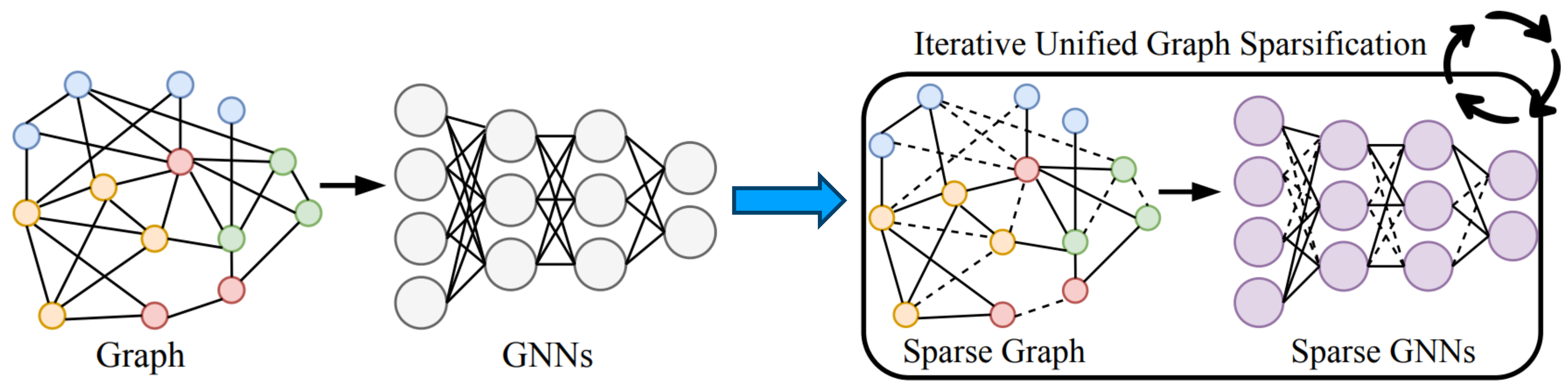
Sparse but Strong: Crafting Adversarially Robust Graph Lottery Tickets

Subhajit Dutta Chowdhury[1], Zhiyu Ni[1], Qingyuan Peng[1], Souvik Kundu[2], Pierluigi Nuzzo[1]
[1]University of Southern California, Los Angeles, USA
[2]Intel Labs, San Diego, USA

Problem Statement

Are the existing graph lottery ticket (GLT) identification techniques capable of generating adversarially robust GLTs?

