THE WEB CONFERENCE ACM

Lyon, France, 25-29 April 2022

AUC-oriented Graph Neural Network for Fraud Detection

Mengda Huang[1], Yang Liu[1], Xiang Ao[1*], Kuan Li[1],
Jianfeng Chi[2], Jinghua Feng[2], Hao Yang[2], Qing He[1*]

1 中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

2 Alibaba Group 阿里巴巴集团

* denotes corresponding author(s)

CONTENT

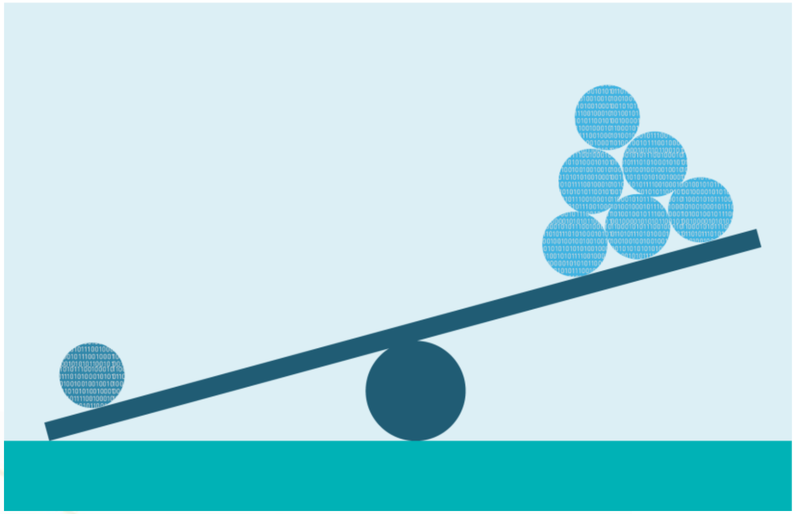
- Background
- Method: AO-GNN
- Experiments
- Conclusion

BACKGROUND: FRAUD DETECTION TASK

Fraud detection tasks become easier when discrete entities are built as graphs.

Task: Distinguish whether a node in a graph is a fraudulent node or a benign node.

BACKGROUND: CHALLENGES



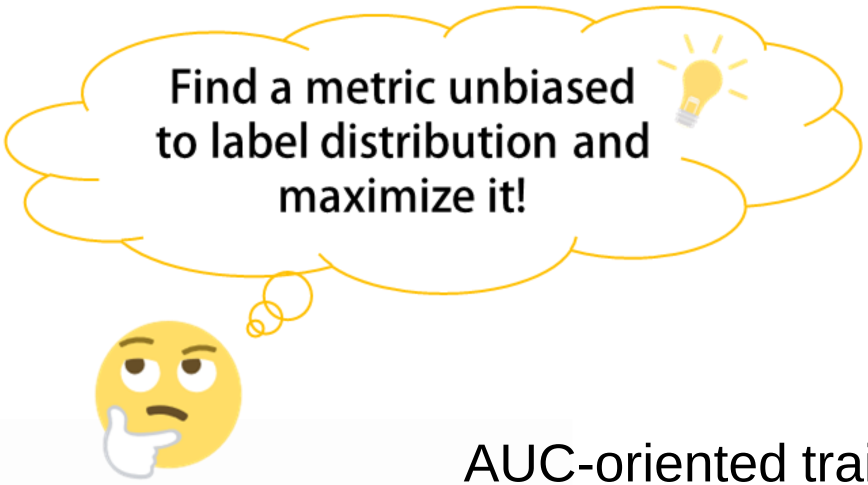
I got 99.9% accuracy by classifying all nodes as benign nodes!

Label Imbalance:
In practical scenarios, fraud samples are often very rare.

→

Overfitting on Majority:
Such models are highly accurate but fail to learn from the data to identify fraudsters.

