Debiased Graph Poisoning Attack via Contrastive Surrogate Objective

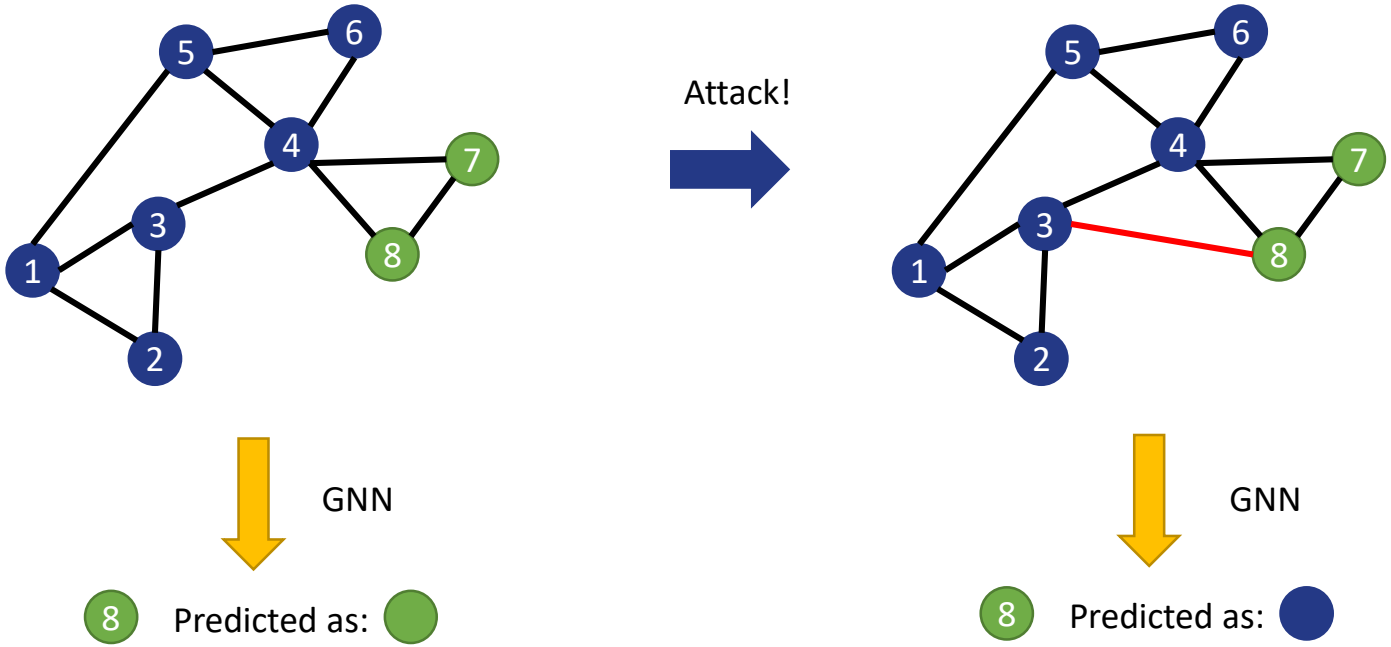
Kanghoon Yoon, Yeonjun In, Namkyeong Lee, Kibum Kim, Chanyoung Park

Department of Industrial & Systems Engineering
KAIST
ykhoon08@kaist.ac.kr

Background

Adversarial Attacks on Graph Structure

- GNNs are vulnerable to adversarial attacks that perturb node features or graph structures, affecting their predictions.



- Recent defense methods have been proposed to make GNNs more robust to such adversarial perturbations

Despite progress in defenses, there has been little focus on thoroughly understanding the methods used for attacking graphs

Background

Meta-Gradient-Based Attack

- Recent graph attacking methods, such as MetaAttack, EpoAtk, GraD, utilize **Meta-Gradient** to perturb the graph structure that effectively degrade the performance of GNNs $\nabla_{\mathbf{A}} \mathcal{L}_{\text{atk}}$

- The optimization problem of attacking graph can be described as :

$$\min_{\tilde{\mathbf{A}}} \mathcal{L}_{\text{atk}}(f_{\theta^*}(\tilde{\mathbf{A}},\mathbf{X}), \hat{\mathbf{Y}}) \quad \text{s.t.} \quad \theta^* = \arg\min_{\theta} \mathcal{L}_{\text{sur}}(f_{\theta}(\tilde{\mathbf{A}}, \mathbf{X}), \mathbf{Y}_L)$$
$$\text{where} \quad \left\|\tilde{\mathbf{A}} - \mathbf{A}\right\|_0 \leq \Delta \qquad (2)$$

