How Faithful are Self-Explainable GNNs?[1]

reporter: Jiale Liu

paper author:Marc Christiansen, Lea Villadsen, Zhiqiang Zhong, Stefano Teso, Davide Mottin

arXiv:2308.15096

May 25, 2024

Catalogs

Dataset

Table 1: Overview of data sets used in our experiments. Numbers are averaged.

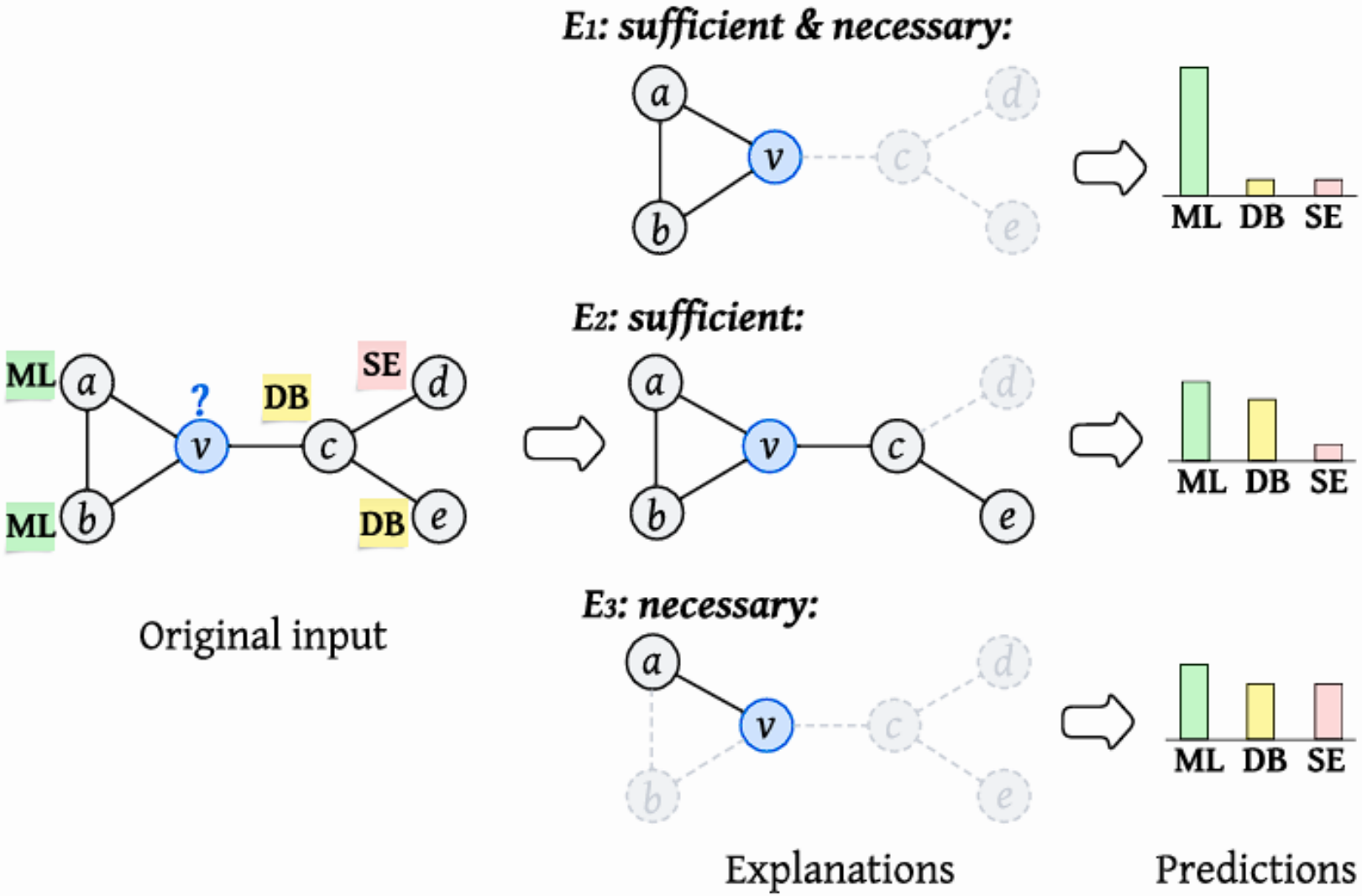
DATASET	# GRAPHS	AVG. # NODES	AVG. # EDGES	# FEATURES	# CLASSES
MUTAG	188	17.9	39.6	7	2
BBBP	2039	24.0	51.9	9	2
Ba2Motif	1000	30.0	57.8	2	2
BaMS	1000	40.0	87.4	2	2