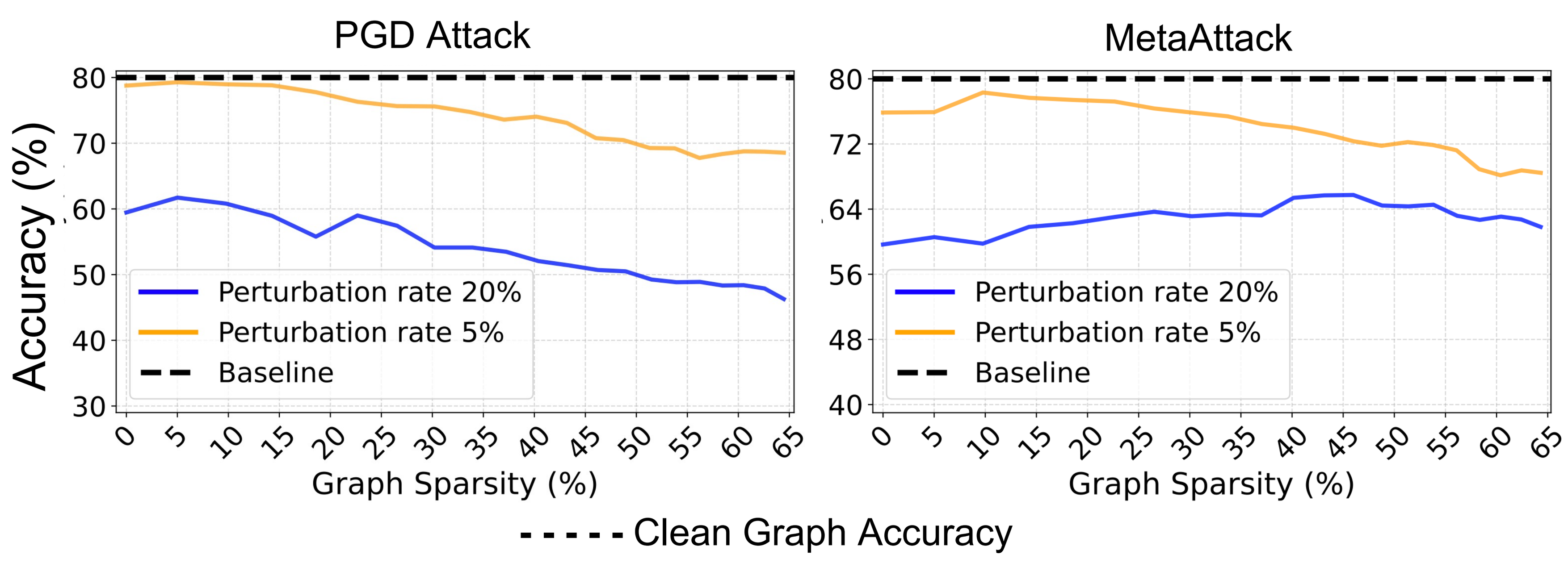
Graph Lottery Tickets in a Nutshell

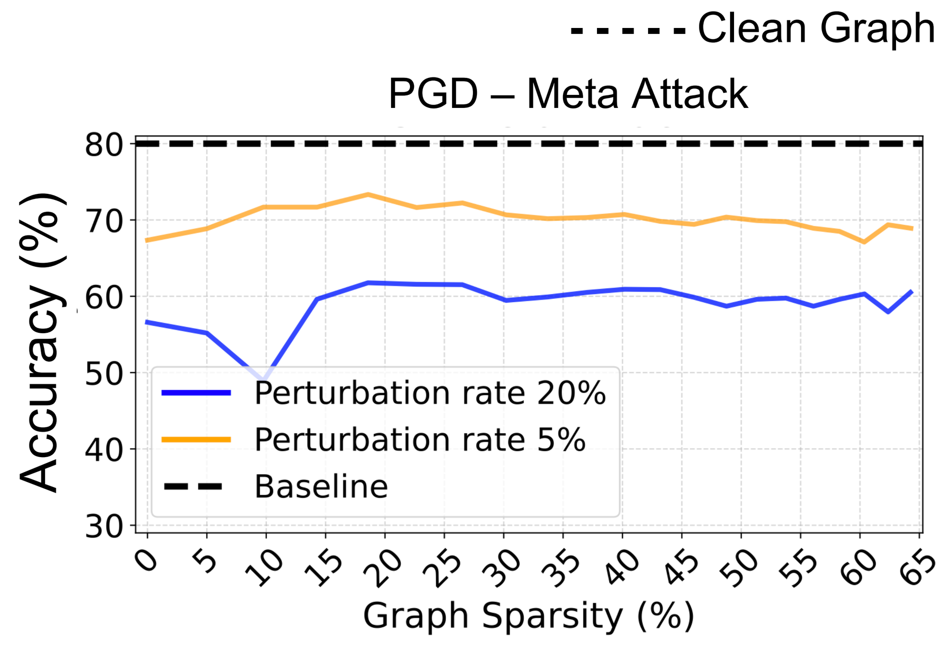


Unified graph sparsification (UGS) iteratively removes edges and weights from the graph and GNN, respectively.
Image courtesy: T. Chen et al., "A Unified Lottery Tickets Hypothesis for Graph Neural Networks," ICML 2021

- A graph lottery ticket (GLT) consists of:
  - A **sparse** graph neural network (GNN)
  - A **sparse** input graph adjacency matrix
- GLTs substantially **reduce** the **inference compute footprint** compared to their dense counterparts while achieving similar performance

