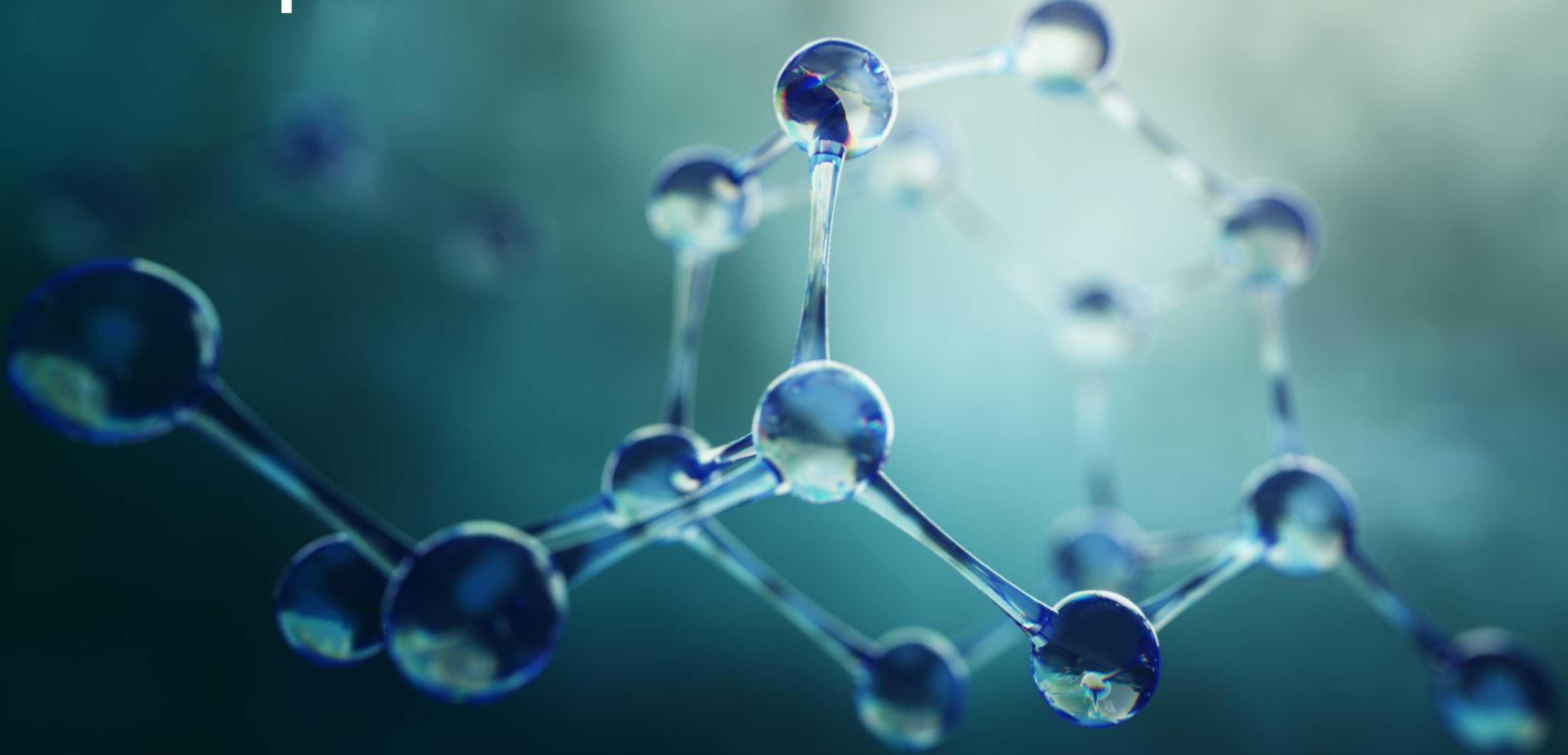
Evaluating Neighbor Explainability for Graph Neural Networks



Oscar Llorente Gonzalez | Rana Fawzy

Ericsson Cognitive Labs · Geometric Artificial Intelligence (GAI) Lab

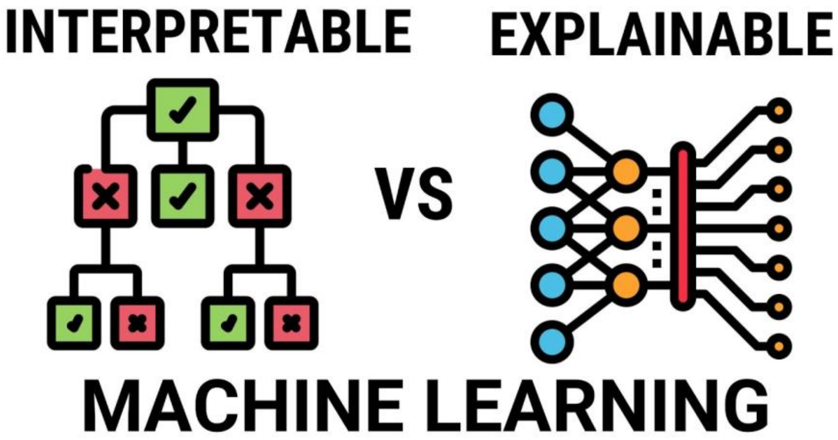
July 2024

Agenda

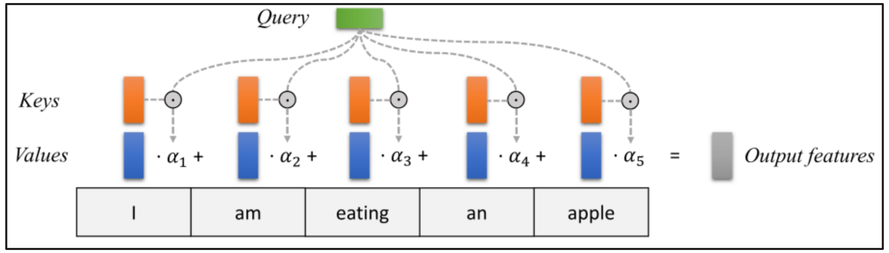
- Introduction to Explainable AI (XAI)
- Introduction to Neighbor XAI
- Loyalty and Inverse Loyalty
- Loyalty and Inverse Loyalty Probabilities
- XAI in self-loops

