CEGA: A Cost-Effective Approach for Graph-Based Model Extraction and Acquisition

Zebin Wang, Menghan Lin, Bolin Shen, Ken Anderson, Molei Liu, Tianxi Cai, Yushun Dong

June 15, 2025

Email: zebinwang@g.harvard.edu

HARVARD UNIVERSITY

FSU | FLORIDA STATE UNIVERSITY

ICML International Conference On Machine Learning

Backgrounds and Research Questions

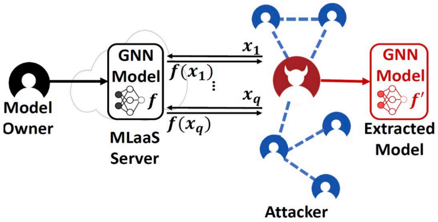
- Graph Neural Networks (GNNs) have many key applications but are vulnerable to malicious model extraction
- Research-driven acquisition of GNN functionality has high potential
- How to depict **structural dependency** between nodes in the graph?
- How to overcome **budget** and **query batch size** constraints?

Introduction to CEGA, A GNN Extraction Strategy

- CEGA: Cost-Effective Graph Acquisition
- CEGA incorporates **historical information** from the initial and previous queries to guide further node selection for querying
- CEGA integrates three key criteria:
  - Nodes' representativeness to the graph structure
  - Nodes' uncertainty on classification based on interim model
  - Nodes' diversity based on distance to queried ones
- CEGA **dynamically** weighs each criterion, as uncertainty and diversity are progressively emphasized in later cycles, resonating with the improved performance of the interim model trained with more queries

Contribution to MEA and Acquisition in Research

- CEGA shows that high-fidelity extraction on graph models is feasible, **even under stringent query budget constraints**
- CEGA alerts the maintainers of proprietary GNNs against MEA and inspires the development of more robust defense mechanisms
- CEGA highlights the potential for ethical, resource-efficient GNN extraction to support researchers with a limited budget



Credit to: Wu, B., Yang, X., Pan, S., and Yuan, X. Model extraction attacks on graph neural networks: Taxonomy and realisation. ASIA CCS '22, pp. 337-350, 2022

Experiment Results of CEGA

- CEGA **consistently outperforms** state-of-the-art active learning (AL) techniques across a **wide range of datasets**, particularly in terms of **fidelity** to the model targeted for extraction

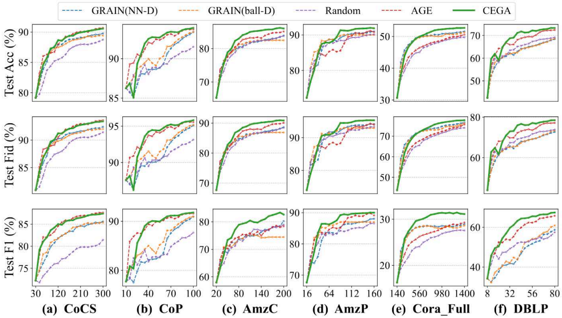


Figure 1. The trajectory of test accuracy, fidelity, and F1 score on different datasets using $2C$ to $20C$ queried nodes. The performance trajectory of CEGA is bolded in green, showing significant superiority over the alternatives across different number of queried nodes.

Ablation Study for CEGA and Future Plans

- Ablation studies demonstrate that representativeness and uncertainty are essential for performance, while diversity controls the variance across tests
- Theoretical analysis shows that the time and space complexity of CEGA's node selection strategy **has a lower order** than that of training the interim model

	CEGA	**No Cen**	**No UnC**	**No Div**
CoCS	$\mathbf{93.4 \pm 0.6}$	93.2 ± 0.2	91.9 ± 0.5	93.4 ± 0.6
CoP	$\mathbf{95.8 \pm 0.5}$	94.9 ± 0.4	90.2 ± 3.3	95.7 ± 0.5
AmzC	$\mathbf{90.8 \pm 0.4}$	90.0 ± 1.2	87.1 ± 2.2	90.7 ± 0.7
AmzP	$\mathbf{95.3 \pm 0.5}$	95.1 ± 0.3	93.7 ± 0.9	95.3 ± 0.7
Cora_Full	77.9 ± 0.9	75.3 ± 0.6	74.9 ± 0.9	$\mathbf{78.3 \pm 1.1}$
DBLP	78.5 ± 0.9	74.2 ± 2.4	65.1 ± 5.5	$\mathbf{78.6 \pm 1.4}$

- Future Directions of Research:
  - Extend CEGA from a *transductive* setting to an *inductive* setting
  - Leverage *edge information* in training interim models

Thank you!