Analyzing the Robustness of GLTs to Adversarial Structure Perturbations





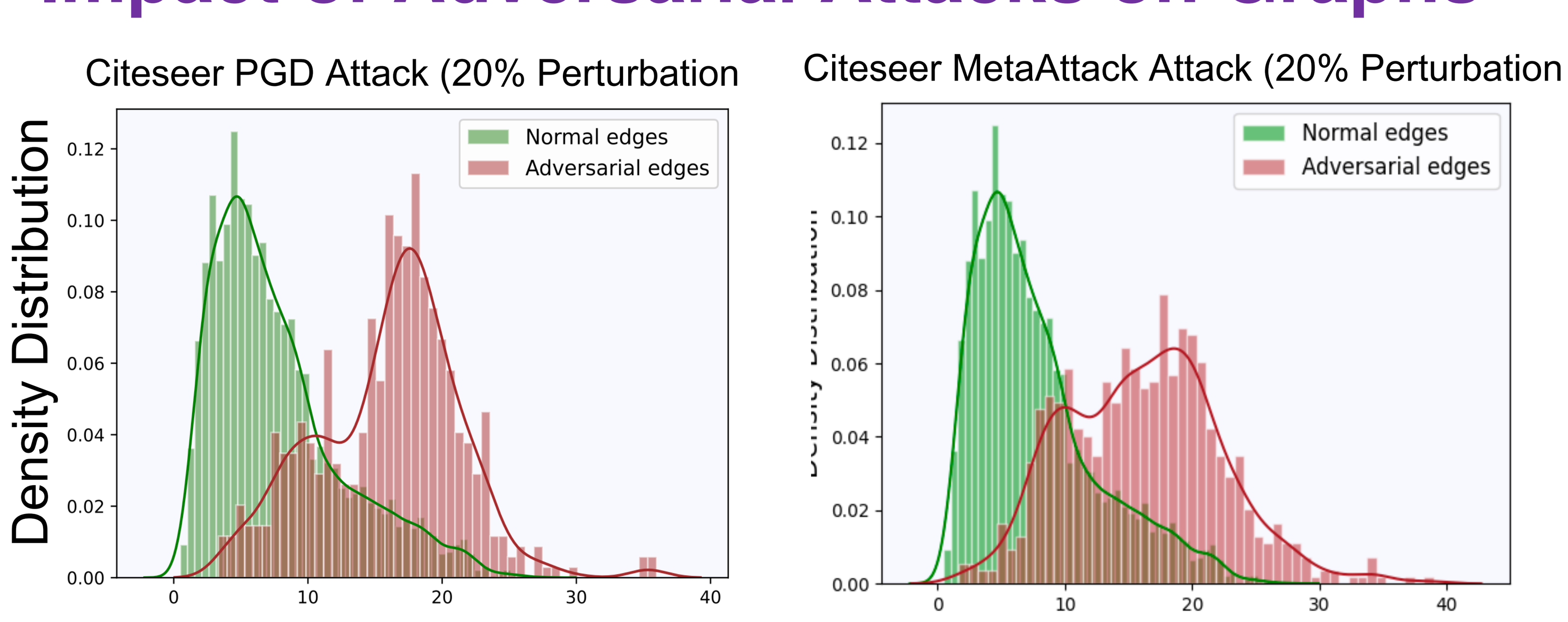
Dataset:
- Cora

Attacks:
- PGD
- MetaAttack
- Meta-PGD Attack

Perturbation Rates:
- 5%, 10%, 15%, and 20%

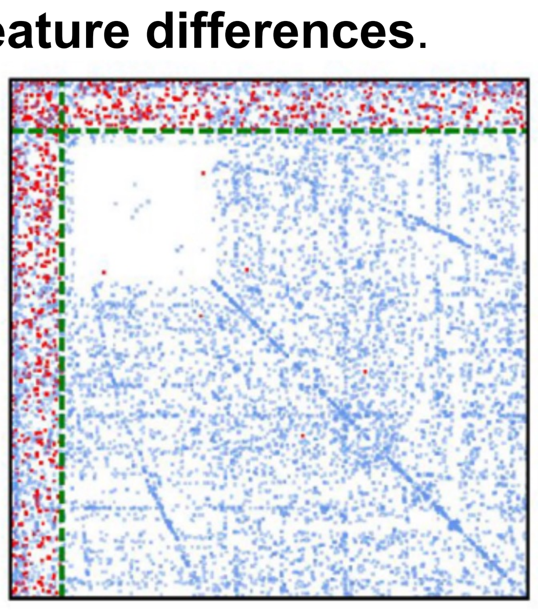
While UGS removes edges from the perturbed adjacency matrix, it **does not** effectively **remove** the **adversarial** edges.

Impact of Adversarial Attacks on Graphs



Attribute feature differences between connected nodes in perturbed Citeseer graph dataset.

- Attacks tend to connect nodes with significant **attribute feature differences**.



Adjacency matrix of Cora attacked by MetaAttack (10%). Blue dots - Clean edges, Red dots – Adversarial edges, Green dotted line - Boundary of train and test nodes.

