Unveiling the Secrets without Data: Can Graph Neural Networks Be Exploited through Data-Free Model Extraction Attacks?

Yuanxin Zhuang Chuan Shi * Mengmei Zhang
Beijing University of Posts and Telecommunications
Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education

Jinghui Chen
The Pennsylvania State University

Lingjuan Lyu
SONY AI

Pan Zhou
Huazhong University of Science and Technology

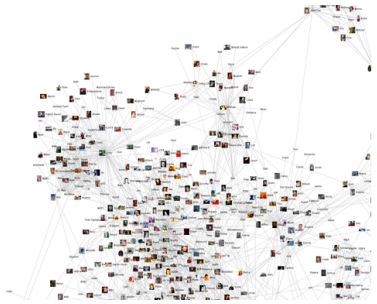
Lichao Sun
Lehigh University

Yuanxin ZHUANG

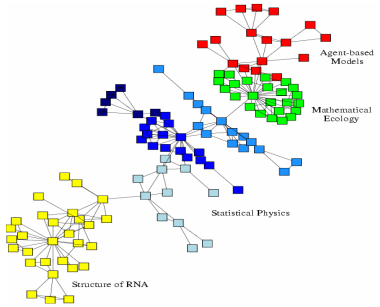
yzhuang436@connect.hkust-gz.edu.cn

Why Do Network Representation?

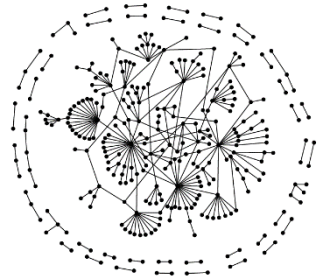
Networks are a general language for describing and modeling complex systems



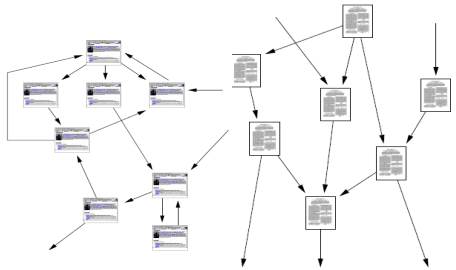
Social networks



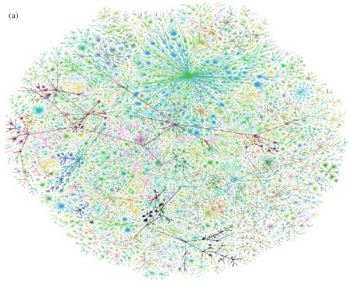
Economic networks



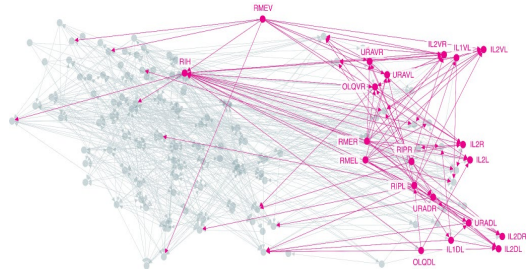
Biomedical networks



Information networks



Internet



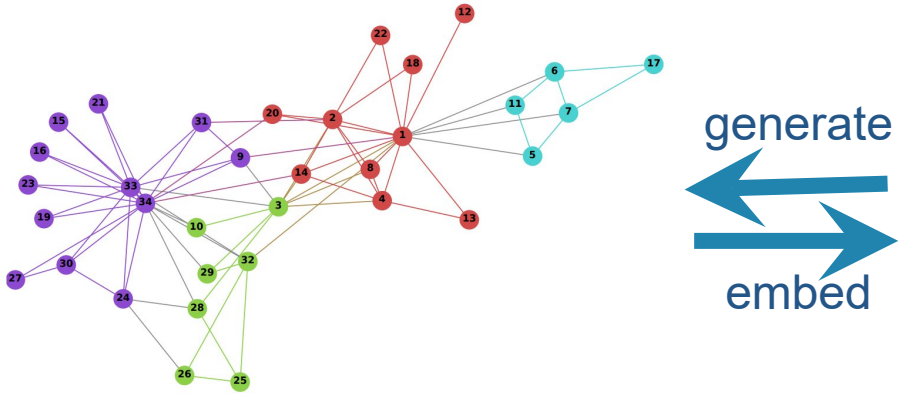
Networks of neurons

Network Representation

Network Representation

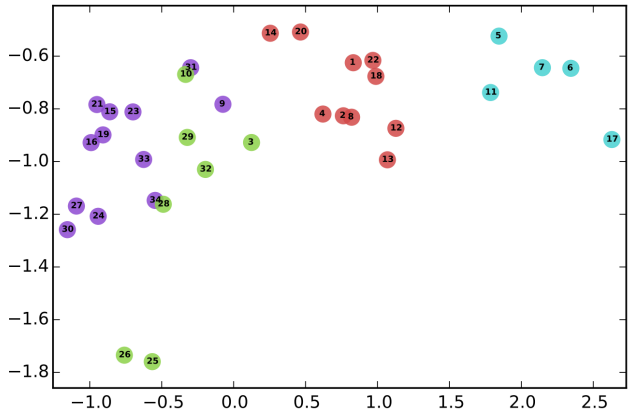
Embed each node of a network into a low-dimensional vector space

- Easy to parallel
- Can apply classical ML methods



Applications

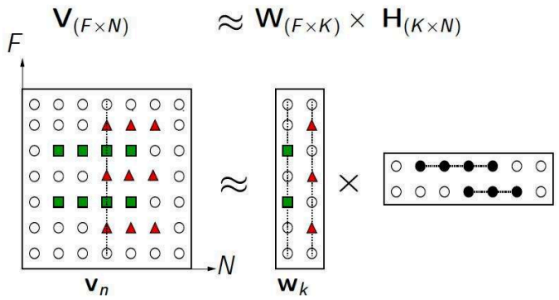
- Node classification
- Link predication
- Community detection
- Network evolution
- …