  - The attack loss is the negative cross entropy
  - The surrogate loss is the cross entropy

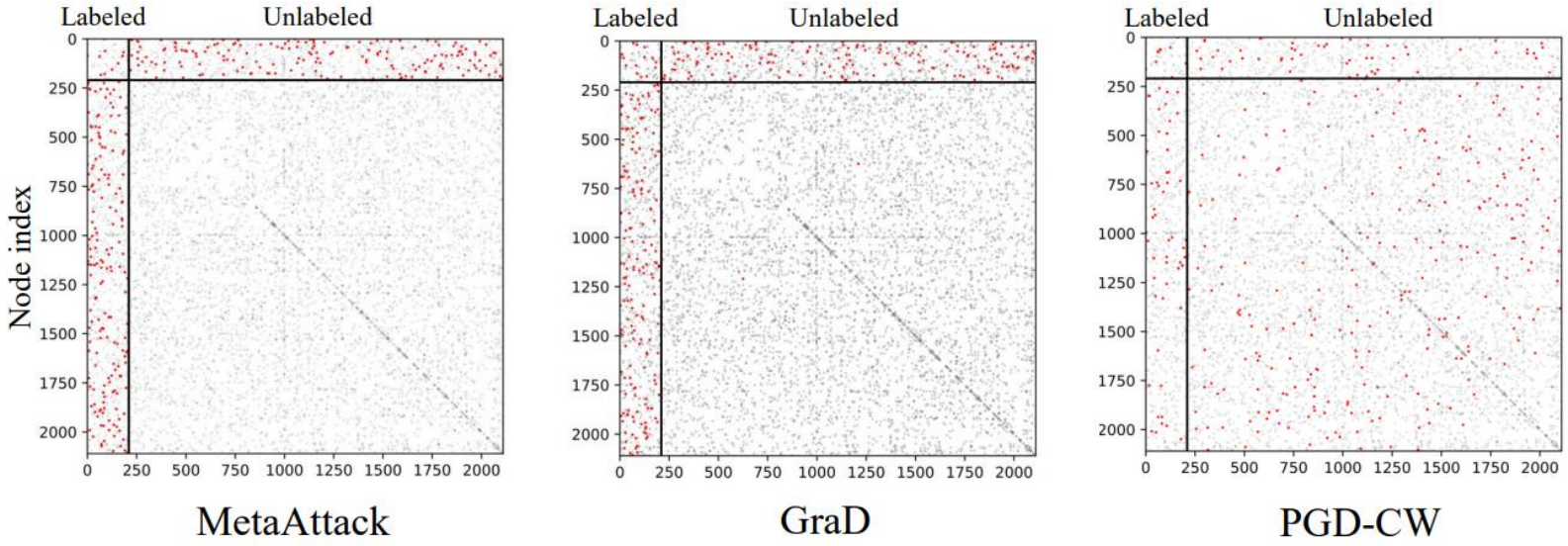
- Based on the meta-gradient matrix, we add the edges with the largest meta-gradient $\nabla_{\mathbf{A}} \mathcal{L}_{\text{atk}}$

Motivation

Meta-Gradient-based Attacks exhibits uneven perturbation between labeled and unlabeled nodes

- **However, we found that existing meta-gradient-based attacks unevenly perturb the graph structure between labeled and unlabeled nodes**

	Model	L-L	L-U	U-U
(a) Num. Attacks	MetaAttack	4.5	177.5	0.5
	EpoAtk	31.5	151.0	0.0
	AtkSE	20.0	162.5	0.0
	GraD	6.0	175.0	0.5
	PGD-CE	4.0	54.0	121.5
	PGD-CW	7.0	55.0	115.5
(b) Attack Ratio (%)	MetaAttack	15.1	29.1	**0.0**
	EpoAtk	93.9	24.8	**0.0**
	AtkSE	64.0	26.6	**0.0**
	GraD	10.2	28.7	**0.0**
	PGD-CE	13.1	8.9	4.0
	PGD-CW	23.9	9.1	3.8



- Existing attack methods consider only Tr-Tr space as a target to attack, which is very small portion of entire space to attack

- Attack performance of existing attack methods is suboptimal as they do not consider the large attack space between unlabeled nodes

Analysis

Investigating Meta-Gradient-based Attacks to find the root cause of uneven perturbation

Unrolling Meta-Gradients

- Recent methods add edge with the largest meta-gradient. We investigate what is the attribute of the attacked edges by unrolling the meta-gradients