BACKGROUND: MOTIVATION



AUC-oriented training tends to obtain a model with the competitive ability for classifying both benign nodes and fraud nodes.

$$AUC(\mathcal{M}_{\mathcal{G}}^{\omega}) = \mathbb{P}(\mathcal{M}_{\mathcal{G}}(\omega; \mathbf{v}) \geq \mathcal{M}_{\mathcal{G}}(\omega; \mathbf{v}') | y_{\mathbf{v}} = 1, y_{\mathbf{v}'} = 0)$$

BACKGROUND : STOCHASTIC AUC MAXIMIZATION

$$AUC(\mathcal{M}_{\mathcal{G}}^{\omega}) = \mathbb{P}(\mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v}) \geq \mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v}')|y_{\mathbf{v}} = 1, y_{\mathbf{v}'} = 0)$$

Maximize its unbiased estimation

$$\max_{\omega} \mathbb{E}_{\mathbf{v},\mathbf{v}'}[\mathbb{I}(\mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v}) \geq \mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v}')|y_{\mathbf{v}} = 1, y_{\mathbf{v}'} = 0]$$

L2 convex surrogates

$$\min_{\omega\in\mathbb{R}^d} \mathbb{E}_{\mathbf{v},\mathbf{v}'}[(1-(\mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v})-\mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v}'))^2|y_{\mathbf{v}}=1, y_{\mathbf{v}'}=0]$$

BACKGROUND : STOCHASTIC AUC MAXIMIZATION

$$\min_{\omega \in \mathbb{R}^d} \mathbb{E}_{\mathbf{v},\mathbf{v}'}[(1-(\mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v})-\mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v}'))^2|y_{\mathbf{v}}=1, y_{\mathbf{v}'}=0]$$

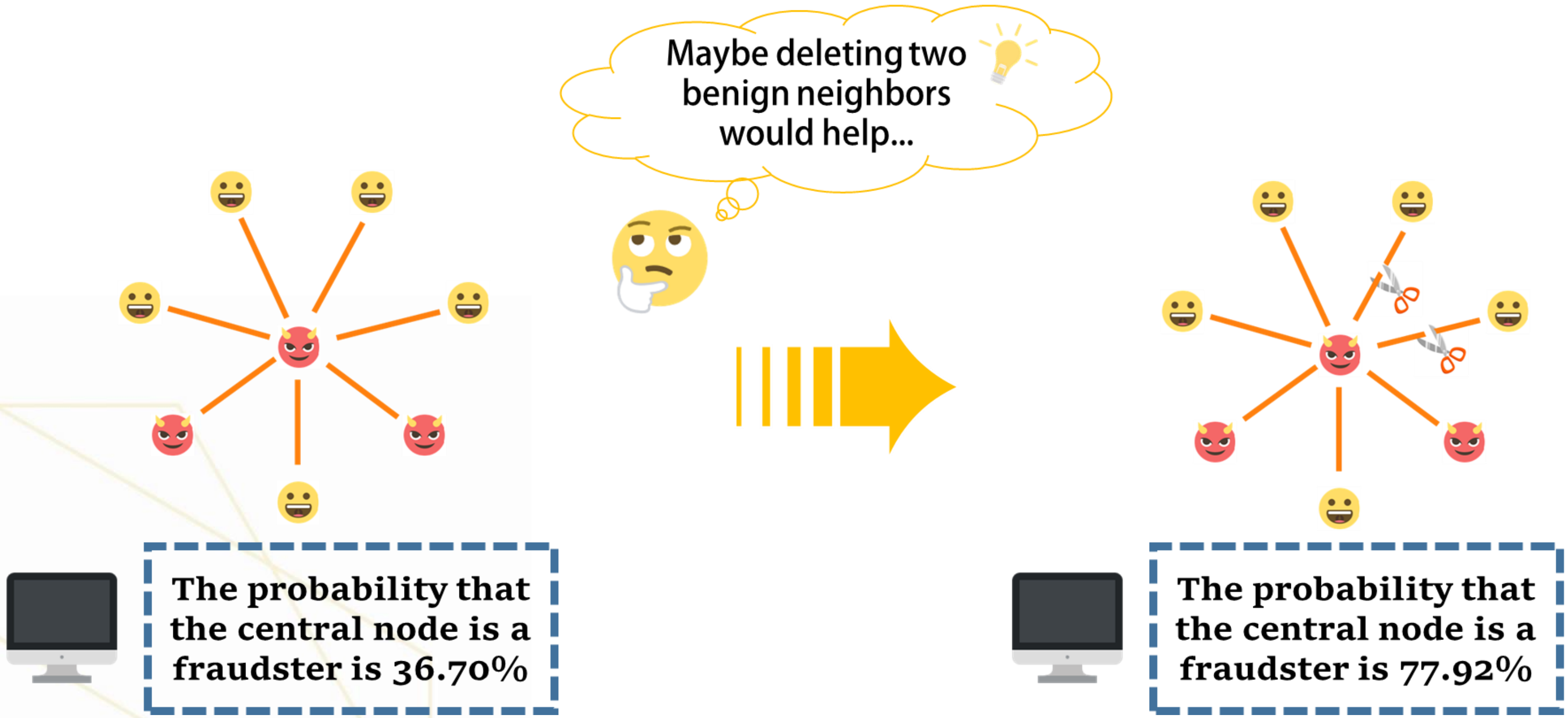
THEOREM 1. *The optimizing problem above is equivalent to*