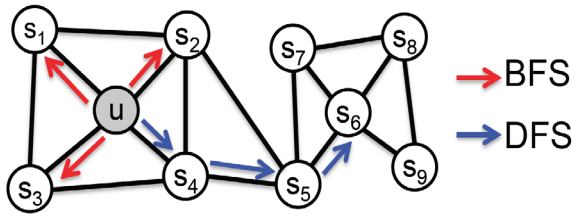
Network Representation

Shallow model

- Factorization based
  - e.g., Laplacian eigenmaps

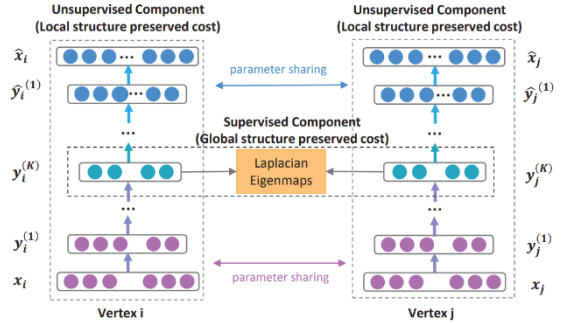


- Random walk based
  - e.g., DeepWalk, node2vec

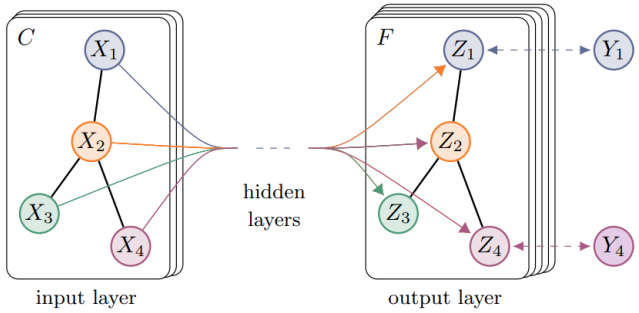


Deep model

- Autoencoder based
  - e.g., DNGR and SDNE

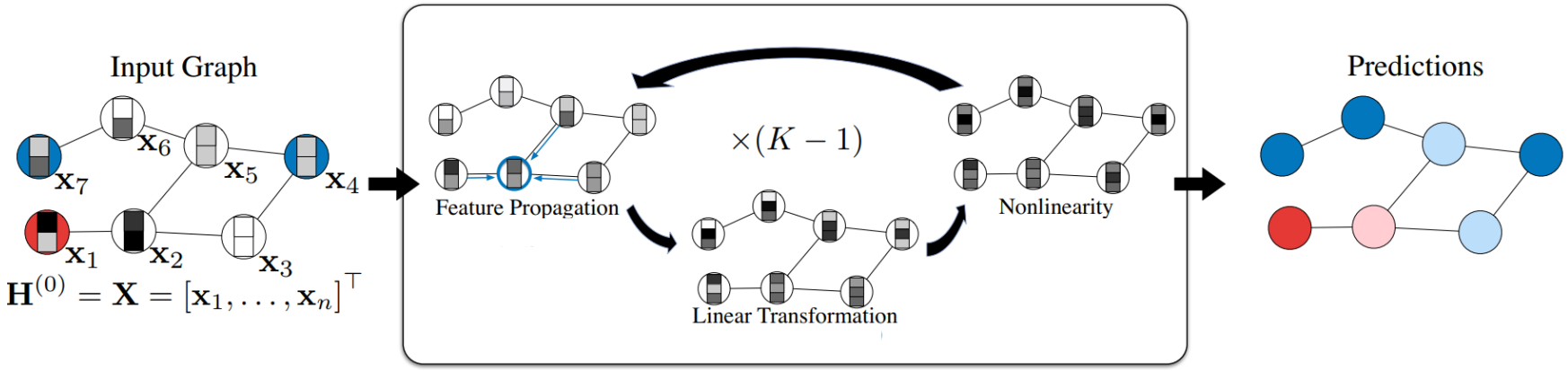


- GNN based
  - e.g., GCN, GraphSage, GAT



Graph Neural Networks

Learning Process of GNN



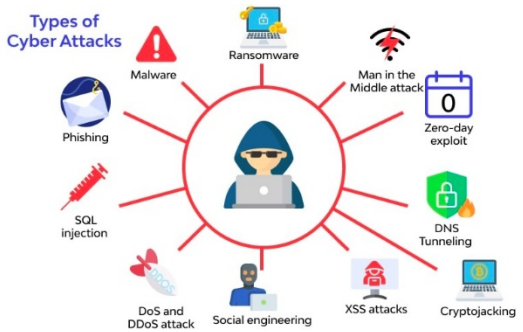
1. Input graph $\mathbf{A}$ and node feature $\mathbf{X}$
2. Initialization: $h_i^{(0)} = \sigma(W_0 X_i), \forall i$
3. k-th iteration: $h_i^{(k)} \leftarrow \sigma\left(W_1 h_i^{(k-1)} + W_2 \sum_{j \in \mathcal{N}(i)} h_j^{(k-1)}\right), \forall i$
4. Prediction: $\hat{\mathbf{Y}} = \mathrm{softmax}(\mathbf{A}\mathbf{H}^{(\mathrm{K}-1)}\mathbf{W})$

AI Security

- Focus on self security of AI algorithms
- Guarantee the integration and privacy of AI model and data
  - Privacy: The ability of protecting model and data from leakage
  - Robustness: Model adaptability to uncertainty, noise, and attacks