Introduction to Explainable AI (XAI)

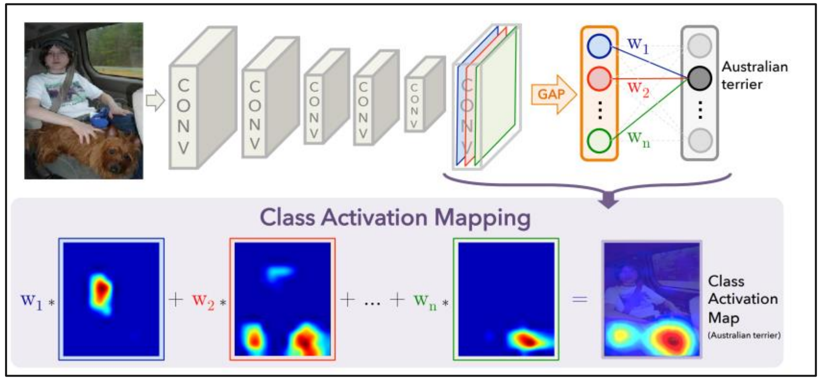
What is XAI?



INTERPRETABLE
VS
EXPLAINABLE
MACHINE LEARNING

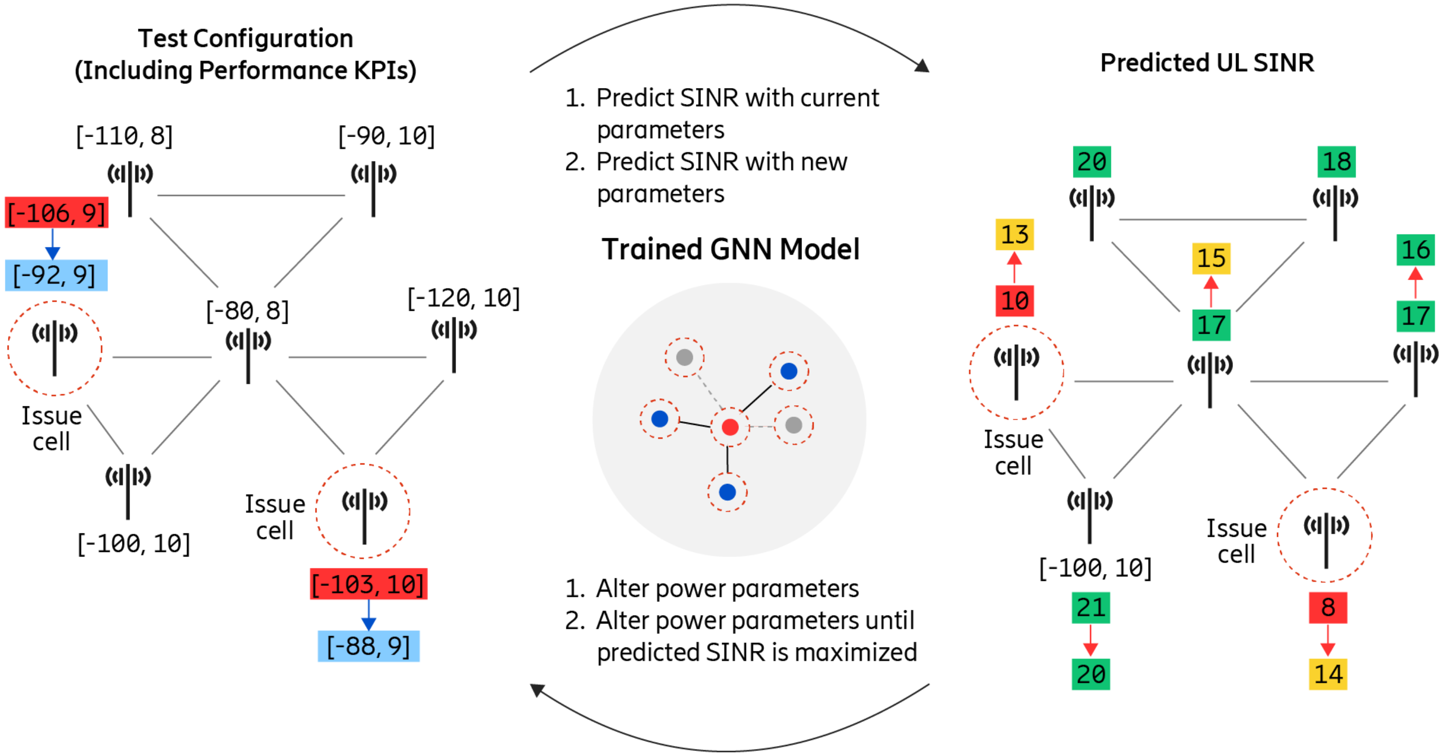


Query
Keys
Values
· α_1 +
· α_2 +
· α_3 +
· α_4 +
· α_5
=
Output features
I
am
eating
an
apple



CONV
GAP
w_1
w_2
w_n
Australian terrier
Class Activation Mapping
w_1 *
+ w_2 *
+ ... + w_n *
=
Class Activation Map
(Australian terrier)

