V-InFoR: A Robust Graph Neural Networks Explainer for Structurally Corrupted Graphs

Senzhang Wang*, Jun Yin*, Chaozhuo Li, Xing Xie, Jianxin Wang

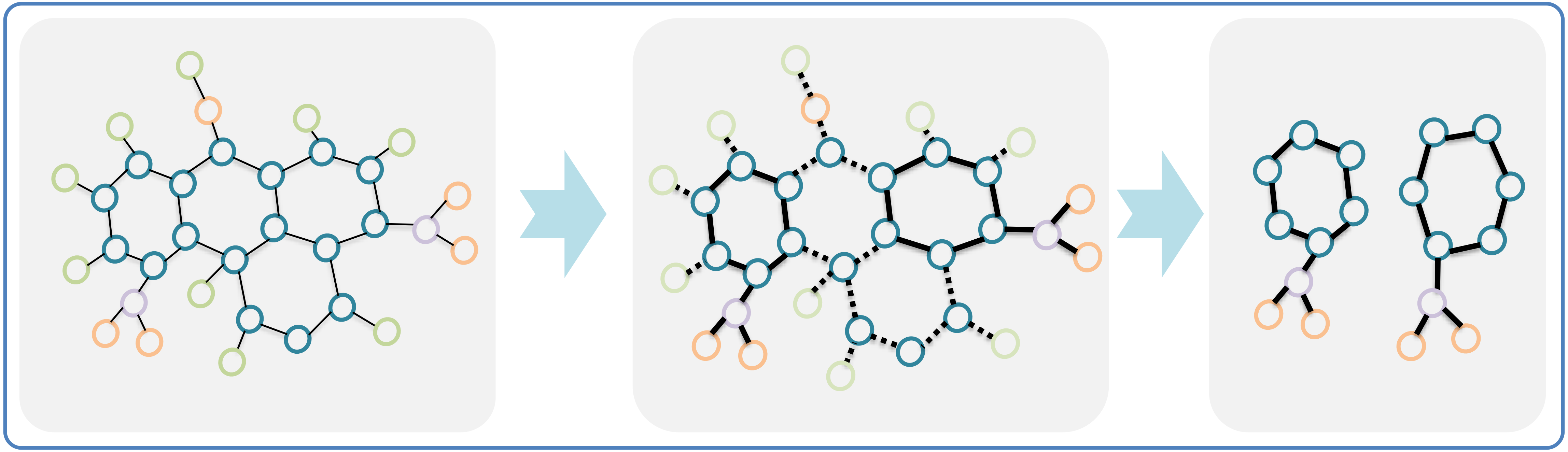
Central South University, Changsha, China