Catalogs

Unfaithfulness

Unfaithfulness measures how strongly a prediction depends on *irrelevant* nodes, edges, and features:

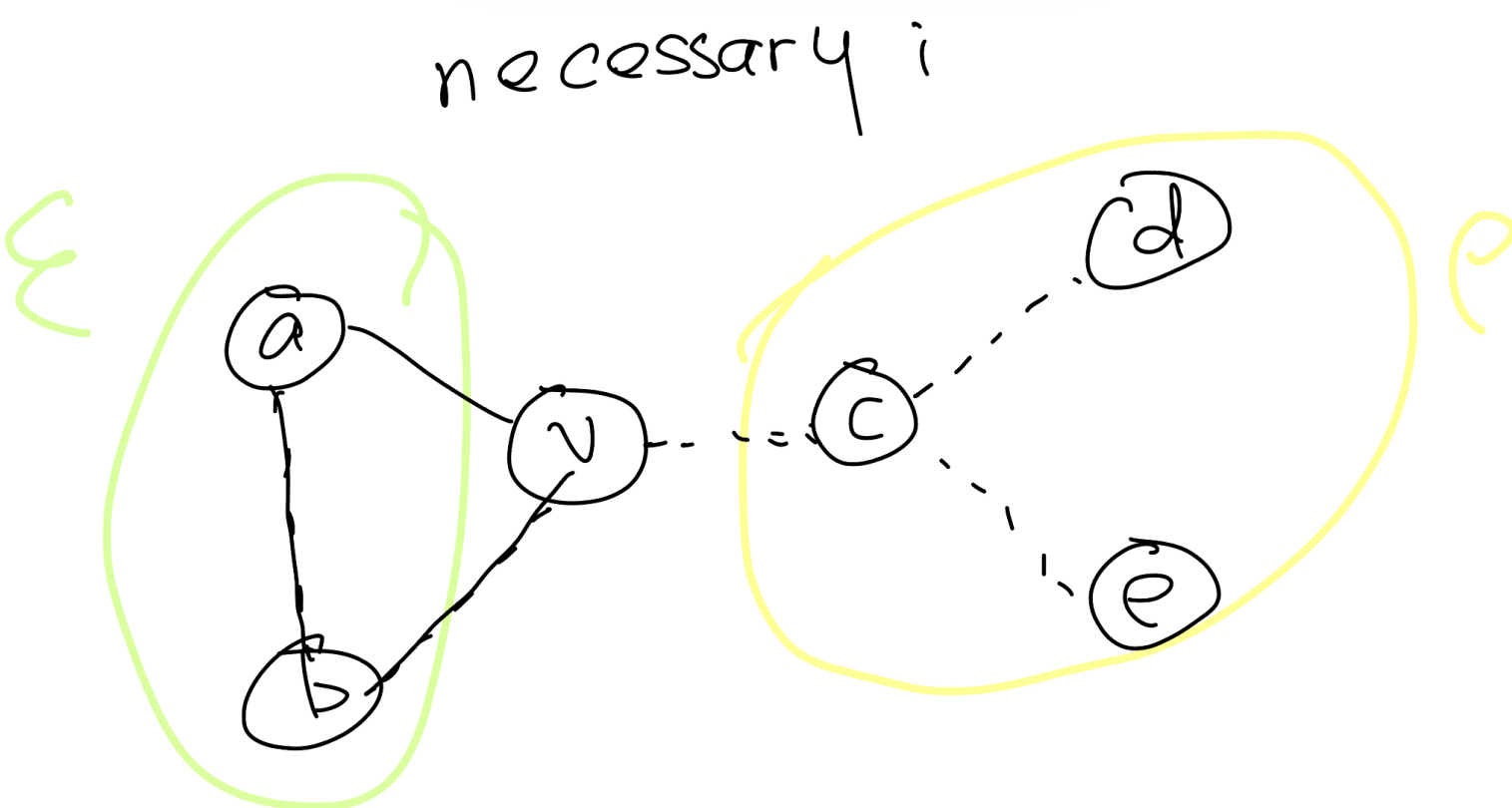
$$\text{Unf} = 1 - \exp\Big(- \text{KL}\Big(p_\theta(Y \mid G) \parallel p_\theta(Y \mid \mathcal{E}) \Big) \Big)$$