Cyber Security

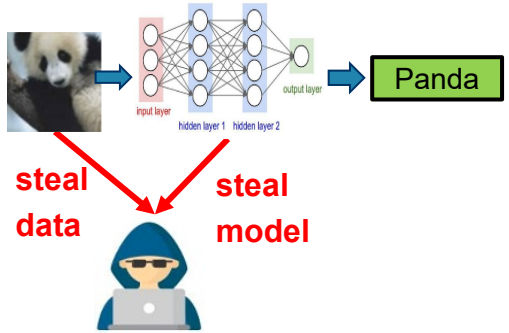
- Network facility security



AI Security

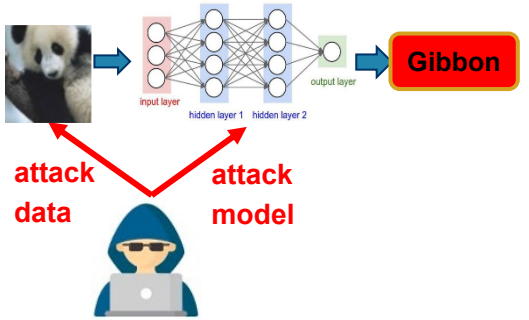
Privacy

- Data/Model steal



Robustness

- Adversarial attack/defense



Examples of AI Security

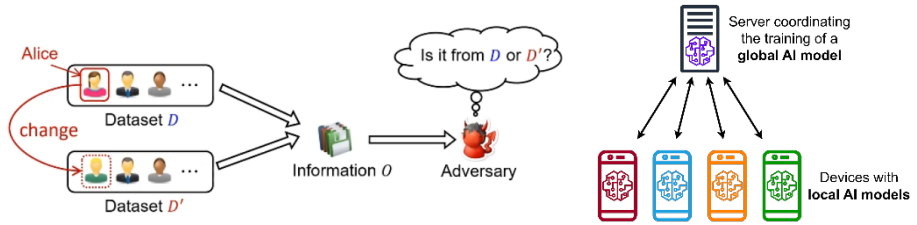
Privacy

Privacy Attack

Membership inference

Model extraction

Privacy Defense

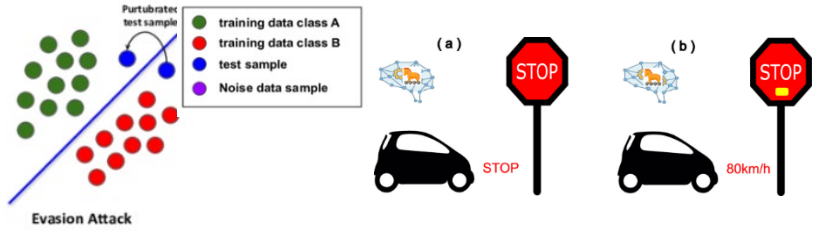


Differential privacy

Federated learning

Robustness

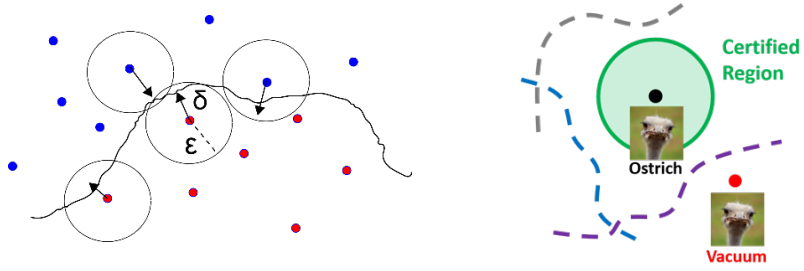
Adversarial Attack



Evasion attack

Backdoor attack

Adversarial Defense



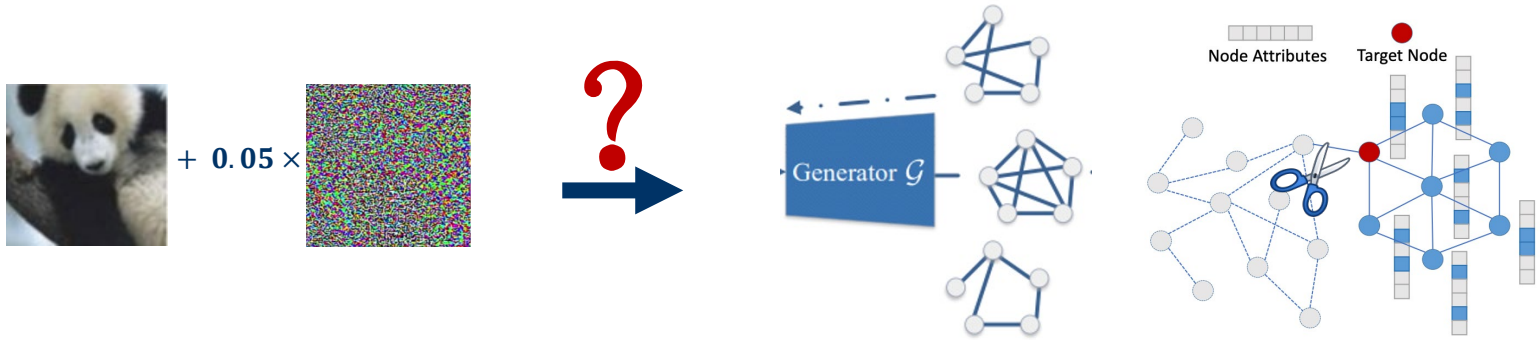
Adversarial training

Certifiable robustness

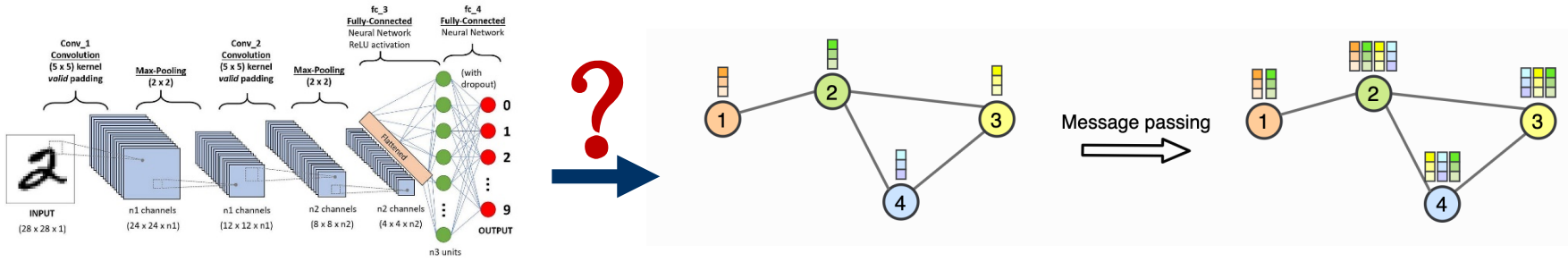
Security of GNNs

Challenges

- Unique data type
  - Discreteness graph structure



- Unique model design
  - Message-passing mechanism



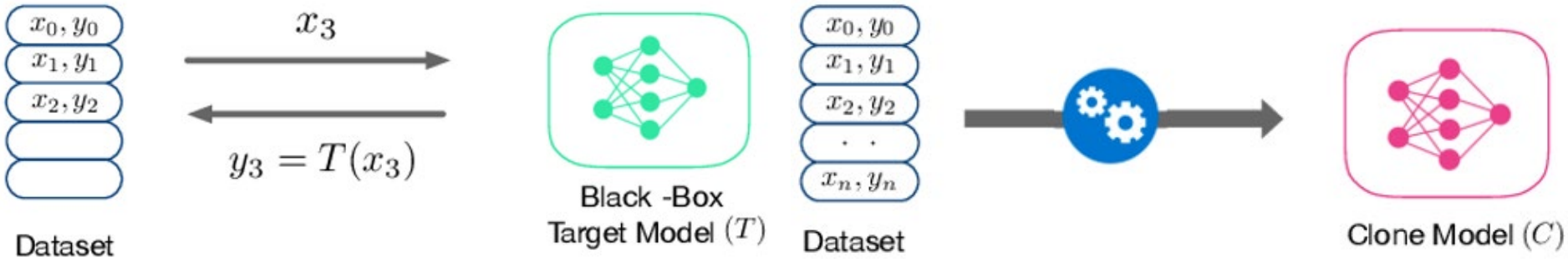
Model Extraction

Model Extraction

- Extracting model's parameters and structure for analysis, replication, and security assessment.

Steps of Model Extraction

- Generate data to capture the behavior of the target model.
- Train a new clone model using the constructed dataset.



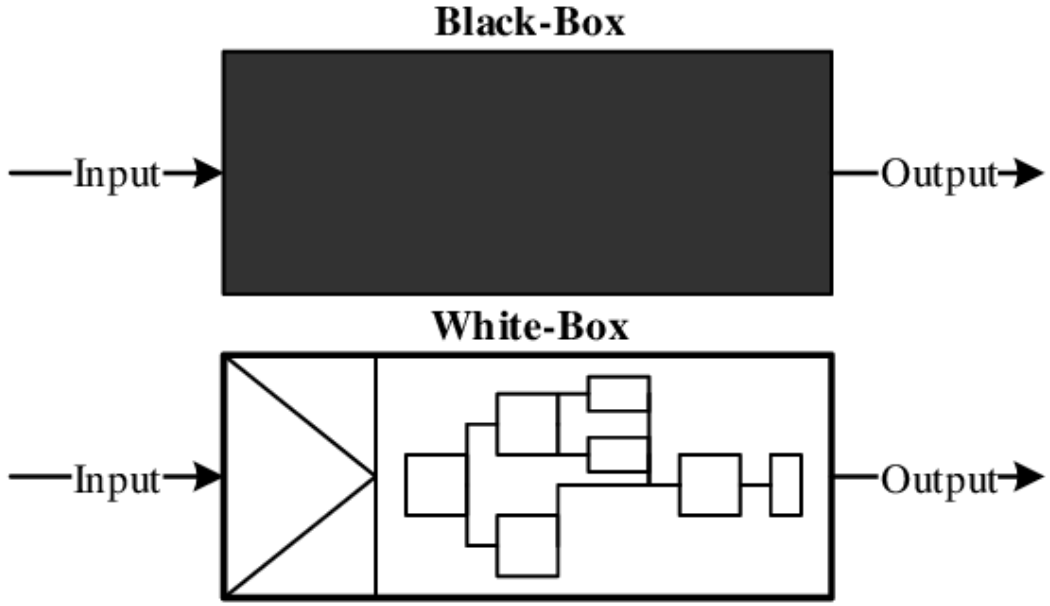
Step1: Construct Training Dataset **Step2: Train the Clone Model**

Kariyappa S, Prakash A, Qureshi M K. Maze: Data-free model stealing attack using zeroth-order gradient estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13814-13823.