NeurIPS 2023

CONTENT

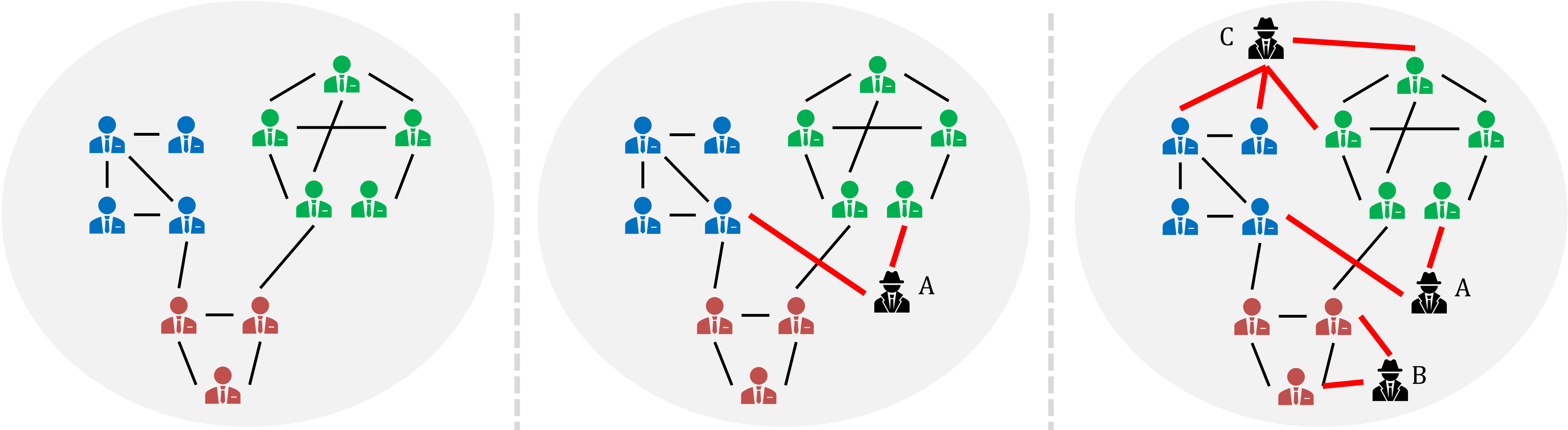
BACKGROUND

- GNN has manifested **promising performance** in many graph tasks.
- However, GNN prediction results **lack human-intelligible explanation**.
- Hence, GNN explanation method aims to identify **explanatory subgraph**.



BACKGROUND

- Existing GNN explanation methods introduce **noise-free assumption**.
- **Minor** corruption and **severe** corruption.