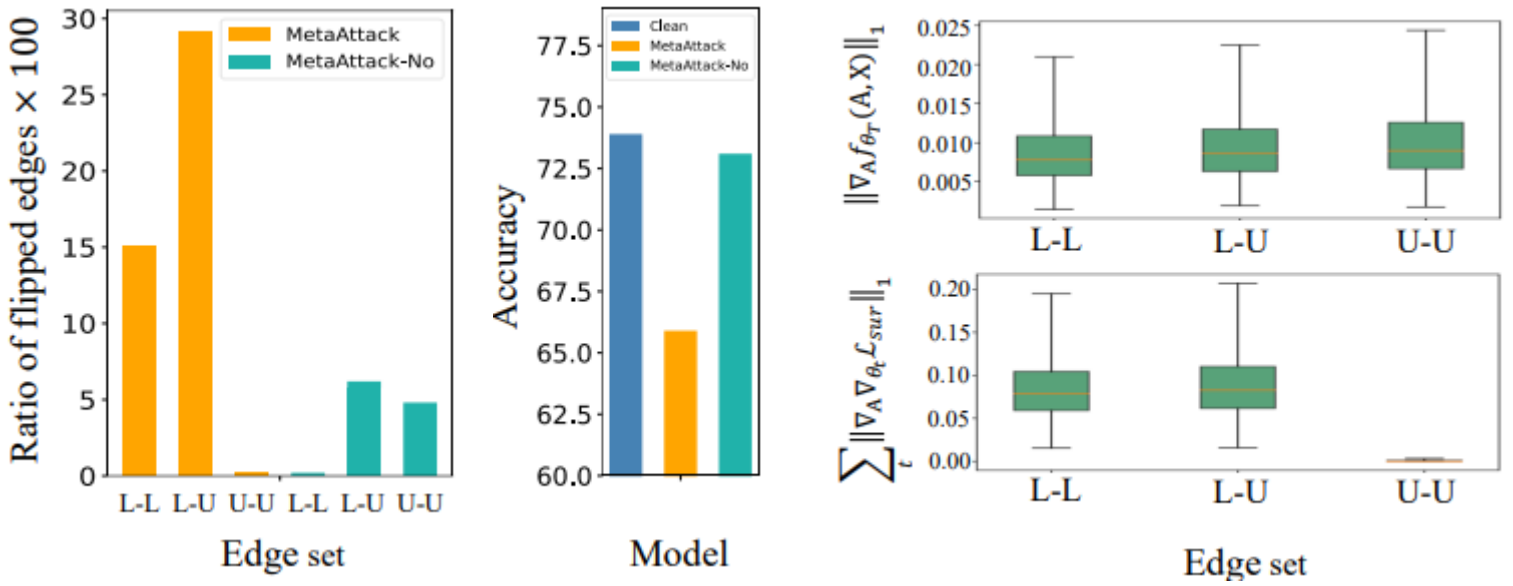
$$\nabla_{\mathbf{A}} \mathcal{L}_{\text{atk}} = \nabla_f \mathcal{L}_{\text{atk}} \cdot \{\nabla_{\mathbf{A}} f_{\theta_T} + \nabla_{\theta_T} f_{\theta_T} \cdot \nabla_{\mathbf{A}} \theta_T\}$$

$$= \nabla_f \mathcal{L}_{\text{atk}} \cdot \{\underbrace{\nabla_{\mathbf{A}} f_{\theta_T}}_{\text{(i)}} + \nabla_{\theta_T} f_{\theta_T} \cdot \underbrace{(\nabla_{\mathbf{A}} \theta_0 - \alpha \sum_{t=1}^{T} \nabla_{\mathbf{A}} \nabla_{\theta_t} \mathcal{L}_{\text{sur}})}_{\text{(ii)}}\}$$

- The first term (i) means how much the model prediction logit changes when we perturb the input graph.
- The second term (ii) means how much the sum of the gradient of loss w.r.t model parameters, given perturbed graph.
  - This term implies that the second term include the surrogate model's training procedure

Analysis

Investigating Meta-Gradient-based Attacks to find the root cause of uneven perturbation



Empirical Study

- First, we remove the training procedure of the surrogate model, to confirm the effect of the second term (training procedure)
- Second, we visualize the first term and the second term in the meta-gradient

Proposed Method

Meta-Gradient-Based Attack via Contrastive Surrogate Loss

We need to design a new attack method with these fact:

- The utilization of meta-gradients is crucial for achieving strong attack performance
- We need to alleviate the inherent bias towards labeled nodes, which results in suboptimal attack performance,

➔ We propose **new surrogate loss for meta-gradient-based attacks** to expand the attack search space to a broader range of edge sets, not just limited to L-L and L-U edges, but all of edges

Challenges.