Model Extraction

☐ Categories of Model Extraction

- White-box involves direct access to the internal parameters and structure of a model
- Black-box relies on observing the model's input-output behavior to infer its internal workings



Data-Free Model Extraction Attacks

Motivation

- Data is unobtainable and characterized by privacy concerns
- In real-world, pre-trained models are proprietary and black-box.

Comparison with Existing Work

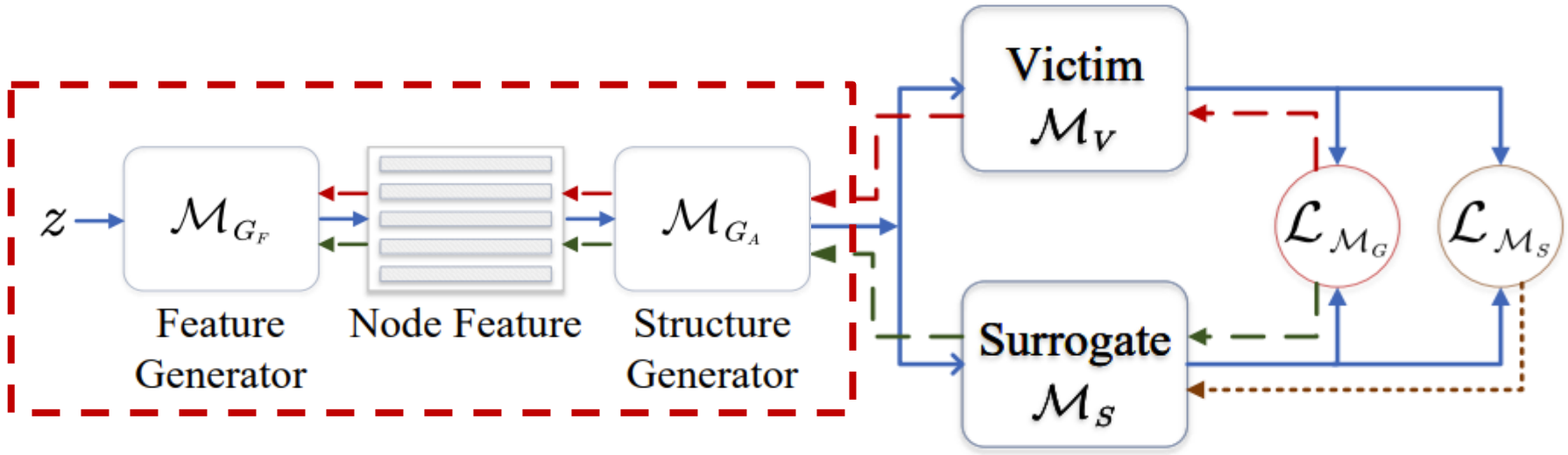
Method	Data-Free	Node Classification		Link Prediction
		Transductive	Inductive	
DeFazio et al.'s work [6]		✓		
Wu et al.'s work [42]		✓		
Shen et al.'s work [36]			✓	
StealGNN	✓	✓	✓	✓

Yuanxin Zhuang,Chuan Shi, Mengmei Zhang, Jinghui Chen, Lingjuan Lyu, Pan Zhou and Lichao Sun. Unveiling the Secrets without Data: Can Graph Neural Networks Be Exploited through Data-Free Model Extraction Attacks? USENIX Security 2024

Framework of StealGNN

Data-free Model Extraction Attack against GNN (StealGNN)

Model Framework



Step 1:

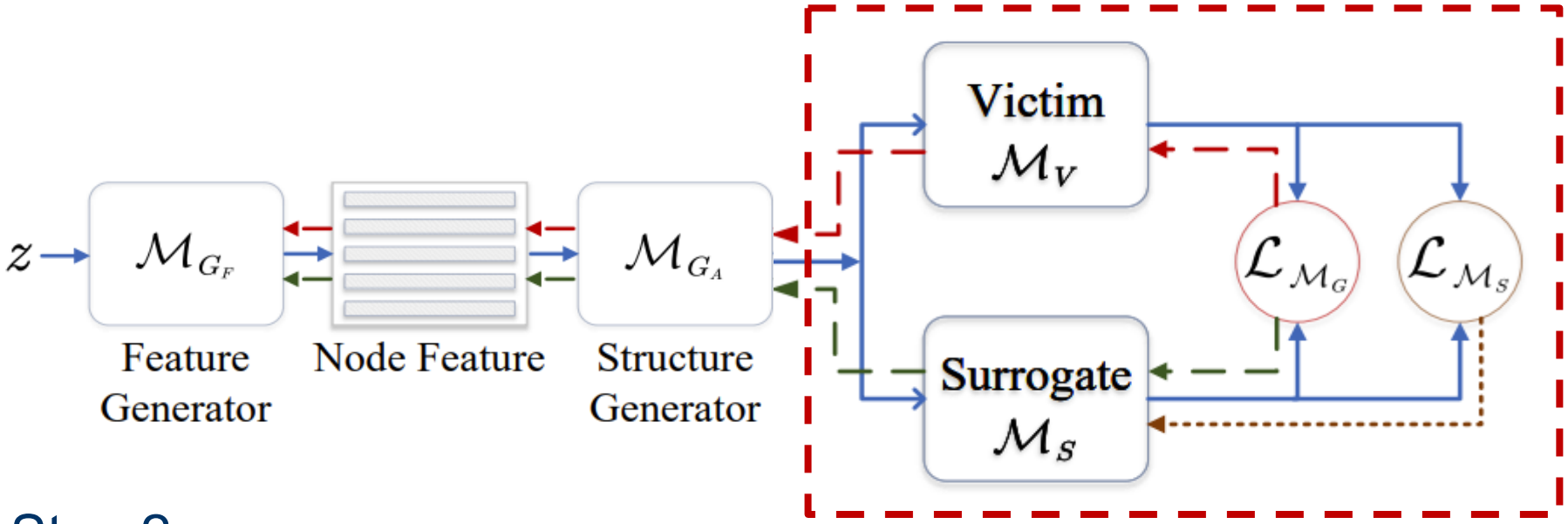
Generate graph from noise:

$$\mathbf{G} = \mathcal{M}_G(z;\theta_G); \quad z \sim \mathcal{N}(0, I),$$

Framework of StealGNN

Data-free Model Extraction Attack against GNN (StealGNN)