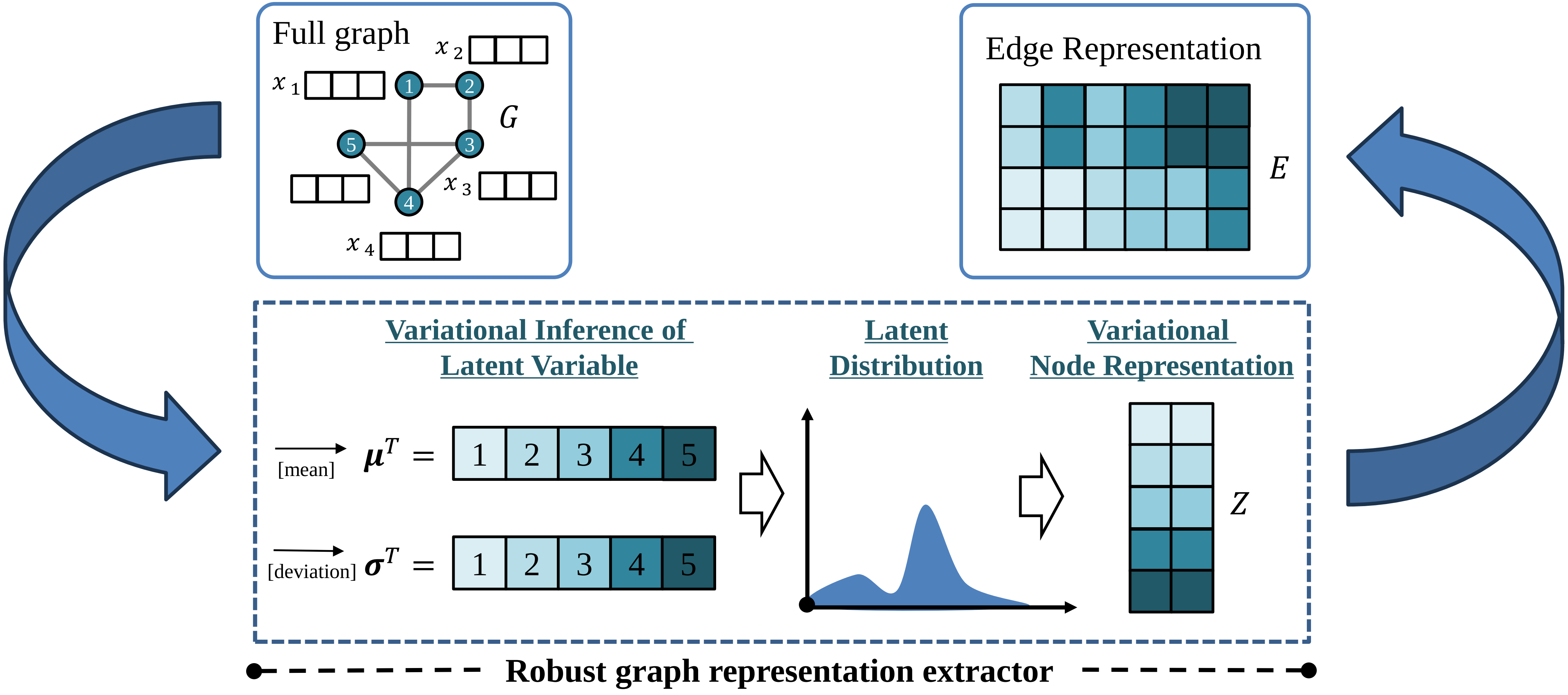
(a) Raw graph G
$f(G) = \hat{y}$

(b) Minor corruption
$f(G \cup \{A\}) = \hat{y}$

(c) Severe corruption
$f(G \cup \{A, B, C\}) \neq \hat{y}$

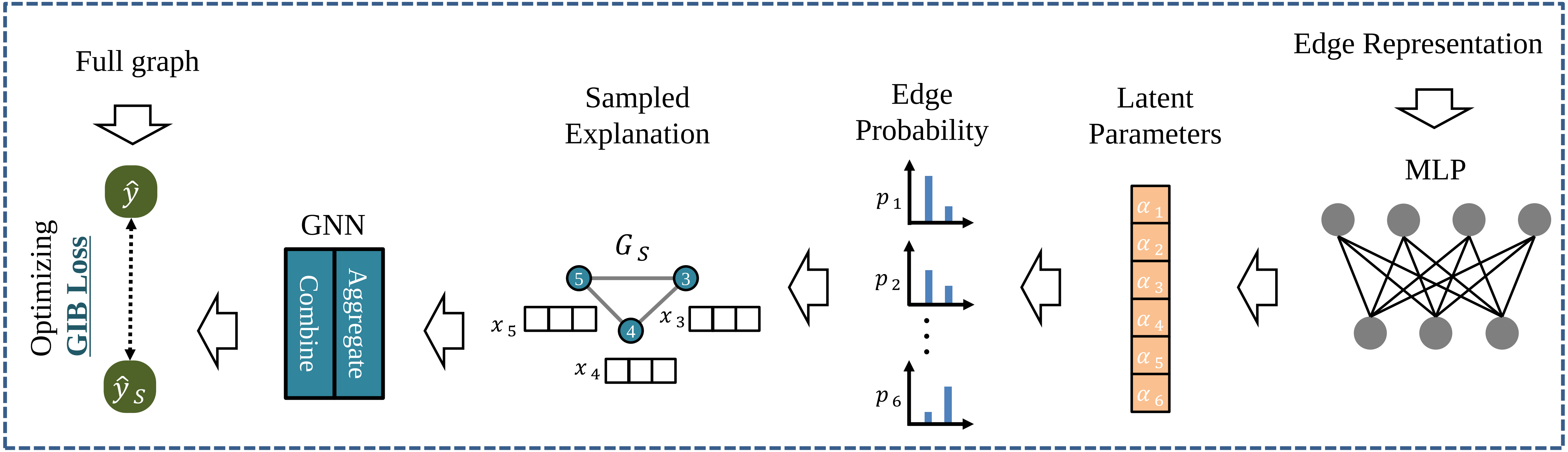
MODEL FRAMEWORK

- Robust graph representation extractor based on variational inference.



MODEL FRAMEWORK

- Adaptive explanation generator based on graph information bottleneck.



Adaptive explanation generator

$$\min_{G_S \subset G} \mathrm{GIB}(G, \hat{y}; G_S) = -\mathrm{MI}(\hat{y}, G_S) + \beta \mathrm{MI}(G, G_S)$$

EXPERIMENT

- Graphs with **random noise**.

Table 1: The comparison of V-InfoR and baselines under random structural noise. We use bold font to mark the highest score. The second highest score is marked with underline. The Impro. is defined as ([V-InfoR]-[Best Baseline]) / [Best Baseline].

