonchi, and A. Gionis. *Explainable classification of brain networks via contrast subgraphs*. KDD 2020.

• C. Abrate and F. Bonchi. *Counterfactual graphs for explainable classification of brain networks*. KDD 2021.

• A. Feng, C. You, S. Wang, and L. Tassiulas. *KerGNNs: interpretable graph neural networks with graph kernels*. AAAI 2022.

• 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. LoG 2022

• M. Bajaj, L. Chu, Z. Y. Xue, J. Pei, L. Wang, P. Cho-Ho Lam, Y. Zhang. *Robust counterfactual explanations on graph neural networks*. NeurIPS 2021.

• B. Sánchez-Lengeling, J. N. Wei, B. K. Lee, E. Reif, P. Wang, W. W. Qian, K. McCloskey, L. J. Colwell, A. B. Wiltschko. *Evaluating attribution for graph neural networks*. NeurIPS 2020

AALBORG UNIVERSITY

IEEE | Cosponsers ACM/ASA/CCF DSAA 2023

AALBORG UNIVERSITY

Reach us at

arijitk@cs.aau.dk
ebmo@cs.aau.dk

https://twitter.com/rijitk