- Edge perturbations are unevenly distributed on the graph
- Most modifications are around train nodes

Adversarially Robust Graph Sparsification

The proposed technique performs unified graph–GNN sparsification such that:
- Promotes feature smoothness
- Leverages self-training to remove edges from the locality of the test nodes

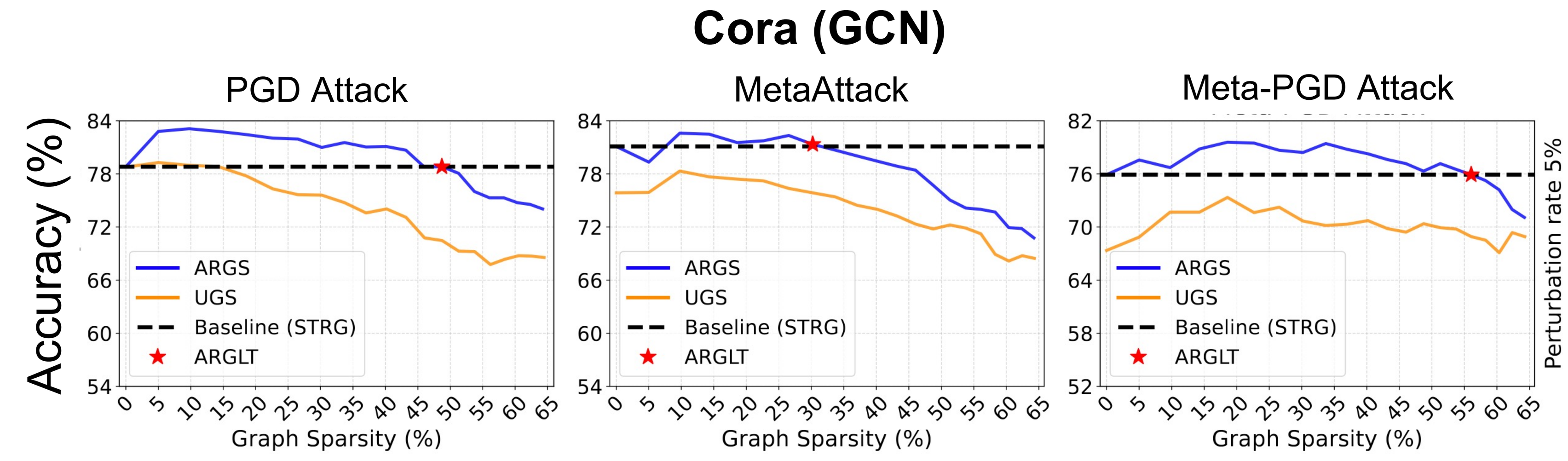
CE loss for train nodes — Feature smoothness

$$\mathcal{L}_{ARGS} = \mathcal{L}_0(f(\{\boldsymbol{m_g} \odot \boldsymbol{A}', \boldsymbol{X}\}, \boldsymbol{m_\theta} \odot \boldsymbol{\Theta})) + \beta \mathcal{L}_{fs}(\boldsymbol{m_g} \odot \boldsymbol{A}', \boldsymbol{X})$$
$$+ \mathcal{L}_1(f(\{\boldsymbol{m_g} \odot \boldsymbol{A}', \boldsymbol{X}\}, \boldsymbol{m_\theta} \odot \boldsymbol{\Theta})) + \lambda_1 ||\boldsymbol{m_g}||_1 + \lambda_2 ||\boldsymbol{m_\theta}||_1$$