Graph Neural Networks in telecom



GNNs: Capabilities and Interpretability challenges

Rich representation of relational data

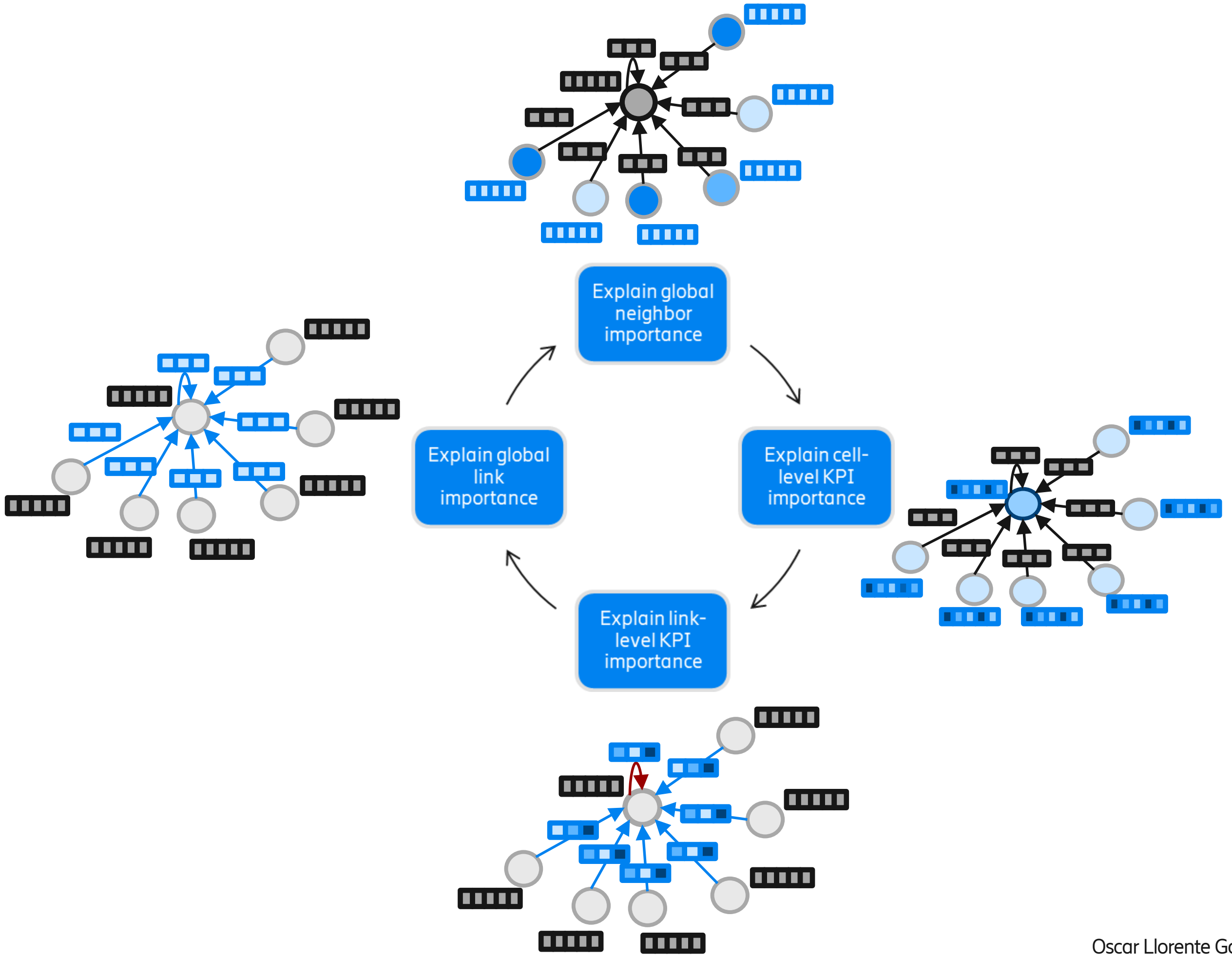
Superior performance in node and graph-level tasks

Interpretability challenges and need for specialized explanation methods

Objectives of XAI in GNNs

Innovative Explanatory Framework for GNNs

Explainable AI applied to graph level requires to explain different variables and relation levels



Explainability techniques for GNNs

- <u>Traditional techniques:</u>
  - SHAP
  - LIME
- <u>Gradient-based techniques:</u>
  - Saliency map
  - SmoothGrad
  - Integrated Gradients
- <u>GNN-specific techniques:</u>
  - GNNExplainer
  - PGExplainer

Introduction to Neighbor XAI

Paper Introduction

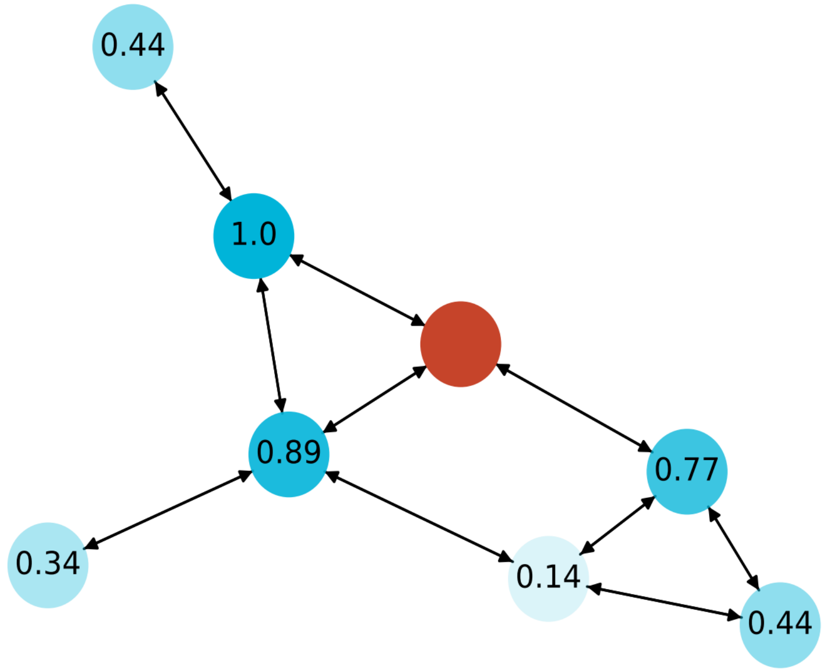
Evaluating Neighbor Explainability for Graph Neural Networks