$$\min_{\omega \in \mathbb{R}^d, \{a,b\} \in \mathbb{R}^2} \max_{\alpha \in \mathbb{R}} \mathbf{E}_{\mathbf{v}}[\mathcal{L}_{AUC}(\omega, a, b, \alpha, \mathbf{v}|p)],$$

where p is the ratio of fraud nodes, and

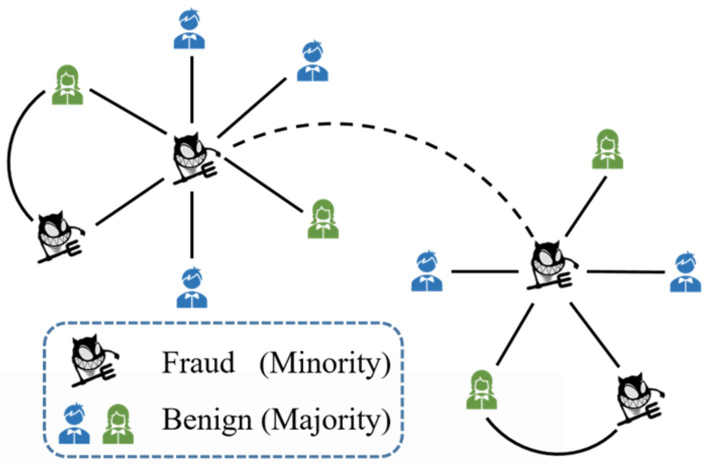
$$\begin{aligned}\mathcal{L}_{AUC}(\omega, a, b, \alpha, \mathbf{v}|p) = \mathbb{I}(y=1)[(1-p)(\mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v})-a)^2 \\ + 2(p-1)(1+\alpha)\mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v})] \\ + \mathbb{I}(y=0)[p(\mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v})-b)^2 \\ + 2p(1+\alpha)\mathcal{M}_{\mathcal{G}}(\omega;\mathbf{v})] + p(1-p)\alpha^2.\end{aligned}$$

BACKGROUND: OBSERVATION



AUC is not maximized!

BACKGROUND: CHALLENGES



$$\arg\max_{\omega} AUC(\mathcal{M}_{\mathcal{G}}^{\omega})$$

$$\arg\max_{\omega,\Pi} AUC(\mathcal{M}_{\mathcal{G}}^{\omega}|\Pi)$$

$\prod$ is a topological cleaner

Polluted Topology:
Fraudulent nodes often confuse their identities by interacting with other nodes.

AUC-oriented GNN: SCHEMATIC

Algorithm 1: AUC maximization with GNN

input : A multi-relation graph $\mathcal{G}$, a GNN architecture $\mathcal{M}_{\mathcal{G}}$

Initialize $\mathcal{M}_{\mathcal{G}} \leftarrow \mathcal{M}_{\mathcal{G}}^{\omega}$ with random parameters;

Initialize Π as a random function;