Dataset	Metric	GradCAM	IG	GNNExplainer	PGExplainer	PGM-Explainer	ReFine	V-InfoR	Rank	Impro.
BA-3Motifs	P_S	<u>0.8725</u>	0.8625	0.8535	0.8510	0.8505	0.8300	**0.8820**	1	1.09%
	P_N	0.2605	<u>0.2795</u>	0.2410	0.2095	0.2235	0.2625	**0.3021**	1	8.09%
	F_{NS}	0.4012	<u>0.4222</u>	0.3758	0.3362	0.3540	0.3989	**0.4610**	1	9.19%
Mutag	P_S	0.8760	0.8880	<u>0.8916</u>	0.8640	0.8900	0.8900	**0.8964**	1	0.54%
	P_N	0.0996	0.1068	0.1080	<u>0.1260</u>	0.1020	<u>0.1260</u>	**0.1696**	1	34.60%
	F_{NS}	0.1789	0.1907	0.1920	0.2199	0.1830	<u>0.2207</u>	**0.2852**	1	29.23%
Ogbg-molhiv	P_S	<u>0.9230</u>	0.9200	0.8925	0.8390	0.8860	0.9105	**0.9386**	1	1.69%
	P_N	0.0680	0.0400	0.0940	<u>0.1265</u>	0.0980	0.1020	**0.1470**	1	16.21%
	F_{NS}	0.1267	0.0767	0.1701	<u>0.2198</u>	0.1765	0.1834	**0.2542**	1	15.65%
Ogbg-ppa	P_S	0.4340	0.5820	<u>0.6616</u>	0.6260	0.6192	0.6344	**0.6700**	1	1.27%
	P_N	<u>0.4720</u>	0.4600	0.3480	0.2856	0.3780	0.4406	**0.4930**	1	4.45%
	F_{NS}	0.4522	0.5139	0.4561	0.3922	0.4694	<u>0.5200</u>	**0.5680**	1	9.23%

EXPERIMENT

- Graphs with **adversarial attack**.

Table 2: The comparison of V-InfoR and baselines under GRABNEL attack [18]. We use bold font to mark the highest score. The second highest score is marked with underlines.

Attack budegt	Dataset	Metric	GradCAM	IG	GNNExplainer	PGExplainer	PGM-Explainer	ReFine	V-InfoR	Rank	Impro.
5%	BA-3Motifs	P_S	0.6980	0.6925	0.5625	0.6225	0.5950	0.6700	**0.7075**	1	1.36%
		P_N	0.3625	0.4675	0.4200	0.3700	0.3925	0.3925	**0.5450**	1	16.58%
		F_{NS}	0.4772	0.5582	0.4809	0.4641	0.4730	0.4950	**0.6157**	1	10.30%
	Mutag	P_S	0.5740	0.6600	0.6140	0.6610	0.5820	0.6340	**0.6760**	1	2.27%
		P_N	0.4200	0.3875	0.3800	0.4003	0.4060	0.4100	**0.4588**	1	9.24%
		F_{NS}	0.4851	0.4883	0.4695	0.4986	0.4783	0.4980	**0.5466**	1	9.63%
10%	BA-3Motifs	P_S	0.8720	0.8495	0.8605	0.9020	0.8125	0.8800	**0.9185**	1	1.83%
		P_N	0.0800	0.2105	0.2615	0.2815	0.1925	0.2100	**0.3332**	1	18.37%
		F_{NS}	0.1466	0.3374	0.4011	0.4291	0.3113	0.3391	**0.4890**	1	13.96%
	Mutag	P_S	0.5848	0.7370	0.6616	0.6524	0.6392	0.6140	**0.7424**	1	0.73%
		P_N	0.4160	0.3404	0.3284	0.3928	0.3344	0.4040	**0.4277**	1	2.81%
		F_{NS}	0.4862	0.4657	0.4389	0.4904	0.4391	0.4873	**0.5427**	1	10.66%

SUMMARY

- Category: Minor corruption and severe corruption.
- A robust GNN explainer for structurally corrupted graphs.
- Robust graph representation extractor based on variational inference.
- Adaptive explanation generator based on graph information bottleneck.

NEURAL INFORMATION PROCESSING SYSTEMS

THANKS