Model Framework



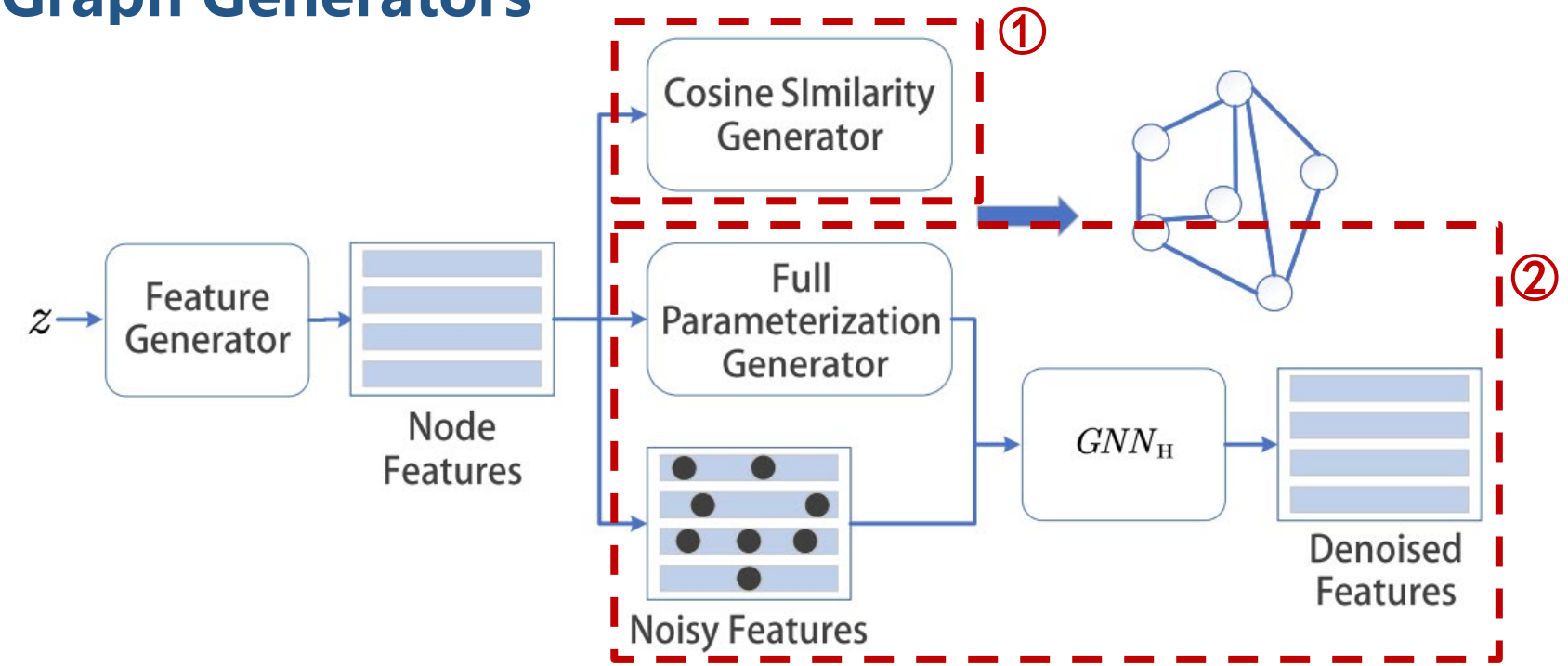
Step 2:

Extract model using generated graph :

$$\overrightarrow{\mathbf{Y}_V} = \mathcal{M}_V(\mathbf{G}, \theta_V); \quad \overrightarrow{\mathbf{P}_S} = \mathcal{M}_S(\mathbf{G}; \theta_S)$$

Framework of StealGNN

Graph Generators



Two structure generation methods:

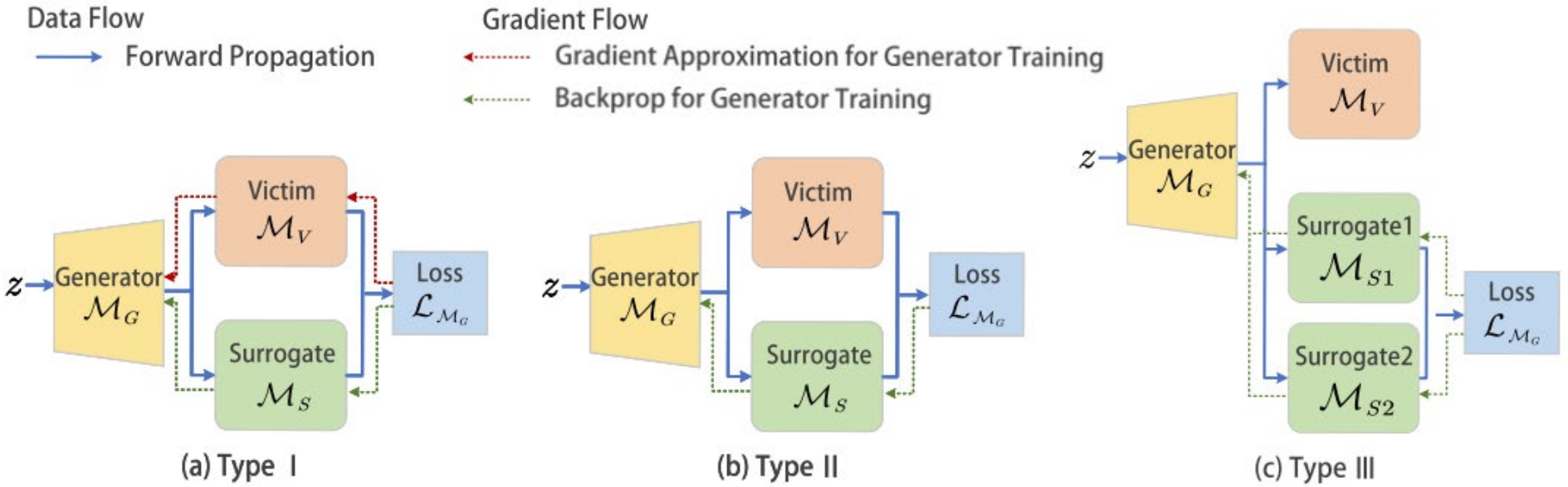
① Cosine Similarity: $\mathbf{A}_{cos_{ij}} = \mathrm{Cosine}(\mathbf{F}_i, \mathbf{F}_j) = \dfrac{\mathbf{F}_i \cdot \mathbf{F}_j}{\max\left(\|\mathbf{F}_i\|_2 \cdot \|\mathbf{F}_j\|_2\right)}$

② Full Parameterization:

$$\mathcal{L}_{\mathrm{denoise}} = \frac{1}{|\mathbf{F}_{idx}|} \sum_{i \in \mathbf{F}_{idx}} \mathcal{L}_{\mathrm{reconstruction}}\left(\mathbf{F}_i, \mathrm{H}\left(\tilde{\mathbf{F}}, \mathbf{A}_{fp}; \boldsymbol{\theta}_{\mathrm{H}}\right)_i\right)$$