while *Break Condition is False* **do**

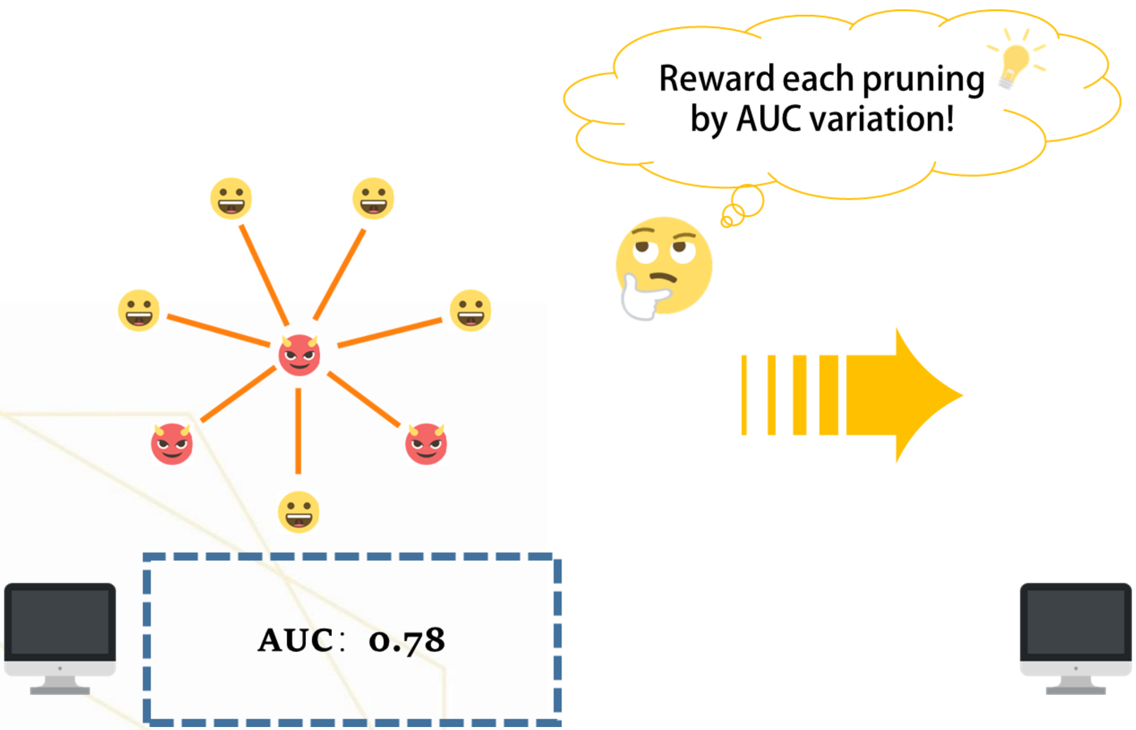
    $\omega \leftarrow \arg\max_{\omega} \; AUC(\mathcal{M}_{\mathcal{G}}^{\omega}|\Pi)$; // Parameter searching

    $\Pi \leftarrow \arg\max_{\Pi} \; AUC(\mathcal{M}_{\mathcal{G}}^{\omega}|\Pi)$; // Policy searching