Oscar Llorente[1], Rana Fawzy[1], Jared Keown[2], Michal Horemuz[2], Péter Vaderna[3], Sándor Laki[4], Roland Kotroczó[4], Rita Csoma[4], and János Márk Szalai-Gindl

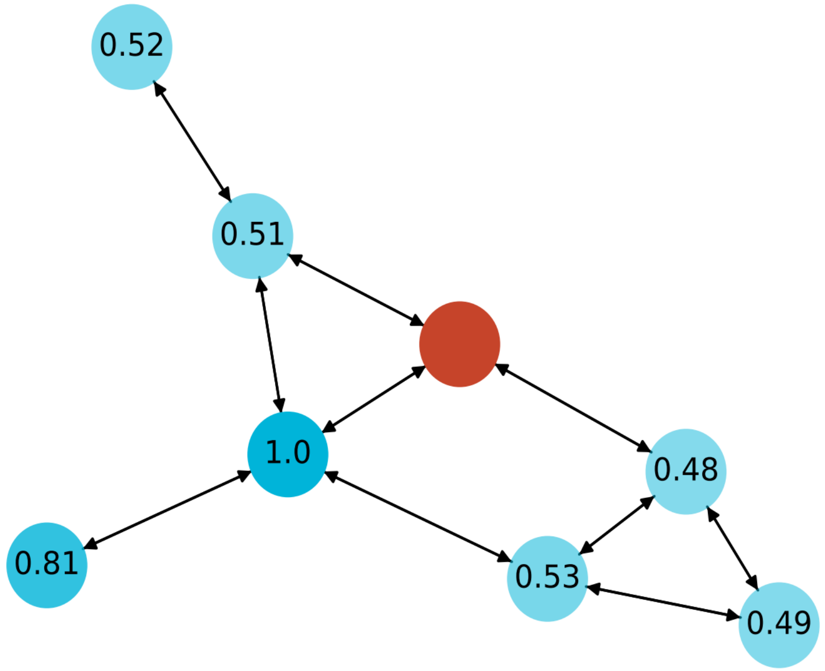
[1] Ericsson Cognitive Network Solutions, Madrid-Cairo, Spain-Egypt
{oscar.llorente.gonzalez,rana.fawzy}@ericsson.com
[2] Ericsson Global Artificial Intelligence Accelerator, Stockholm, Sweden
{jared.keown,michal.horemuz}@ericsson.com
[3] Ericsson Research, Stockholm, Sweden
Peter.Vaderna@ericsson.com
[4] Etvös Loránd University, Budapest, Hungary
{lakis,kotroczo.roland,gq9215,szalaigindl}@inf.elte.hu

Abstract. Graph Neural Networks (GNNs) have rapidly emerged as powerful tools for modeling complex data structures, particularly in the context of telecommunications, chemistry and social networking. Explainability in GNNs holds essential significance as it empowers stakeholders to gain insights into the inner workings of these complex models, fostering trust and transparency in decision-making processes. In this publication, we address the problem of determining how important is each neighbor for the GNN when classifying a node and how to measure the performance for this specific task. To do this, various known explainability methods are reformulated to calculate the neighbor importance and four new metrics, that aid in determining an explainability method's reliability, are presented. Our results show that there is almost no difference between the explanations provided by gradient-based techniques in the GNN domain, in contrast to other areas of deep learning where this is an active area of research. This means that efforts in this direction may not produce such promising results for GNNs. In addition, many explainability techniques failed to identify important neighbors when GNNs without self-loops are used[5].