Framework of StealGNN

☐ Generator Parameter Update Strategy



Type Ⅰ: Gradient Approximation + Surrogate. **Type Ⅱ:** One Surrogate.

$$\mathcal{L}_{\mathcal{M}_G} = -\mathcal{L}_{\mathcal{M}_S}$$

Type Ⅲ: Using the inconsistency between predictions of two surrogate models to generate informative graphs.

$$\mathcal{L}_{\mathcal{M}_G} = -\frac{1}{N}\sum_{i=1}^{N}\left[\mathrm{Std}\left(\mathbf{P}_{S1_i}, \mathbf{P}_{S2_i}\right)\right]$$

Experiments

☐ Model Performance (node classification)

$\mathcal{M}_V$	Dataset	Cora		Pubmed		A-Computers		OGB-Arxiv	
	$\mathcal{M}_S$(GCN)	Accuracy	Fidelity	Accuracy	Fidelity	Accuracy	Fidelity	Accuracy	Fidelity
GAT	Real Data	80.09±0.89 (+0.53)	92.96±0.51	76.94±0.31 (-0.12)	89.96±2.51	79.20±1.67 (-0.89)	90.51±1.12	53.89±1.02 (-0.77)	88.73±1.94
	Random Graph	68.26±3.84 (-11.30)	73.70±2.95	59.35±2.86 (-17.71)	71.79±2.49	55.63±4.61 (-24.46)	52.29±3.62	36.78±5.26 (-17.88)	63.52±4.39
	Attack I-$\mathbf{E}$	81.11±0.50 (+1.55)	93.09±0.26	**77.57±0.78** (+0.51)	92.09±2.56	65.68±1.91 (-14.41)	73.48±2.13	**52.17±2.94** (-2.49)	80.26±2.23
	Attack I-$\mathbf{A}_{cos}$	80.79±0.75 (+1.23)	93.90±0.42	77.15±0.93 (+0.09)	90.34±2.94	70.73±1.90 (-9.36)	81.18±1.18	51.72±2.31 (-2.94)	78.77±1.68
	Attack I-$\mathbf{A}_{fp}$	**81.23±0.68** (+1.67)	94.10±0.38	77.40±0.87 (+0.34)	90.90±2.76	**71.12±1.78** (-8.54)	81.45±1.10	52.15±2.23 (-2.51)	79.22±1.58
	Attack II-$\mathbf{E}$	80.98±0.57 (+1.42)	92.78±0.39	77.1					
	Attack II-$\mathbf{A}_{cos}$	81.17±0.71 (+1.61)	93.19±0.47	77.4					
	Attack II-$\mathbf{A}_{fp}$	**81.35±0.66** (+1.79)	93.50±0.43	**77.58**					
	Attack III-$\mathbf{E}$	81.12±0.52 (+1.56)	93.02±0.3	77.3					
	Attack III-$\mathbf{A}_{cos}$	81.45±0.62 (+1.89)	93.80±0.4	77.6					
	Attack III-$\mathbf{A}_{fp}$	**81.68±0.58** (+2.12)	94.05±0.3	**77.80**					
GCN	Real Data	81.09±0.69 (+0.49)	92.84±0.29	77.06					
	Random Graph	69.83±2.61 (-10.78)	81.79±2.28	62.78					
	Attack I-$\mathbf{E}$	80.67±0.45 (+0.06)	93.57±0.96	75.9					
	Attack I-$\mathbf{A}_{cos}$	80.68±0.29 (+0.07)	93.77±1.37	76.7					
	Attack I-$\mathbf{A}_{fp}$	**80.91±0.24** (+0.30)	93.96±1.24	**76.7**					
	Attack II-$\mathbf{E}$	79.79±0.23 (-0.82)	92.81±0.79	74.1					
	Attack II-$\mathbf{A}_{cos}$	80.07±0.59 (-0.54)	92.84±0.31	76.4					
	Attack II-$\mathbf{A}_{fp}$	**81.12±0.28** (+0.51)	93.65±0.12	**76.7**					
	Attack III-$\mathbf{E}$	80.78±0.31 (+0.17)	93.39±0.15	76.4					
	Attack III-$\mathbf{A}_{cos}$	81.05±0.25 (+0.44)	93.62±0.10	76.6					
	Attack III-$\mathbf{A}_{fp}$	**81.28±0.21** (+0.67)	93.79±0.08	**76.8**					
SAGE	Real Data	79.84±1.25 (+0.51)	92.57±0.28	76.7					
	Random Graph	60.39±2.95 (-18.94)	78.39±3.38	58.7					
	Attack I-$\mathbf{E}$	**81.32±0.65** (+1.99)	93.01±0.12	77.0					
	Attack I-$\mathbf{A}_{cos}$	80.89±0.47 (+1.56)	93.97±0.49	77.2					
	Attack I-$\mathbf{A}_{fp}$	81.10±0.35 (+1.77)	93.85±0.33	**77.4**					
	Attack II-$\mathbf{E}$	**81.28±0.67** (+1.95)	93.92±0.70	76.8					
	Attack II-$\mathbf{A}_{cos}$	81.01±0.51 (+1.68)	95.14±0.17	77.1					
	Attack II-$\mathbf{A}_{fp}$	80.98±0.41 (+1.65)	93.68±0.39	**77.2**					
	Attack III-$\mathbf{E}$	**81.56±0.45** (+2.23)	93.28±0.09	77.2					
	Attack III-$\mathbf{A}_{cos}$	81.45±0.63 (+2.12)	93.10±0.10	77.3					
	Attack III-$\mathbf{A}_{fp}$	81.55±0.61 (+2.22)	93.15±0.09	**77.25±0.14** (+0.02)	92.90±0.41	**70.55±2.43** (-8.69)	86.65±1.32	58.55±3.78 (-10.63)	78.25±1.97

$\mathcal{M}_V$	Dataset	Cora	
	$\mathcal{M}_S$(GCN)	Accuracy	Fidelity
GAT	Real Data	80.09±0.89 (+0.53)	92.96±0.51
	Random Graph	68.26±3.84 (-11.30)	73.70±2.95
	Attack I-$\mathbf{E}$	81.11±0.50 (+1.55)	93.09±0.26
	Attack I-$\mathbf{A}_{cos}$	80.79±0.75 (+1.23)	93.90±0.42
	Attack I-$\mathbf{A}_{fp}$	**81.23±0.68** (+1.67)	94.10±0.38
	Attack II-$\mathbf{E}$	80.98±0.57 (+1.42)	92.78±0.39
	Attack II-$\mathbf{A}_{cos}$	81.17±0.71 (+1.61)	93.19±0.47
	Attack II-$\mathbf{A}_{fp}$	**81.35±0.66** (+1.79)	93.50±0.43
	Attack III-$\mathbf{E}$	81.12±0.52 (+1.56)	93.02±0.36
	Attack III-$\mathbf{A}_{cos}$	81.45±0.62 (+1.89)	93.80±0.41
	Attack III-$\mathbf{A}_{fp}$	**81.68±0.58** (+2.12)	94.05±0.39

Experiments

Model Performance (link prediction)