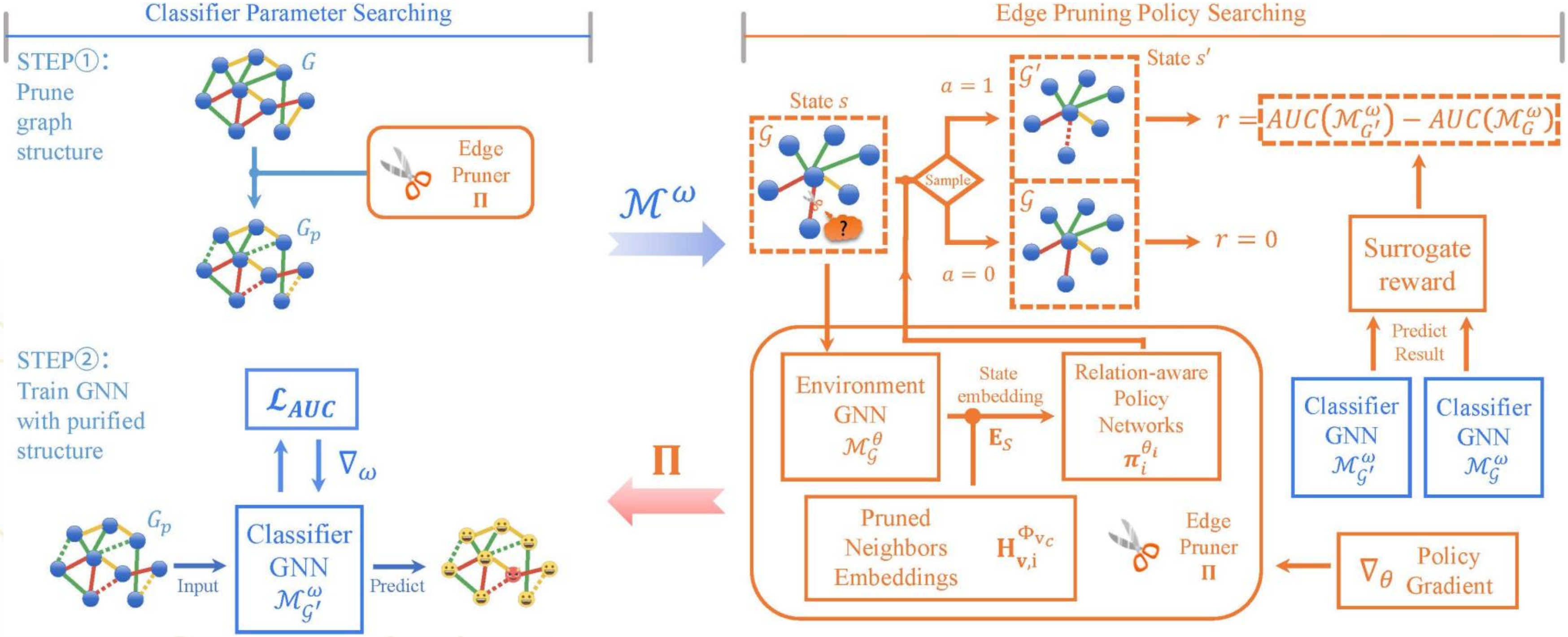
end

return : A GNN $\mathcal{M}_{\mathcal{G}}^{\omega}$, Pruning policy Π

AO-GNN: EDGE PRUNING MDP



AO-GNN: MODEL OVERVIEW



AO-GNN: ACCELERATING MECHANISMS

- Parameter sharing: all policy networks to share a GNN layer
- Halting Mechanism: restrict the maximum number of deleting for each node
- Surrogate reward: reduces the complexity of returning reward from $O(nlogn)$ to $O(n)$

$$\mathcal{R}(s, a, s') = F(y_{\mathrm{v}_c}) \cdot \begin{cases} |\{p_{\mathrm{v}} | p_{\mathrm{v}} \in (p_{\mathrm{v}_c}, p'_{\mathrm{v}_c}], y_{\mathrm{v}} \neq y_{\mathrm{v}_c}\}|, & p'_{\mathrm{v}_c} \geq p_{\mathrm{v}_c} \\ -|\{p_{\mathrm{v}} | p_{\mathrm{v}} \in [p'_{\mathrm{v}_c}, p_{\mathrm{v}_c}), y_{\mathrm{v}} \neq y_{\mathrm{v}_c}\}|, & p'_{\mathrm{v}_c} < p_{\mathrm{v}_c} \end{cases}.$$

EXPERIMENT: DATASET

Our datasets vary in many dimensions.
YelpChi: spam reviews detection
Amazon: fraud users detection
Books: fake-order item detection

Dataset	#Node	#Edge	Relations	Relation#Edges
YelpChi	45,954; 14.5% Fraud	3,846,979	R-U-R	49,315
			R-T-R	573,616
			R-S-R	3,402,743
Amazon	11,944; 9.5% Fraud	4,398,392	U-P-U	165,608
			U-S-U	3,566,479
			U-V-U	1,036,737
Books	1,418; 1.9%	3,695	Co-purchase	3,695

EXPERIMENT: PERFORMANCE COMPARISON

Method	Dataset	YelpChi		
	Metric	AUC	F1-macro	GMean
Baselines	GCN	0.5983±0.0049	0.5620±0.0067	0.4365±0.0262
	GAT	0.5715±0.0029	0.4879±0.0230	0.1659±0.0789
	GraphSAGE	0.5439±0.0025	0.4405±0.1066	0.2589±0.1864
	DR-GCN	0.5921±0.0195	0.5523±0.0231	0.4038±0.0742
	GraphConsis	0.6983±0.0302	0.5870±0.0200	0.5857±0.0385
	CARE-GNN	0.7619±0.0292	0.6332±0.0094	0.6791±0.0359
	PC-GNN	0.8178±0.0014	0.6400±0.0230	0.7395±0.0130
Ablation	AO-GNN$_{woP}$	0.8680±0.0020	**0.7182±0.0177**	0.7484±0.0125
	AO-GNN$_{woC}$	0.8545±0.0177	0.7063±0.0129	0.7305±0.0241
	AO-GNN$_{R-P}$	0.8302±0.0286	0.6936±0.0351	0.7192±0.0586
Ours	AO-GNN	**0.8805±0.0008**	0.7042±0.0051	**0.8134±0.0232**