To mitigate the bias present in meta-gradient-based attacks, **it becomes crucial to incorporate the influence of both labeled and unlabeled nodes in the surrogate model's training procedure.**

However, **simply incorporating the unlabeled nodes in the surrogate loss falls short of generating effective attacks if the goal of the loss (e.g,., link reconstruction) does not align with that of the victim GNNs (e.g., the node classification)**

Proposed Method

Meta-Gradient-Based Attack via Contrastive Surrogate Loss

Metacon-S (with Sample Contrastive Surrogate Loss)

- The sample contrastive surrogate loss is computed on testing nodes

$$\mathcal{L}_{\text{s-con}} = \sum_{u \in \mathcal{V}_U} l_{\text{s-con}}(u)$$

$$l_{\text{s-con}}(u) = -\log \frac{e^{s(p_u, \hat{p}_u)}}{e^{s(p_u, \hat{p}_u)} + \sum_{k=1}^{n} \left[\mathbb{1}_{[k \neq u]} e^{s(p_u, \hat{p}_k)} + \mathbb{1}_{[k \neq u]} e^{s(\hat{p}_u, \hat{p}_k)} \right]}$$

THEOREM 5.1. *Assume that a surrogate model $f_\theta(\cdot)$ consists of multiple GCN layers without non-linear activation, and the mean of the probabilities obtained from p_u and $\hat{p}_u$ is the same. When the node classes are balanced, $\mathcal{L}_{s\text{-}con}$ is the upper bound of the cross entropy.*

- The goal of the sample contrastive loss align with the cross entropy loss as it is the upper bound of the cross entropy

Proposed Method

Meta-Gradient-Based Attack via Contrastive Surrogate Loss

Metacon-D (with Dimension Contrastive Surrogate Loss)

- The sample contrastive loss requires quadratic computation on the number of nodes.
- To scale the method, we also propose the dimension contrastive surrogate loss is computed on testing nodes

$$\mathcal{L}_{\text{d-con}} = m(\mathbf{P}, \hat{\mathbf{P}}) + \mu_1 (v(\mathbf{P}) + v(\hat{\mathbf{P}})) + \mu_2 (c(\mathbf{P}) + c(\hat{\mathbf{P}}))$$

$$m(\mathbf{P}, \hat{\mathbf{P}}) = \frac{1}{n} \sum_{u \in \mathcal{V}} ||p_u - \hat{p}_u||_2^2$$

$$v(\mathbf{P}) = \frac{1}{K} \sum_{j=1}^{K} \sqrt{\max\{0, \gamma - Var(p^j)\}}$$

$$c(\mathbf{P}) = \frac{1}{K} \sum_{i \neq j} [C(\mathbf{P})]_{i,j}^2$$

$$C(\mathbf{P}) = \frac{1}{n-1} \sum_{u \in \mathcal{V}} (p_u - \bar{p})(p_u - \bar{p})^T.$$

- We propose the theorem about the goal alignment of the dimension contrastive loss with that of the original surrogate loss in the paper too.

Experiment

Experimental settings and datasets

Dataset

- Citation Network, Co-purchase Network, Social Network

Evaluation Protocol

- Training Time Attack
- Untargeted Attack
- Node Classification Task

Baselines

- Random attack
- PGD attack
- Meta-gradient-based attack
- Meta-gradient attack for self-supervised learning models

Table 2: Statistics of datasets.

Dataset	# Nodes	# Edges	# Features	# Classes
Cora	2,485	5,069	1,433	7
Cora ML	2,810	7,981	2,879	7
Citeseer	2,110	3,668	3,703	6
Polblogs	1,222	16,714	0	2
Am. Photo	7,650	119,081	745	8
Am. Computers	13,752	245,861	767	10
Reddit	231,443	11,606,919	602	41

Experiment

Node classification on adversarial attack

Table 3: Node classification accuracy under the strong transfer scenario. 5% of edges are flipped on Cora, Citeseer, Polblogs and Cora ML datasets. Bold represents the best performance. Underline represents the second place.