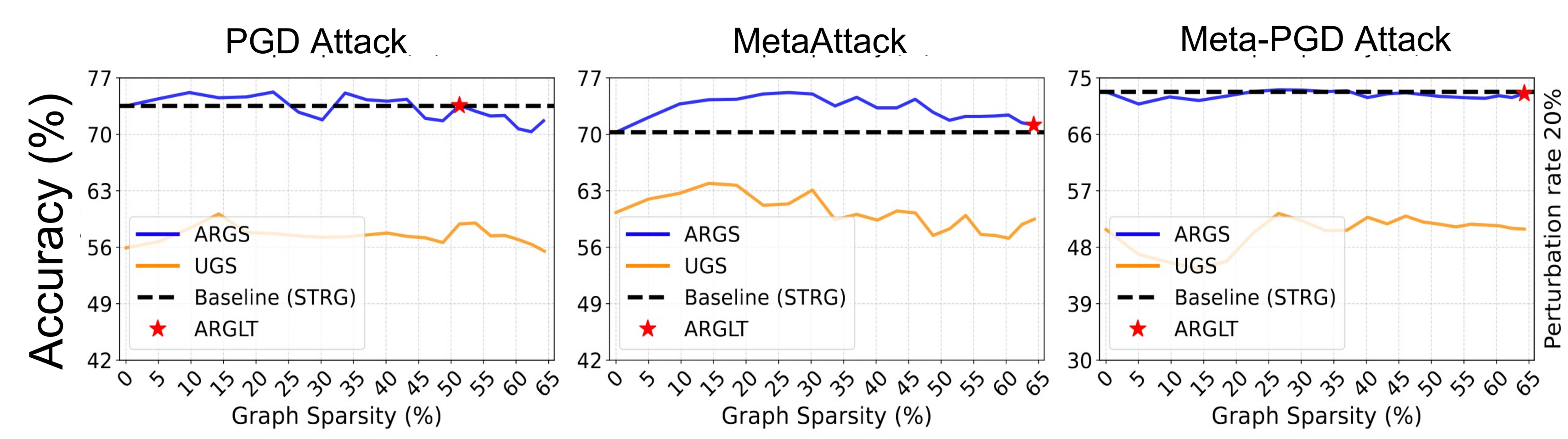
CE loss for test nodes — l_1 regularizers for the masks

Experimental Results

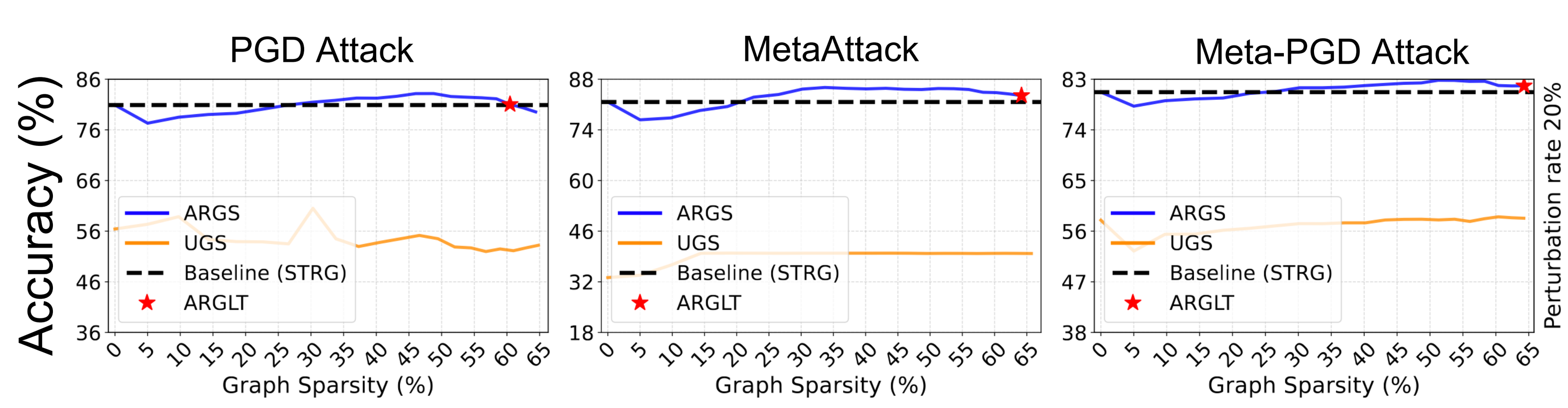
Cora (GCN)