Method	Dataset	Amazon		
	Metric	AUC	F1-macro	GMean
Baselines	GCN	0.8369±0.0125	0.6408±0.0694	0.5718±0.1951
	GAT	0.8102±0.0179	0.6464±0.0387	0.6675±0.1345
	GraphSAGE	0.7589±0.0046	0.6416±0.0079	0.5949±0.0349
	DR-GCN	0.8295±0.0079	0.6488±0.0364	0.7963±0.0091
	GraphConsis	0.8741±0.0334	0.7512±0.0325	0.7677±0.0486
	CARE-GNN	0.9067±0.0112	**0.8990±0.0073**	0.8962±0.0018
	PC-GNN	0.9586±0.0014	0.8956±0.0077	0.9030±0.0044
Ablation	AO-GNN$_{woP}$	0.9588±0.0008	0.8956±0.0026	0.8740±0.0137
	AO-GNN$_{woC}$	0.9392±0.0166	0.8914±0.0041	0.8828±0.0267
	AO-GNN$_{R-P}$	0.9197±0.0238	0.8827±0.0135	0.8602±0.0164
Ours	AO-GNN	**0.9640±0.0020**	0.8921±0.0045	**0.9096±0.0105**

RQ1 Does AO-GNN outperform state-of-the-art GNN-based fraud detection models?

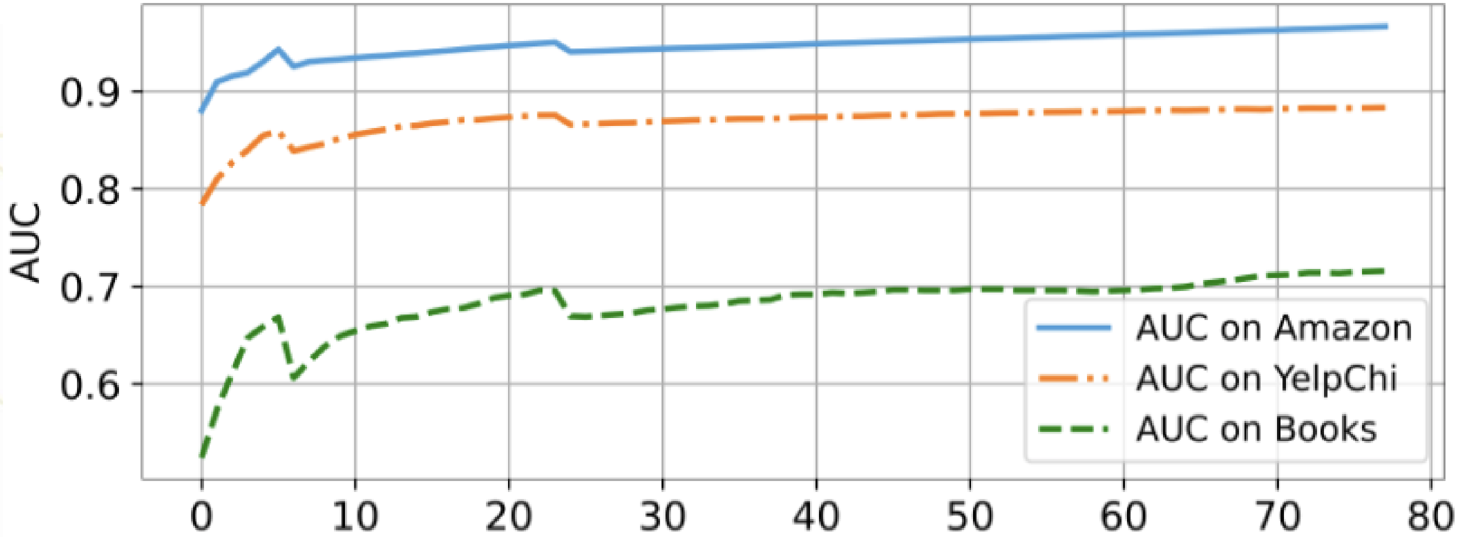
RQ2 How significant are the classifier parameter searching and the edge pruning policy searching in boosting AUC?