Explainability Objective – Identify Neighbors



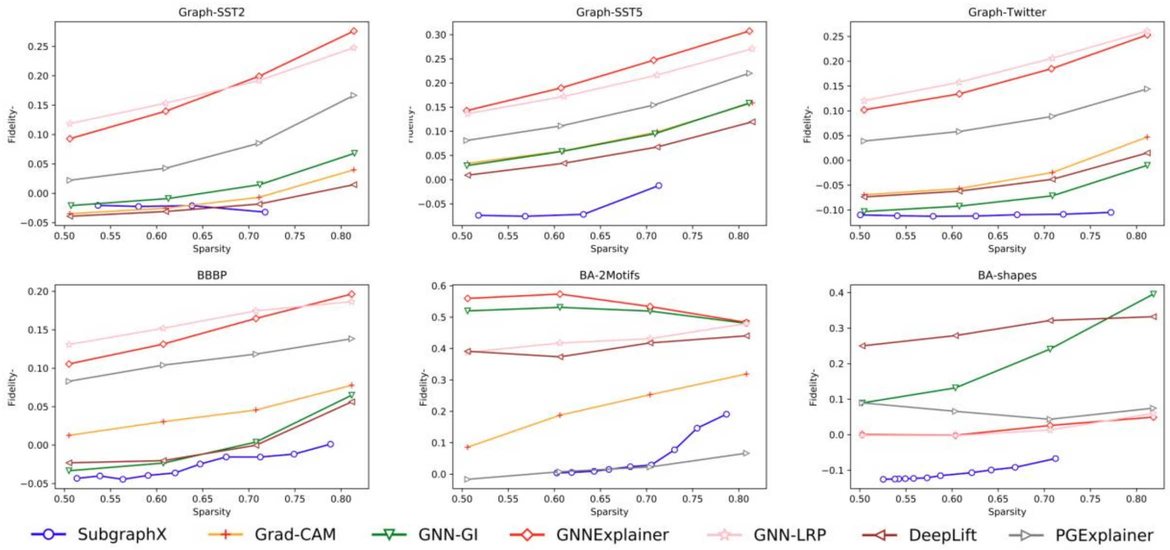
(a) Saliency Map



(b) GNNExplainer

Loyalty and Inverse Loyalty

Existing Metrics



Phenomenon

$$fid_+ = \frac{1}{N}\sum_{i=1}^{N}\left|\mathbb{1}(\hat{y}_i = y_i) - \mathbb{1}(\hat{y}_i^{G_{C\setminus S}} = y_i)\right|$$

$$fid_- = \frac{1}{N}\sum_{i=1}^{N}\left|\mathbb{1}(\hat{y}_i = y_i) - \mathbb{1}(\hat{y}_i^{G_S} = y_i)\right|$$

Model

$$fid_+ = 1 - \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{y}_i^{G_{C\setminus S}} = \hat{y}_i)$$

$$fid_- = 1 - \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}(\hat{y}_i^{G_S} = \hat{y}_i)$$

Loyalty

$$l_k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\hat{y}_{oi} = \hat{y}_{ki}),$$

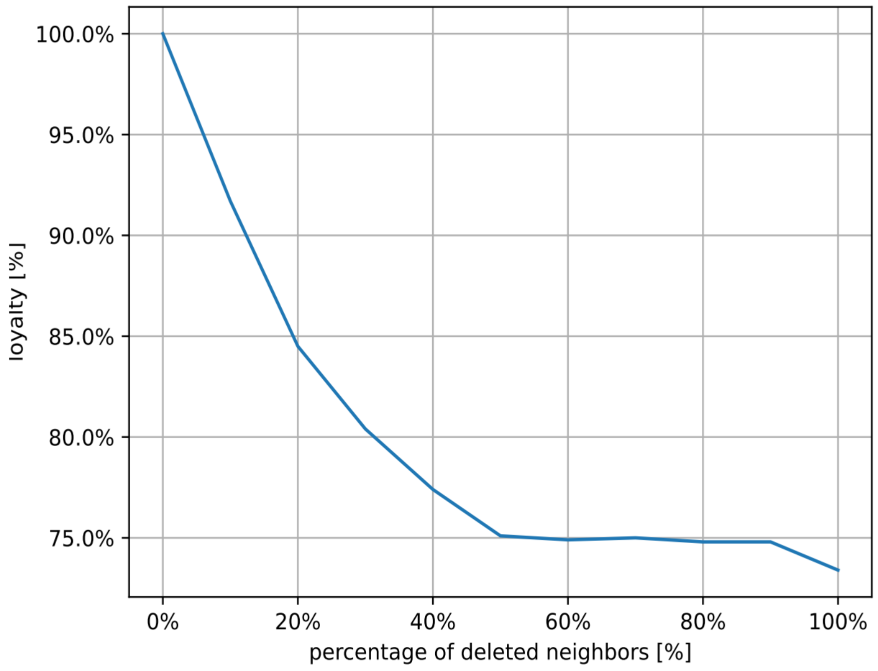
- Neighbors with non-zero importance values will be sorted in decreasing order.
- The better the technique, the greater the drop in classification accuracy at the beginning and the smoother at the end.
- Check that Explainability methods identity correctly most important neighbors.

Inverse Loyalty

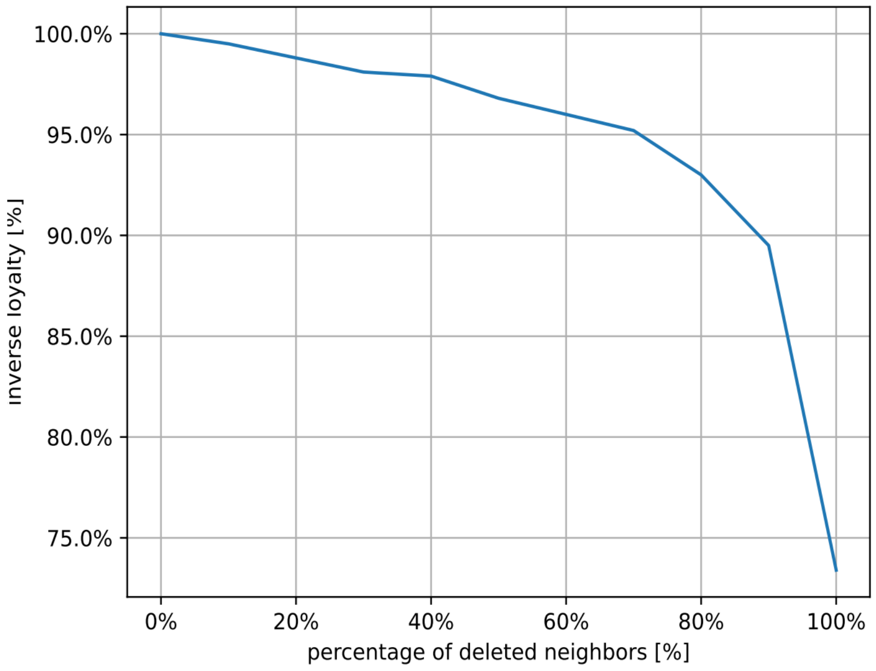
$$l_k = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\hat{y}_{oi} = \hat{y}_{ki}),$$

- Neighbors with non-zero importance values will be sorted in ascending order.
- The better the technique, the smoother the drop in classification accuracy at the beginning and the greater at the end.
- Check that Explainability methods identity correctly least important neighbors.