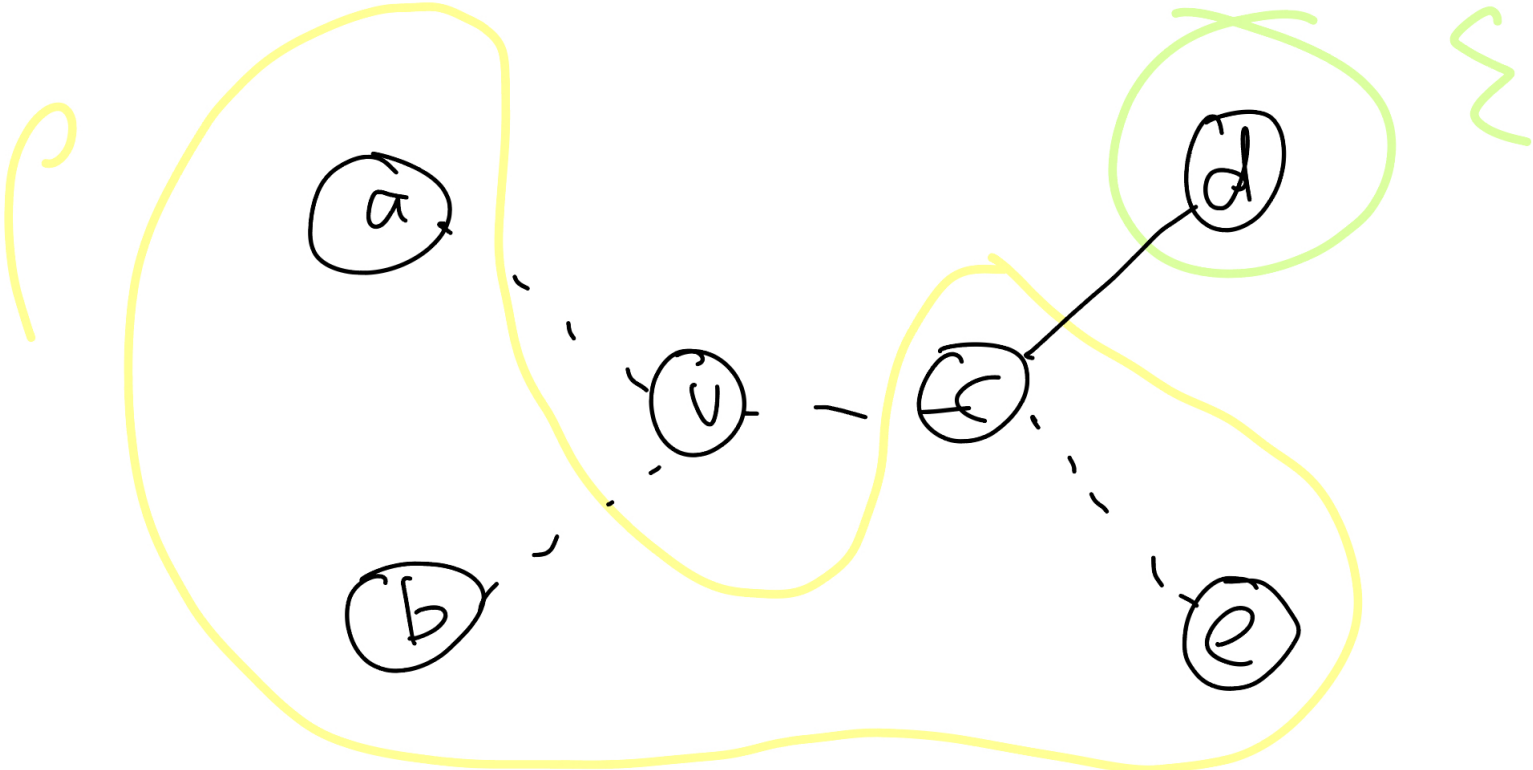

E1: sufficient & necessary:
E2: sufficient:
E3: necessary:
ML
DB
SE
?
Original input
Explanations
Predictions

Fidelity

Fidelity includes two metrics assessing to what degree explanations are *necessary* (Fid^+) and *sufficient* (Fid^-).

$$\text{Fid}^+ = \left| \mathbb{1}\left\{\hat{y} = y\right\} - \mathbb{1}\left\{\hat{y}^{\mathcal{C}} = y\right\} \right|, \quad \text{Fid}^- = \left| \mathbb{1}\left\{\hat{y} = y\right\} - \mathbb{1}\left\{\hat{y}^{\mathcal{E}} = y\right\} \right|$$


necessary i
a
v
b
c
d
e


a
d
v
c
b
e

Catalogs

Results

Table 2: Results for all models and data sets. We report average and std. dev. of classification accuracy, faithfulness of the model explanations $\mathcal{E}$, and faithfulness of random subgraphs $\mathcal{R}$. **Bold** entries indicate best result, **red** ones that the random subgraph is at least as faithful as the model's explanation, **orange** ones that it is at most 0.10 worse, and **green** the rest.