Dataset		Cora	Citeseer	Polblogs	Cora ML
Clean		83.6±0.3	73.9±0.5	95.0±0.8	85.3±1.1
Rand	Random	82.7±0.2	73.3±0.8	91.6±1.2	84.0±1.2
	DICE	82.3±0.6	73.1±1.0	89.7±0.3	84.2±0.9
Self	BBGA	82.7±0.5	73.4±1.2	87.7±0.4	84.9±0.7
	CLGA	81.2±0.3	72.4±1.1	88.2±1.4	84.8±0.7
PGD	PGD-CE	83.7±0.6	73.3±0.7	83.5±0.6	85.1±0.6
	PGD-CW	80.6±0.7	70.9±0.8	78.2±1.6	81.7±0.9
Meta	MetaAttack	76.9±0.6	65.9±1.3	76.6±0.5	76.4±1.3
	EpoAtk	82.9±0.3	73.0±1.4	94.4±0.5	84.6±1.0
	AtkSE	79.5±2.3	72.0±0.9	78.7±1.1	80.6±1.4
	GraD	76.8±2.4	66.4±2.0	**75.1±0.9**	76.1±1.1
	Metacon-S	75.4±1.5	64.1±0.7	**75.1±0.5**	76.0±1.2
	Metacon-D	**75.3±1.1**	**63.9±0.8**	75.2±0.6	**75.7±0.8**

Table 4: Node classification accuracy under the weak transfer scenario. 5% of edges are flipped on Cora, Citeseer, Polblogs and Cora ML datasets. Bold represents the best performance. Underline represents the second place.

Dataset		Cora		Citeseer		Polblogs		Cora ML	
Victim Model		GraphSAGE	GAT	GraphSAGE	GAT	GraphSAGE	GAT	GraphSAGE	GAT
Clean		81.6±1.6	83.8±0.6	72.7±1.1	73.5±0.8	95.1±1.0	94.9±0.3	84.5±1.0	85.2±0.8
Self	BBGA	82.3±0.6	82.2±0.8	72.4±0.9	73.7±0.6	93.3±0.7	92.1±0.9	84.4±0.4	84.3±0.7
	CLGA	80.7±0.8	81.3±0.5	71.8±0.4	72.6±0.6	92.6±1.1	90.0±1.3	82.6±0.6	82.5±0.9
PGD	PGD-CE	82.7±0.8	83.8±0.5	73.3±1.0	74.1±0.8	**85.6±0.4**	83.5±1.0	85.1±0.4	85.5±0.8
	PGD-CW	80.5±0.7	80.7±0.9	71.4±0.9	70.9±1.0	87.1±0.8	**79.8±1.2**	82.6±0.5	81.5±1.0
Meta	MetaAttack	78.2±1.1	78.7±0.9	69.4±1.0	69.1±0.0	86.6±1.8	80.9±0.9	80.9±1.2	79.8±1.2
	EpoAtk	81.6±1.2	82.7±0.4	72.4±0.5	73.4±0.6	94.5±0.6	94.5±0.3	84.2±0.3	84.3±0.8
	AtkSE	80.5±1.1	81.9±1.2	72.4±1.3	73.9±0.8	91.7±1.4	88.8±3.0	82.4±1.2	81.5±1.4
	GraD	78.5±0.8	79.1±1.0	69.3±1.8	69.2±1.0	86.9±0.5	81.2±0.7	80.7±1.1	**78.6±1.4**
	Metacon-S	**77.3±1.0**	**77.0±1.1**	**66.4±2.1**	68.0±0.9	86.8±1.3	81.5±1.4	80.4±1.0	78.8±1.4
	Metacon-D	77.9±0.7	**77.0±1.0**	66.6±1.9	**66.8±1.2**	86.9±1.3	82.1±0.8	**80.2±0.5**	78.9±0.9