Loyalty and Inverse Loyalty Results



(a) Loyalty



(b) Inverse Loyalty

AUC Loyalty and Inverse Loyalty

Table 1: AUC Loyalty (L) and Inverse (I) Loyalty

Self-Loops	Method	Cora				CiteSeer				PubMed			
		GCN		GAT		GCN		GAT		GCN		GAT	
		L	I	L	I	L	I	L	I	L	I	L	I
With	Saliency Map	0.80	0.95	0.67	**0.94**	0.86	0.86	0.81	0.94	0.83	0.96	0.84	**0.97**
	Smoothgrad	0.80	0.95	0.70	0.89	0.87	0.96	0.81	0.93	0.83	0.96	0.84	0.96
	Deconvnet	0.79	0.95	0.67	**0.94**	0.86	0.86	0.81	0.94	0.83	0.96	0.84	**0.97**
	Guided Backprop	0.79	0.95	0.67	**0.94**	0.86	0.86	0.81	0.94	0.83	0.96	0.84	**0.97**
	GNNExplainer	**0.74**	**0.97**	**0.64**	**0.94**	**0.83**	**0.97**	**0.78**	**0.96**	**0.79**	**0.97**	**0.81**	**0.97**
	PGExplainer	0.88	0.88	0.75	0.86	0.91	0.90	0.85	0.89	0.90	0.86	0.88	0.91
Without	Saliency Map	0.81	0.90	0.58	0.77	0.89	0.93	0.78	0.70	0.89	0.9	0.58	0.81
	Smoothgrad	0.81	0.90	0.57	0.78	0.89	0.92	0.77	0.71	0.89	0.93	**0.54**	**0.85**
	Deconvnet	0.81	0.90	0.58	0.79	0.90	0.92	0.79	0.69	0.89	0.93	0.58	0.80
	Guided Backprop	0.81	0.90	0.58	0.79	0.90	0.92	0.79	0.69	0.89	0.93	0.58	0.80
	GNNExplainer	**0.74**	**0.94**	**0.56**	**0.80**	0.86	**0.96**	**0.74**	**0.76**	0.77	**0.97**	0.7	0.71
	PGExplainer	0.76	0.73	0.66	0.73	**0.80**	0.73	0.75	0.75	**0.74**	0.72	0.67	0.74

Loyalty and Inverse Loyalty Probabilities

Loyalty Probabilities

$$l_k = \frac{1}{N}\sum_{i=1}^{N} |(P(\hat{y}_i = \hat{y}_{oi} \mid \mathbf{G} = \mathbf{G_{ki}}) - (P(\hat{y}_i = \hat{y}_{oi} \mid \mathbf{G} = \mathbf{G_o})|$$

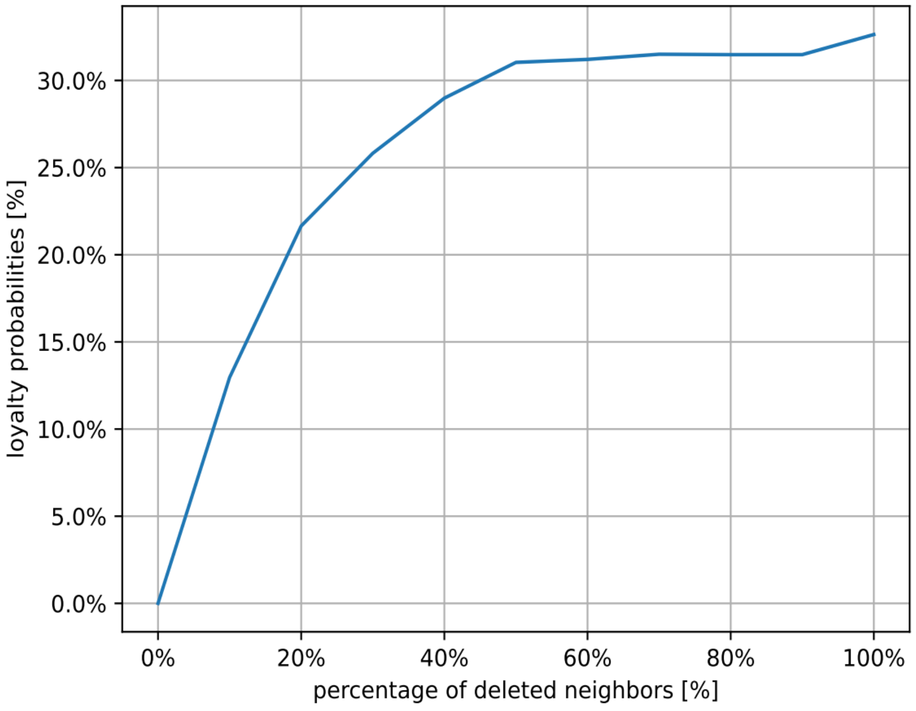
- Neighbors with non-zero importance values will be sorted in descending order.
- The better the technique, the sharper the increase in probabilities difference at the beginning and the smoother at the end.
- Check that Explainability methods identity correctly most important neighbors when neighbors are not highly-important.