Dataset	Model	Acc (↑)	Unf($\mathcal{E}$)(↓)	Unf($\mathcal{R}$)(↓)	Fid$^-$($\mathcal{E}$)(↓)	Fid$^-$($\mathcal{R}$)(↓)	Fid$^+$($\mathcal{E}$)(↑)	Fid$^+$($\mathcal{R}$)(↑)
BBBP	GISST	**0.87 ± 0.01**	0.22 ± 0.15	0.15 ± 0.02	0.29 ± 0.25	0.21 ± 0.04	0.16 ± 0.01	0.20 ± 0.03
	PIGNN+P	0.86 ± 0.01	**0.07 ± 0.02**	0.16 ± 0.03	**0.07 ± 0.03**	0.20 ± 0.03	**0.33 ± 0.06**	0.22 ± 0.03
	PIGNN+T	0.86 ± 0.01	0.10 ± 0.01	0.15 ± 0.01	0.14 ± 0.02	0.22 ± 0.02	0.30 ± 0.05	0.27 ± 0.03
	ProtGNN	0.85 ± 0.01	0.15 ± 0.08	0.27 ± 0.05	0.14 ± 0.02	0.23 ± 0.05	0.21 ± 0.06	0.20 ± 0.05
MUTAG	GISST	**0.86 ± 0.04**	0.11 ± 0.12	0.13 ± 0.06	0.31 ± 0.17	0.41 ± 0.16	**0.50 ± 0.19**	0.38 ± 0.13
	PIGNN+P	0.82 ± 0.02	0.12 ± 0.08	0.18 ± 0.08	0.26 ± 0.26	0.59 ± 0.08	0.42 ± 0.17	0.52 ± 0.22
	PIGNN+T	0.84 ± 0.07	0.14 ± 0.13	0.08 ± 0.07	0.40 ± 0.30	0.51 ± 0.12	0.37 ± 0.05	0.40 ± 0.15
	ProtGNN	0.82 ± 0.02	**0.00 ± 0.00**	0.00 ± 0.00	**0.10 ± 0.07**	0.18 ± 0.04	0.34 ± 0.05	0.20 ± 0.05
Ba2Motif	GISST	0.98 ± 0.02	0.13 ± 0.16	0.32 ± 0.09	**0.14 ± 0.16**	0.32 ± 0.08	0.51 ± 0.02	0.52 ± 0.02
	PIGNN+P	**1.00 ± 0.00**	0.50 ± 0.04	0.42 ± 0.01	0.59 ± 0.06	0.50 ± 0.00	0.01 ± 0.01	0.50 ± 0.00
	PIGNN+T	0.98 ± 0.01	**0.03** ± 0.02	0.08 ± 0.05	0.21 ± 0.06	0.57 ± 0.06	**0.58 ± 0.14**	0.56 ± 0.08
	ProtGNN	0.51 ± 0.11	–	–	–	–	–	–
BaMS	GISST	0.79 ± 0.13	**0.11** ± 0.07	0.13 ± 0.06	**0.20 ± 0.13**	0.27 ± 0.12	0.42 ± 0.20	0.39 ± 0.19
	PIGNN+P	**0.96 ± 0.00**	0.32 ± 0.00	0.32 ± 0.00	0.54 ± 0.00	0.54 ± 0.00	0.54 ± 0.00	0.54 ± 0.00
	PIGNN+T	0.92 ± 0.00	0.13 ± 0.01	0.30 ± 0.00	0.46 ± 0.01	0.56 ± 0.00	**0.56 ± 0.00**	0.56 ± 0.00
	ProtGNN	0.64 ± 0.10	–	–	–	–	–	–

Q1: Faithfulness of self-explainable GNNs is not perfect and varies across tasks.

All models tend to perform better at Unf and Fid^-(best results are 0 and 0.07, respectively, the lower the better) than at Fid^+(best result is 0.58, the higher the better), indicating they are better suited at generating *sufficient*, rather than *necessary*, explanations. Moreover, explanation faithfulness varies widely across learning tasks.

Q2: Absolute faithfulness measures can be misleading. It is true, however, that faithfulness is excellent for some combinations of models and tasks. This is especially the case for ProtGNN on MUTAG and for PIGNNT on BA2MOTIF, which both attain near-perfect Unf and Fid^-.