Method	Dataset	Books		
	Metric	AUC	F1-macro	GMean
Baselines	GCN	0.4538±0.1977	0.2374±0.2065	0.0000±0.0000
	GAT	0.4006±0.2023	0.2058±0.1623	0.0000±0.0000
	GraphSAGE	0.4761±0.1508	0.2464±0.2004	0.0000±0.0000
	DR-GCN	0.5131±0.1579	0.3048±0.2454	0.0000±0.0000
	GraphConsis	0.5647±0.1281	0.2912±0.1325	0.0000±0.0000
	CARE-GNN	0.6267±0.0462	0.4050±0.0996	0.4861±0.0811
	PC-GNN	0.6431±0.0189	0.4951±0.0037	0.5244±0.1012
Ablation	AO-GNN$_{woP}$	0.6720±0.0111	0.4131±0.0102	0.4829±0.0519
	AO-GNN$_{woC}$	0.5821±0.1397	0.2901±0.2102	0.3711±0.1919
	AO-GNN$_{R-P}$	0.5604±0.1733	0.2845±0.2329	0.3068±0.1240
Ours	AO-GNN	**0.7174±0.0158**	**0.5503±0.0141**	**0.6127±0.0252**

EXPERIMENT: AUC EVOLVING PROCESS STUDY

RQ3 How does AUC evolved in parameter searching?

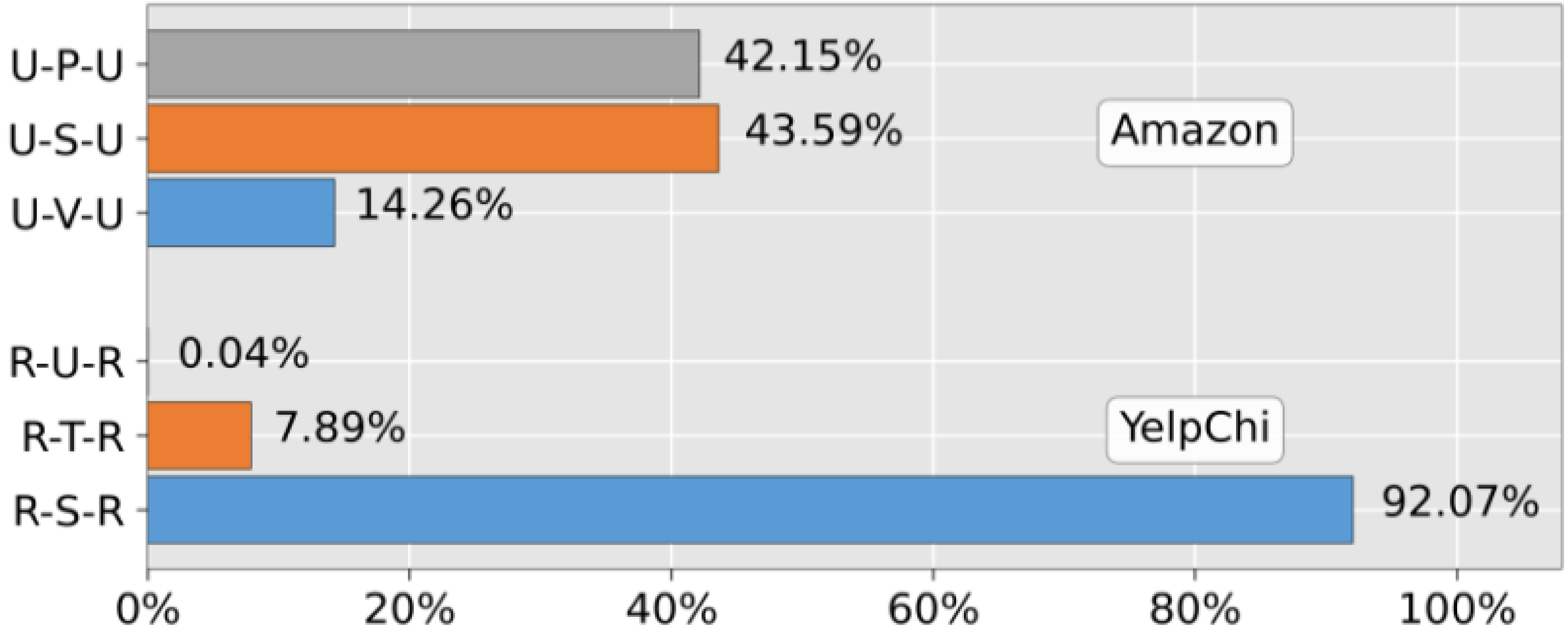
- After each falling, the AUC curves gain a longer-lasting growth and higher upper bound



EXPERIMENT: PRUNED EDGES STUDY

RQ4 What kind of edges are easier to be pruned?