Inverse Loyalty Probabilities

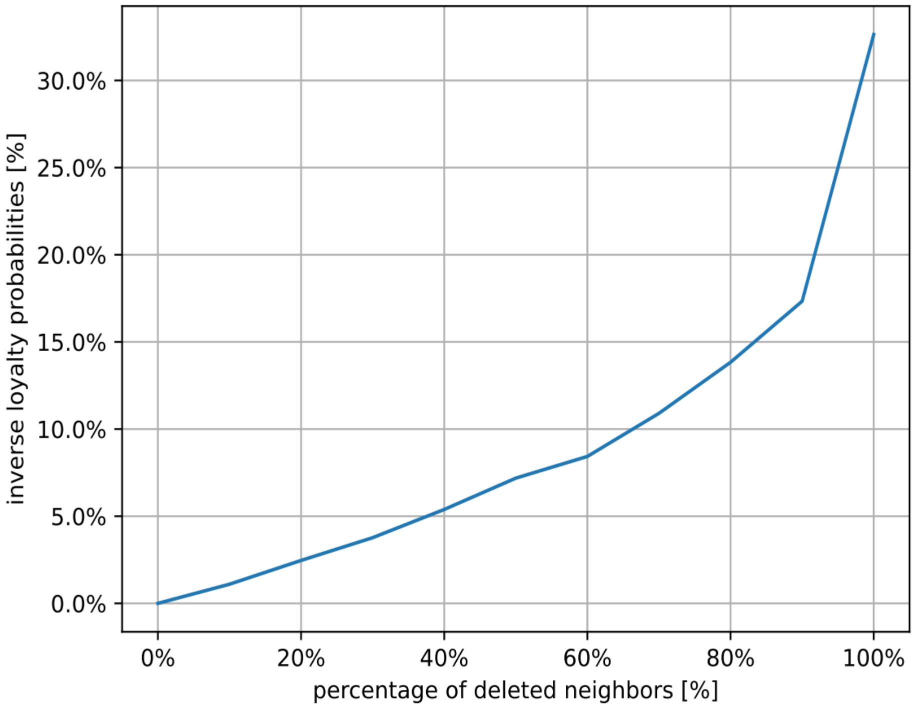
$$l_k = \frac{1}{N}\sum_{i=1}^{N} |(P(\hat{y}_i = \hat{y}_{oi} \mid \mathbf{G} = \mathbf{G_{ki}}) - (P(\hat{y}_i = \hat{y}_{oi} \mid \mathbf{G} = \mathbf{G_o})|$$

- Neighbors with non-zero importance values will be sorted in ascending order.
- The better the technique, the smoother the increase in probabilities difference at the beginning and the sharper at the end.
- Check that Explainability methods identity correctly least important neighbors when neighbors are not highly-important.

Loyalty and Inverse Loyalty Probabilities Results



(a) Loyalty Probabilities



(b) Inverse Loyalty Probabilities

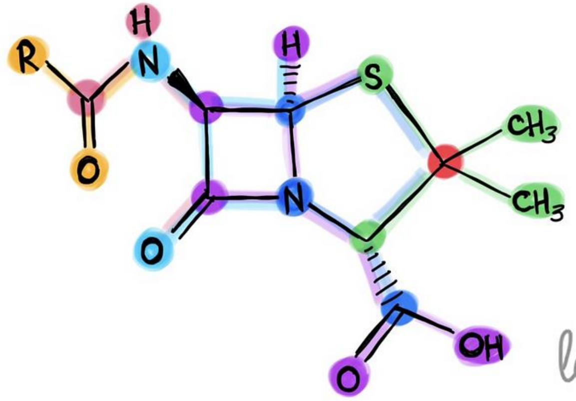
AUC Loyalty and Inverse Loyalty Probabilities

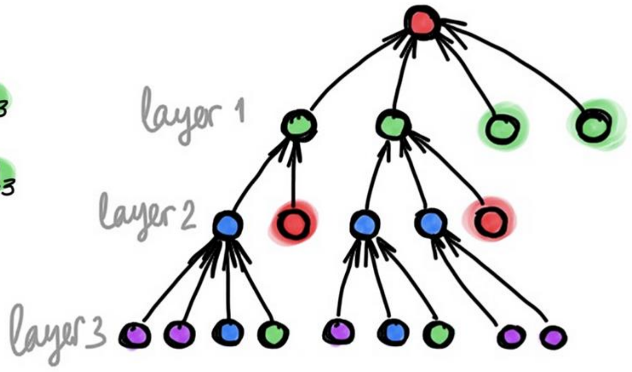
Table 2: AUC Loyalty (L) and Inverse (I) Loyalty Probabilities

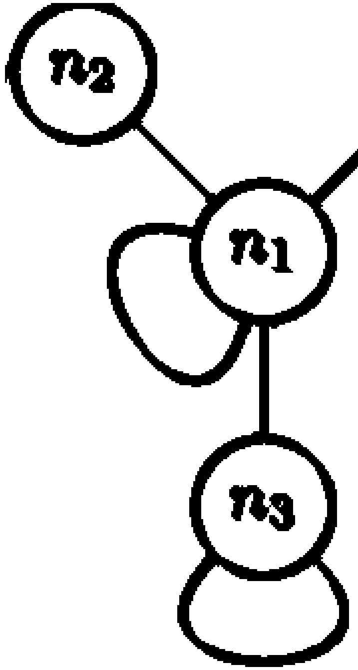
Self-Loops	Method	Cora				CiteSeer				PubMed			
		GCN		GAT		GCN		GAT		GCN		GAT	
		L	I	L	I	L	I	L	I	L	I	L	I
With	Saliency Map	0.26	**0.09**	0.40	0.08	0.17	**0.07**	0.24	**0.07**	**0.22**	**0.06**	0.20	**0.04**
	SmoothGrad	0.26	**0.09**	0.37	0.14	0.17	**0.07**	0.23	0.08	**0.22**	**0.06**	0.19	0.05
	Deconvnet	0.26	**0.09**	0.40	**0.07**	0.17	**0.07**	0.24	**0.07**	**0.22**	**0.06**	0.20	**0.04**
	Guided Backprop	0.26	**0.09**	0.40	**0.07**	0.17	**0.07**	0.24	**0.07**	**0.22**	**0.06**	0.20	**0.04**
	GNNExplainer	**0.28**	0.15	**0.41**	0.10	**0.18**	0.11	**0.25**	0.08	**0.22**	0.10	**0.21**	0.06
	PGExplainer	0.18	0.17	0.31	0.19	0.12	0.13	0.19	0.13	0.13	0.17	0.15	0.11
Without	Saliency Map	0.24	**0.14**	0.46	0.24	0.13	**0.09**	0.24	0.28	0.15	**0.10**	0.40	0.18
	SmoothGrad	0.24	**0.14**	0.45	0.26	0.13	**0.09**	0.24	0.27	0.15	**0.10**	**0.43**	**0.14**
	Deconvnet	0.24	**0.14**	0.46	**0.23**	0.13	**0.09**	0.24	0.28	0.15	**0.10**	0.39	0.19
	Guided Backprop	0.24	**0.14**	0.46	**0.23**	0.13	**0.09**	0.24	0.28	0.15	**0.10**	0.39	0.19
	GNNExplainer	**0.28**	0.17	**0.47**	**0.23**	**0.14**	0.11	**0.26**	**0.24**	0.23	0.13	0.32	0.26
	PGExplainer	0.26	0.30	0.39	0.30	0.20	0.26	**0.26**	**0.24**	**0.24**	0.28	0.34	0.23