Table 2: Results for all models and data sets. We report average and std. dev. of classification accuracy, faithfulness of the model explanations $\mathcal{E}$, and faithfulness of random subgraphs $\mathcal{R}$. **Bold** entries indicate best result, **red** ones that the random subgraph is at least as faithful as the model's explanation, **orange** ones that it is at most 0.10 worse, and **green** the rest.

DATASET	MODEL	ACC (↑)	Unf($\mathcal{E}$)(↓)	Unf($\mathcal{R}$)(↓)	Fid$^-$($\mathcal{E}$)(↓)	Fid$^-$($\mathcal{R}$)(↓)	Fid$^+$($\mathcal{E}$)(↑)	Fid$^+$($\mathcal{R}$)(↑)
BBBP	GISST	**0.87 ± 0.01**	0.22 ± 0.15	0.15 ± 0.02	0.29 ± 0.25	0.21 ± 0.04	0.16 ± 0.01	0.20 ± 0.03
	PIGNN+P	0.86 ± 0.01	**0.07 ± 0.02**	0.16 ± 0.03	**0.07 ± 0.03**	0.20 ± 0.03	**0.33 ± 0.06**	0.22 ± 0.03
	PIGNN+T	0.86 ± 0.01	0.10 ± 0.01	0.15 ± 0.01	0.14 ± 0.02	0.22 ± 0.02	0.30 ± 0.05	0.27 ± 0.03
	ProtGNN	0.85 ± 0.01	0.15 ± 0.08	0.27 ± 0.05	0.14 ± 0.02	0.23 ± 0.05	0.21 ± 0.06	0.20 ± 0.05
MUTAG	GISST	**0.86 ± 0.04**	0.11 ± 0.12	0.13 ± 0.06	0.31 ± 0.17	0.41 ± 0.16	**0.50 ± 0.19**	0.38 ± 0.13
	PIGNN+P	0.82 ± 0.02	0.12 ± 0.08	0.18 ± 0.08	0.26 ± 0.26	0.59 ± 0.08	0.42 ± 0.17	0.52 ± 0.22
	PIGNN+T	0.84 ± 0.07	0.14 ± 0.13	0.08 ± 0.07	0.40 ± 0.30	0.51 ± 0.12	0.37 ± 0.05	0.40 ± 0.15
	ProtGNN	0.82 ± 0.02	**0.00 ± 0.00**	0.00 ± 0.00	**0.10 ± 0.07**	0.18 ± 0.04	0.34 ± 0.05	0.20 ± 0.05
Ba2Motif	GISST	0.98 ± 0.02	0.13 ± 0.16	0.32 ± 0.09	**0.14 ± 0.16**	0.32 ± 0.08	0.51 ± 0.02	0.52 ± 0.02
	PIGNN+P	**1.00 ± 0.00**	0.50 ± 0.04	0.42 ± 0.01	0.59 ± 0.06	0.50 ± 0.00	0.01 ± 0.01	0.50 ± 0.00
	PIGNN+T	0.98 ± 0.01	**0.03** ± 0.02	0.08 ± 0.05	0.21 ± 0.06	0.57 ± 0.06	**0.58 ± 0.14**	0.56 ± 0.08
	ProtGNN	0.51 ± 0.11	–	–	–	–	–	–
BaMS	GISST	0.79 ± 0.13	**0.11** ± 0.07	0.13 ± 0.06	**0.20 ± 0.13**	0.27 ± 0.12	0.42 ± 0.20	0.39 ± 0.19
	PIGNN+P	**0.96 ± 0.00**	0.32 ± 0.00	0.32 ± 0.00	0.54 ± 0.00	0.54 ± 0.00	0.54 ± 0.00	0.54 ± 0.00
	PIGNN+T	0.92 ± 0.00	0.13 ± 0.01	0.30 ± 0.00	0.46 ± 0.01	0.56 ± 0.00	**0.56 ± 0.00**	0.56 ± 0.00
	ProtGNN	0.64 ± 0.10	–	–	–	–	–	–

Catalogs

Our initial results indicate that, at least when it comes to graph classification, it is *hard to assert that self-explainable GNNs are faithful by design*. Specifically, faithfulness of these models seems to be very **data set dependent**. This becomes even more clear when we compare the faithfulness of their explanations to that of a completely uninformed baseline.

One possible motivation behind these results is that the explanations offered by self-explainable models are faithful **at the concept level**.

Catalogs

[1] Marc Christiansen et al. *How Faithful are Self-Explainable GNNs?* 2023. arXiv: 2308.15096 [cs.LG].