$\mathcal{M}_V$	Dataset	Cora	Pubmed	A-computers
GAT	Real Data	89.30±0.59 (-0.38)	65.70±1.58 (-4.67)	84.64±1.78 (+1.89)
	Random Graph	75.32±2.12 (-14.36)	50.21±3.89 (-20.16)	66.79±2.78 (-15.96)
	Attack I- $\mathbf{E}$	89.01±0.56 (-0.67)	64.78±1.61 (-5.59)	80.29±1.48 (-2.46)
	Attack I- $\mathbf{A}_{cos}$	88.71±0.32 (-0.97)	**64.98±1.92** (-5.39)	**80.62±1.19** (-2.13)
	Attack I- $\mathbf{A}_{fp}$	**89.12±0.22** (-0.56)	64.25±1.38 (-6.12)	79.92±0.92 (-2.83)
	Attack II- $\mathbf{E}$	88.96±0.63 (-0.72)	64.58±1.82 (-5.79)	**80.19±1.19** (-2.56)
	Attack II- $\mathbf{A}_{cos}$	**89.38±0.47** (-0.30)	64.39±1.25 (-5.98)	79.89±0.92 (-2.86)
	Attack II- $\mathbf{A}_{fp}$	88.95±0.62 (-0.73)	**66.12±1.61** (-4.25)	78.98±1.28 (-3.77)
	Attack III- $\mathbf{E}$	88.95±0.50 (-0.73)	**66.78±1.32** (-3.59)	79.47±0.98 (-3.28)
	Attack III- $\mathbf{A}_{cos}$	89.25±0.53 (-0.43)	65.78±1.42 (-4.59)	
	Attack III- $\mathbf{A}_{fp}$	**89.45±0.35** (-0.23)	66.25±1.15 (-4.12)	
GCN	Real data	89.59±0.23 (-0.09)	87.72±0.18 (-0.38)	
	Random Graph	80.12±2.18 (-9.56)	72.85±3.29 (-15.25)	
	Attack I- $\mathbf{E}$	90.78±0.73 (+1.10)	85.54±1.09 (-2.56)	
	Attack I- $\mathbf{A}_{cos}$	**90.83±0.41** (+1.15)	84.20±1.67 (-2.90)	
	Attack I- $\mathbf{A}_{fp}$	89.70±0.60 (-0.02)	**87.45±0.85** (-0.76)	
	Attack II- $\mathbf{E}$	**89.95±0.52** (+0.27)	87.69±0.74 (-0.41)	
	Attack II- $\mathbf{A}_{cos}$	89.09±0.25 (-0.59)	88.05±0.2 (-0.05)	
	Attack II- $\mathbf{A}_{fp}$	89.30±0.22 (-0.38)	**88.25±0.25** (+0.1	
	Attack III- $\mathbf{E}$	90.25±0.45 (+0.57)	87.85±0.62 (-0.25)	
	Attack III- $\mathbf{A}_{cos}$	90.40±0.42 (+0.72)	88.09±0.55 (-0.01)	
	Attack III- $\mathbf{A}_{fp}$	**90.65±0.40** (+0.97)	**88.12±0.52** (+0.02)	
SAGE	Real data	91.15±0.20 (-1.32)	87.90±0.09 (+1.06)	
	Random Graph	77.25±1.35 (-10.54)	75.60±1.75 (-11.34)	
	Attack I- $\mathbf{E}$	**88.68±0.54** (-3.79)	85.79±0.48 (-1.05)	
	Attack I- $\mathbf{A}_{cos}$	87.79±0.96 (-4.68)	86.94±0.51 (+0.00)	
	Attack I- $\mathbf{A}_{fp}$	88.05±0.45 (-4.42)	**87.30±0.55** (+0.36)	
	Attack II- $\mathbf{E}$	89.01±0.56 (-3.46)	86.77±0.39 (-0.07)	
	Attack II- $\mathbf{A}_{cos}$	**90.04±0.47** (-2.43)	**88.26±0.08** (+1.42)	
	Attack II- $\mathbf{A}_{fp}$	88.50±0.35 (-4.97)	87.80±0.50 (+0.96)	
	Attack III- $\mathbf{E}$	90.05±0.50 (-2.42)	87.32±0.30 (+0.48)	
	Attack III- $\mathbf{A}_{cos}$	90.31±0.40 (-2.16)	**88.15±0.10** (+1.31)	
	Attack III- $\mathbf{A}_{fp}$	**90.59±0.40** (-1.88)	87.43±0.25 (+0.59)	**84.78±1.20** (-1.52)

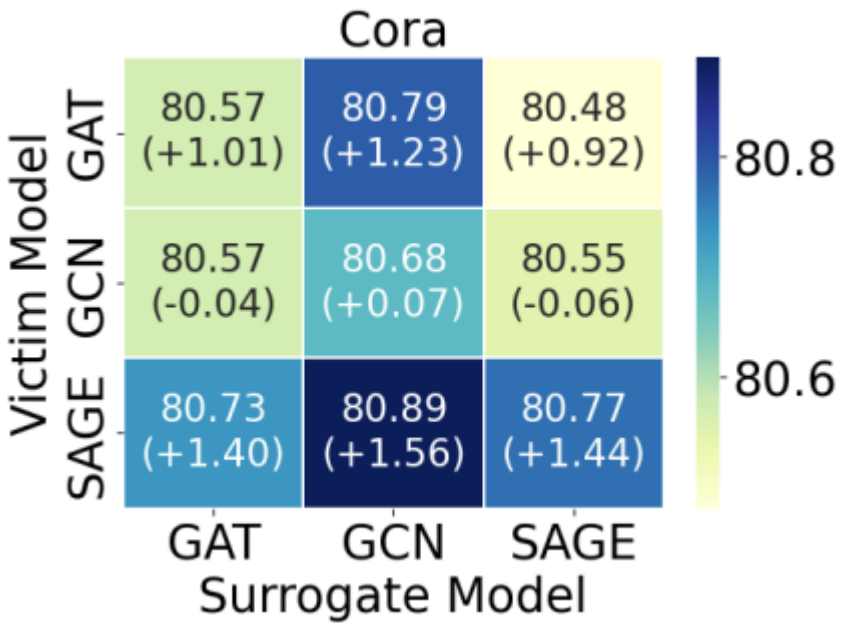
$\mathcal{M}_V$	Dataset	Cora	Pubmed	A-computers
GAT	Real Data	87.70±0.47	68.50±2.20	85.11±1.38
	Random Graph	78.25±0.35	52.10±1.95	72.80±1.05
	Attack I- $\mathbf{E}$	85.91±0.86	68.19±1.20	79.99±1.05
	Attack I- $\mathbf{A}_{cos}$	86.79±0.62	67.08±0.99	81.73±1.49
	Attack I- $\mathbf{A}_{fp}$	85.45±0.92	68.60±1.05	79.75±0.98
	Attack II- $\mathbf{E}$	89.96±0.83	66.29±1.52	79.61±1.38
	Attack II- $\mathbf{A}_{cos}$	90.72±0.67	65.42±1.65	80.59±1.17
	Attack II- $\mathbf{A}_{fp}$	90.40±0.72	65.75±1.58	80.85±1.10
	Attack III- $\mathbf{E}$	91.02±0.64	66.02±1.53	81.13±1.04
	Attack III- $\mathbf{A}_{cos}$			81.08±1.08
	Attack III- $\mathbf{A}_{fp}$	90.73±0.64	88.82±0.36	82.65±1.27