XAI in self-loops

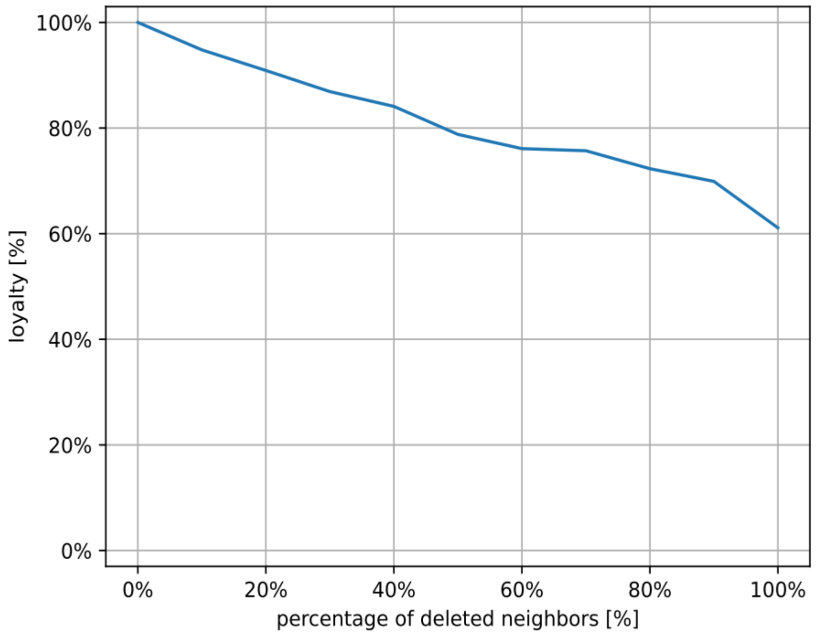
Self-loops



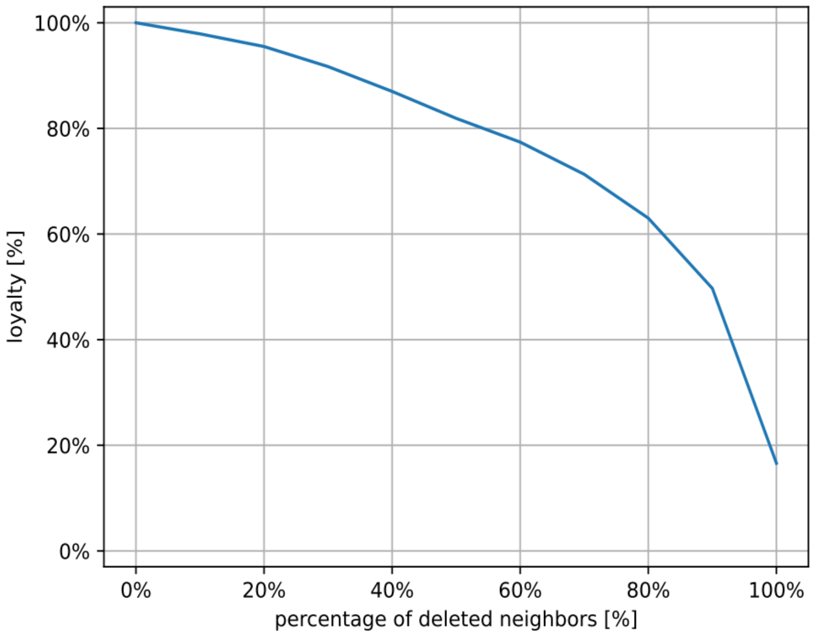




Loyalty without self-loops



(a) Saliency Map



(b) PGExplainer

Accuracy without self-loops

Method	Cora		CiteSeer		PubMed	
	GCN	GAT	GCN	GAT	GCN	GAT
Saliency Map	0.61	0.22	0.75	0.31	0.76	0.29
Smoothgrad	0.61	0.22	0.75	0.31	0.76	0.29
Deconvnet	0.61	0.24	0.75	0.31	0.76	0.29
Guided Backprop	0.61	0.24	0.75	0.31	0.76	0.29
GNNExplainer	0.61	0.22	0.75	0.32	0.76	0.29
PGExplainer	0.17	0.19	0.20	0.20	0.43	0.29
Without Neighbors	0.17	0.19	0.20	0.20	0.43	0.29