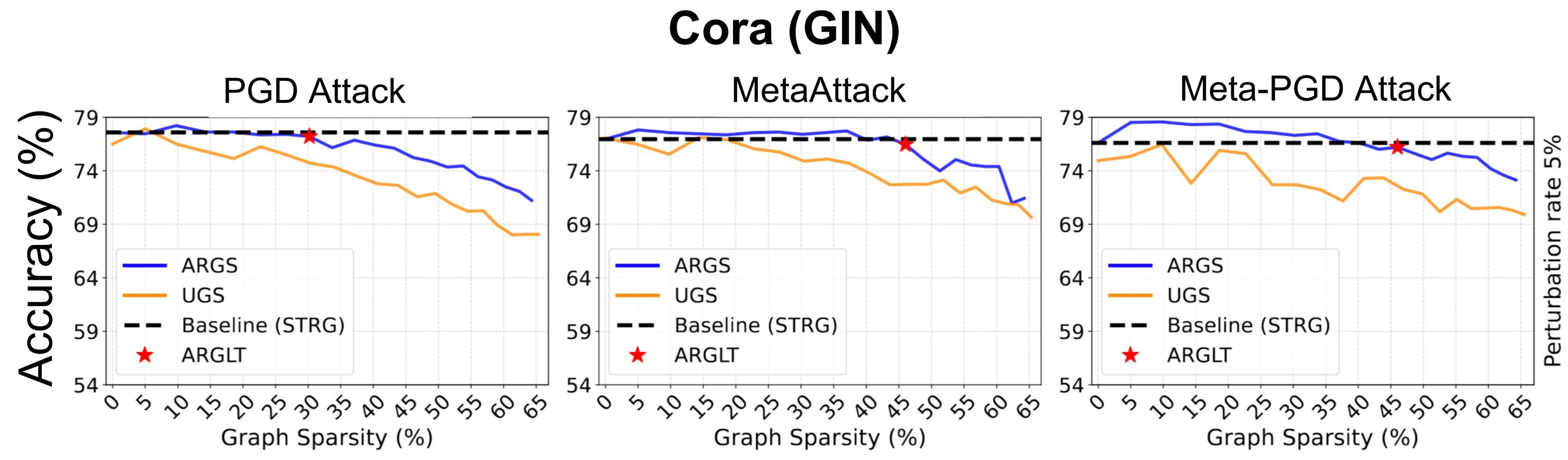
Citeseer (GCN)



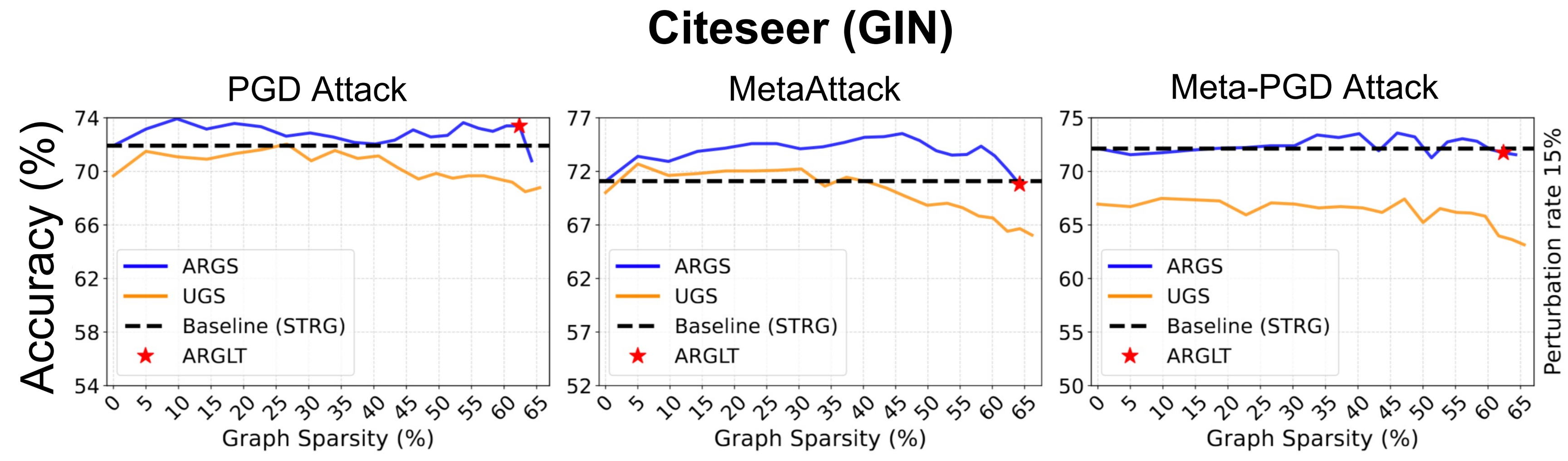
PubMed (GCN)



Cora (GIN)



Citeseer (GIN)



Summary

- Existing techniques for GLT identification are unable to generate adversarially robust GLTs
- The proposed technique ARGS can remove adversarial edges effectively and generate adversarially robust GLTs
- Future work includes extending ARGS to heterophilic graphs

Email contact — Arxiv pre-print

NeurIPS 2023 New Frontiers In Graph Learning (GLFrontiers) Workshop