GCN	Real data	89.59±0.23 (-0.09)
	Random Graph	80.12±2.18 (-9.56)
	Attack I- $\mathbf{E}$	90.78±0.73 (+1.10)
	Attack I- $\mathbf{A}_{cos}$	**90.83±0.41** (+1.15)
	Attack I- $\mathbf{A}_{fp}$	89.70±0.60 (-0.02)
	Attack II- $\mathbf{E}$	**89.95±0.52** (+0.27)
	Attack II- $\mathbf{A}_{cos}$	89.09±0.25 (-0.59)
	Attack II- $\mathbf{A}_{fp}$	89.30±0.22 (-0.38)
	Attack III- $\mathbf{E}$	90.25±0.45 (+0.57)
	Attack III- $\mathbf{A}_{cos}$	90.40±0.42 (+0.72)
	Attack III- $\mathbf{A}_{fp}$	**90.65±0.40** (+0.97)

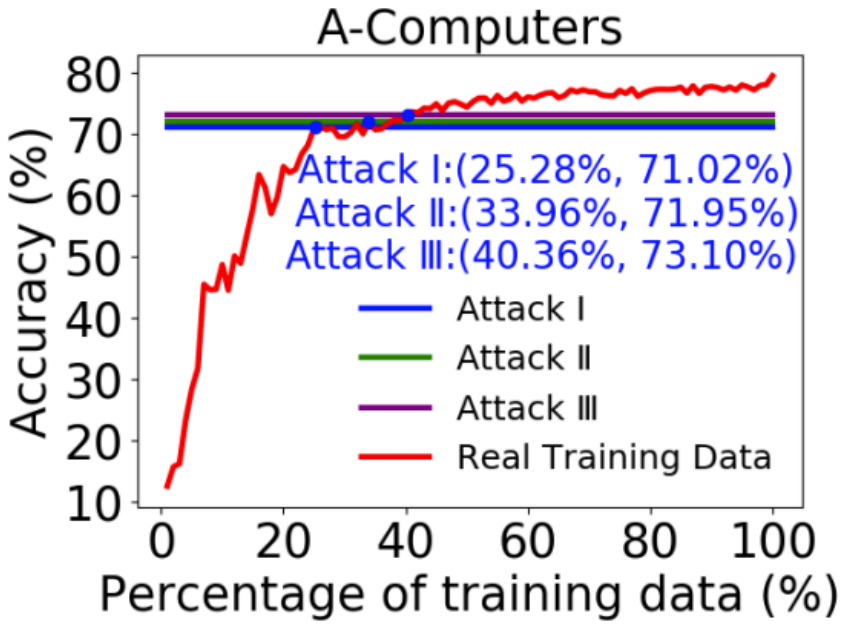
Experiments

☐ Model Analysis

- **Model Comparison**



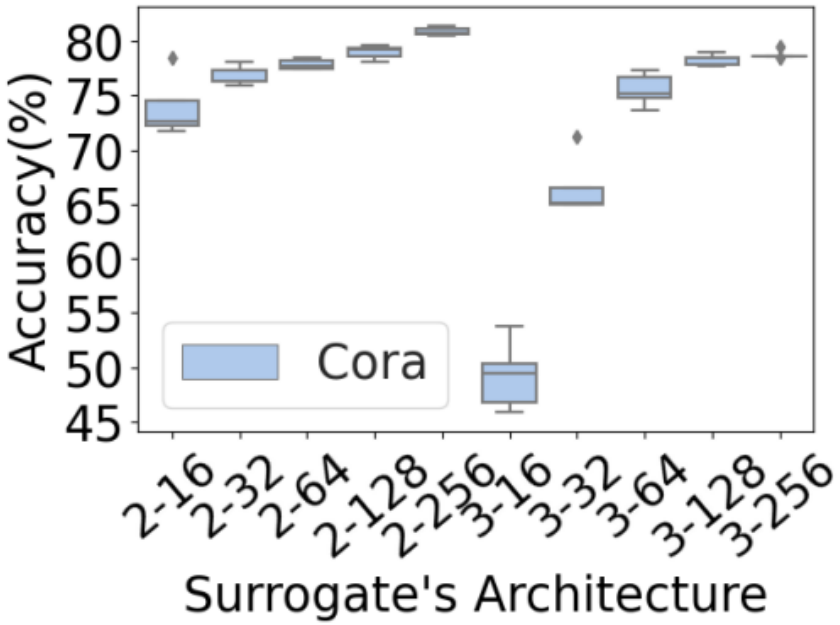
- **Quantitative Analysis**



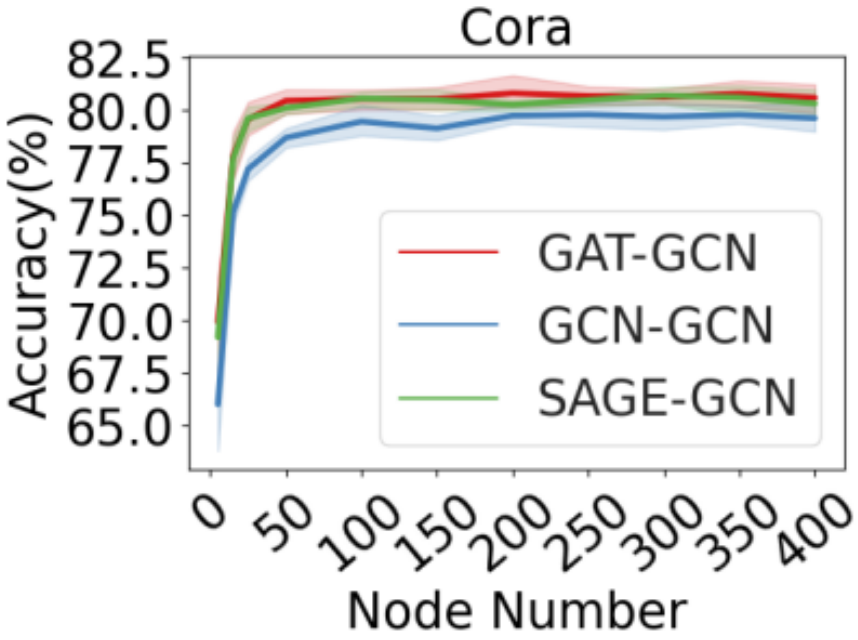
Experiments

☐ Model Analysis

- **Surrogate Architectures**



- **Generated Graph Size**



Conclusion

- **Effectiveness**: StealGNN successfully performs data-free model extraction, replicating GNNs with high accuracy.
- **Innovation**: Introduced the first framework targeting GNNs, showcasing multiple strategies for generator updates
- **Security Implications**: Revealed significant vulnerabilities in GNN models, highlighting the need for stronger defenses.

Yuanxin ZHUANG

yzhuang436@connect.hkust-gz.edu.cn

Thank you !