Experiment

Node classification on adversarial attack

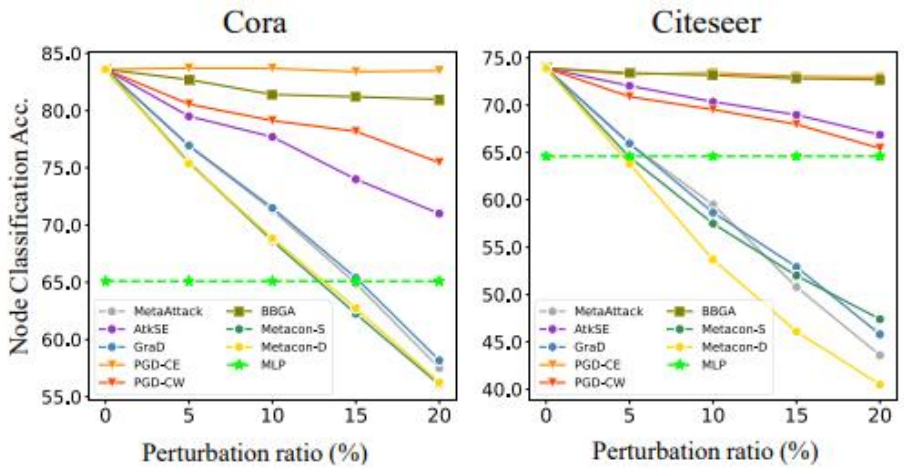


Figure 3: Node classification accuracy over perturbation ratios under the strong transfer scenario.

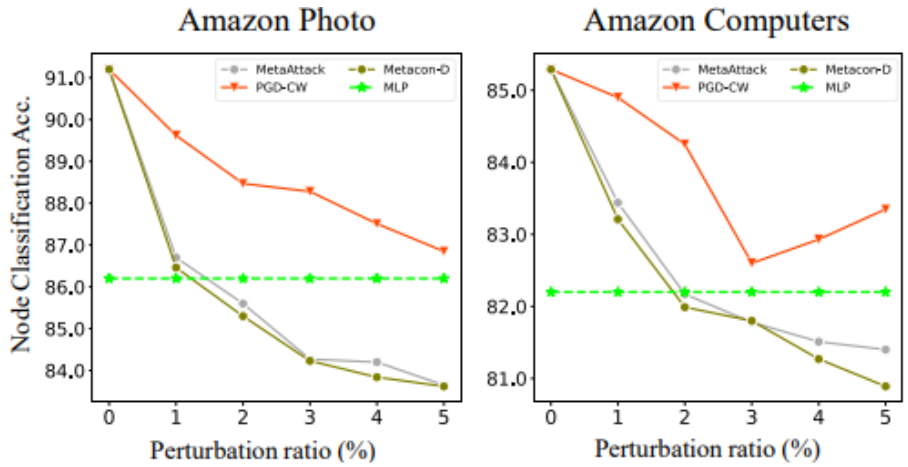


Figure 4: Node classification accuracy over perturbation ratios on large-scale networks under the strong transfer scenario. Amazon Photo and Computers datasets are used.

Experiment

Node classification on adversarial attack

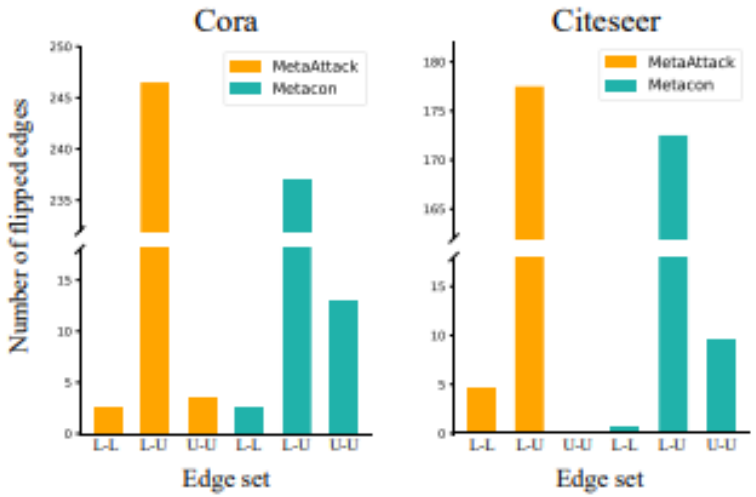


Figure 5: Number of flipped edges on Cora and Citeseer datasets when MetaAttack and Metacon-S are applied.

Table 8: Analysis on the complexity of the attack methods. The memories required for the attack models are reported.

Dataset	Cora	Citeseer
# Nodes	2,485	2,110
# Edges	5,069	3,668
# Feats	1,433	3,703
MetaAttack	5.5 GB	5.8 GB
Metacon-S	22.4 GB	23.6 GB
Metacon-D	6.3 GB	6.5 GB

Conclusion

- We found an unique phenomenon of the graph attack methods, which unevenly perturbs the graph structure between labeled nodes and unlabeled nodes.

- We investigate the root cause of the uneven perturbation that the training procedure of the surrogate model incur the uneven perturbation

- We propose the new surrogate loss for existing attack methods, the sample-contrastive surrogate objective and the dimension-contrastive objective. They address the bias towards labeled nodes and achieve the state-of-the-art attack performance on several benchmark datasets

Thank you for listening!