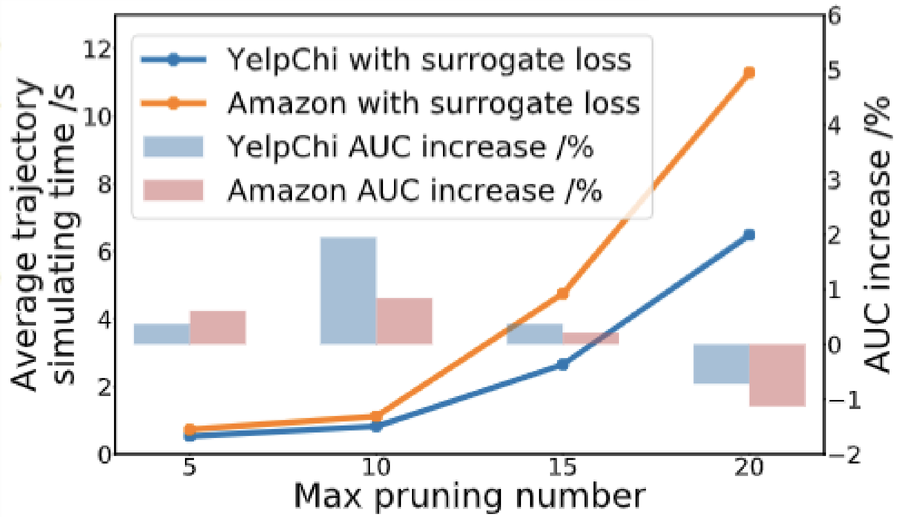
- R-U-R edges are hardly pruned in YelpChi.
- U-P-U edges are very noisy in Amazon.

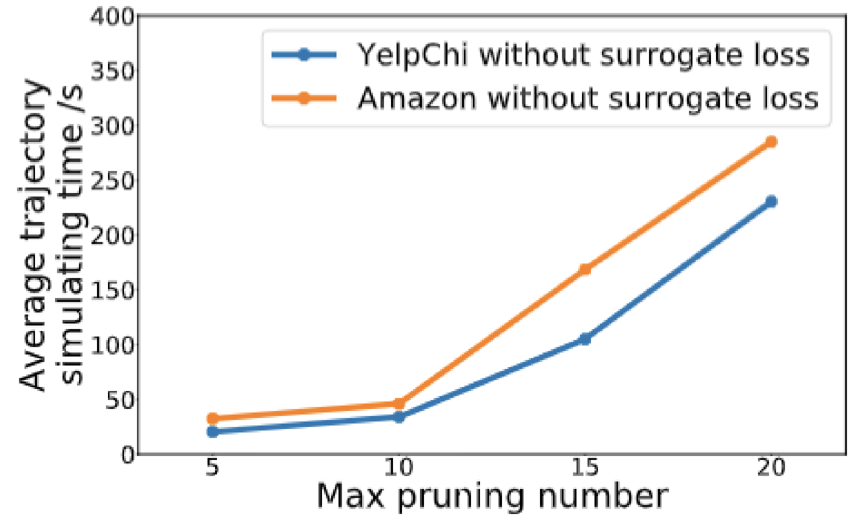


EXPERIMENT: ACCELERATE MECHANISMS STUDY

RQ5 How effective are RL accelerating mechanisms?

- Surrogate loss accelerates trajectory simulating 25-30 times.
- Pruning 10 edges per node achieves best AUC improvement.





CONCLUSION

- We propose a novel GNN-based model for fraud detection from the standpoint of AUC maximization.
- We formulate neighbors choosing as an MDP with a theoretical guarantee of maximizing AUC and solve it by Deep RL.
- Experiments on three public datasets demonstrate that AO-GNN clearly outperforms the state-of-the-art baselines.

AUC-oriented Graph Neural Network for Fraud Detection

Thanks for listening!

Email: huangmengda19s@ict.ac.cn

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Alibaba Group
阿里巴巴集团