The Confidence Trap: Calibration Attacks for Graph Neural Networks

Cuong Dang[1], Jiahao Zhang[2], Hieu Ta Quang[3], Dung Le[3], Lu Cheng[4], Suhang Wang[2]

Virginia Tech[1] • The Pennsylvania State University[2] • VinUniversity[3] • University of Illinois Chicago[4]

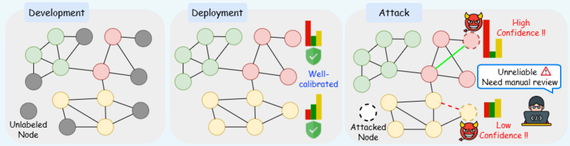
KDD 2026 Jeju, Korea August 9–13, 2026

Abstract

Confidence calibration is essential for trustworthy decision-making in safety-critical applications, yet the robustness of calibrated Graph Neural Networks to adversarial structural perturbations remains largely unexplored. We introduce the Unified Graph Calibration Attack (UGCA), a white-box framework that manipulates confidence scores while preserving predicted labels. UGCA combines a KL-to-uniform loss, reranking, a hybrid label-recovery loss, and beam search to efficiently explore the discrete graph space. Our theoretical and empirical results show that UGCA substantially increases Expected Calibration Error while preserving classification accuracy, revealing significant robustness gaps in existing GNN calibration methods.

Problem Setup

Calibration attacks differ from ordinary adversarial attacks. Instead of changing the predicted class, the attacker perturbs the graph structure to manipulate prediction confidence while preserving the original predicted label.



Why Naive Attacks Fail?

Ineffective UCA Objective

The original underconfidence objective does not sufficiently push predictions toward a uniform distribution.

Sensitivity to Edge Perturbations

GNNs are highly sensitive to edge perturbations, which can easily cause label flips.

Local Optima

Greedy search may terminate early or become trapped in a suboptimal adversarial graph.

Theory Insight & Contributions

Overconfidence Attack (OCA)

$$ECE_{\text{OCA}} = \frac{1}{K}[(K-2)\text{acc}(\mathcal{D}) + 1]$$

Underconfidence Attack (UCA)

$$ECE_{\text{UCA}} = \text{acc}(\mathcal{D}) - \frac{1}{K}$$

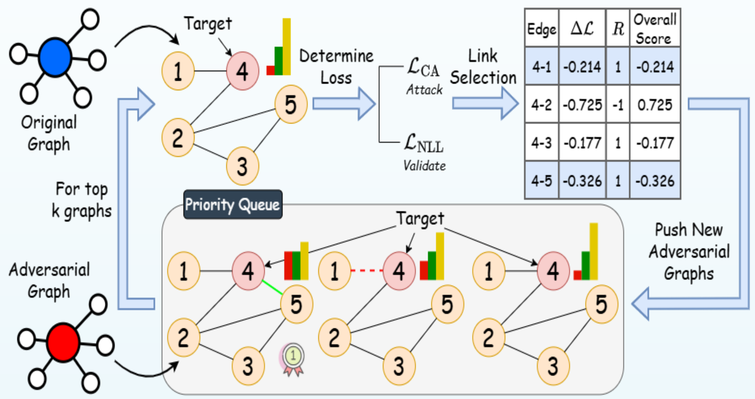
Higher classification accuracy and a larger number of classes imply greater calibration vulnerability under the analyzed threat model.

Our Contributions

- First dedicated framework for calibration attacks on GNNs.
- Formulates graph calibration attack as a constrained discrete optimization problem.
- Hybrid objective with reranking and beam search.
- Empirical analysis across datasets, budgets, and methods.
- Insights into robustness and calibration-data regimes.

Methodology / UGCA Framework

UGCA formulates graph calibration attack as a constrained discrete optimization problem. The framework iteratively searches for graph perturbations that maximize calibration degradation while preserving the original predicted label.



UGCA maintains a priority queue of promising adversarial graphs, evaluates label-aware calibration losses, reranks candidate edges, and uses beam search to explore multiple perturbation paths.

UGCA Components

KL-to-uniform loss
Pushes the predictive distribution toward uniformity for stronger underconfidence attacks.

Reranking
Downweights edge perturbations that are likely to flip the predicted label.

Hybrid loss
Switches to label recovery when the label-preservation constraint is violated.

Beam search
Explores multiple adversarial candidates to reduce local-optimum issues.

Optimization Objective

UGCA solves a constrained discrete optimization problem

$$\mathcal{L}_{\text{total}} = \beta\,\mathcal{L} + (1-\beta)\,\mathcal{L}_{\text{NLL}}$$

- $\mathcal{L}$ is the calibration attack objective: KL-to-uniform for UCA or the corresponding overconfidence objective for OCA.
- $\mathcal{L}_{\text{NLL}}$ is the negative log-likelihood objective used to restore the original predicted label when a constraint violation occurs.
- $\beta \in [0,1]$ controls the trade-off between calibration degradation and label recovery.

Constraint: The predicted label must remain unchanged after the attack.

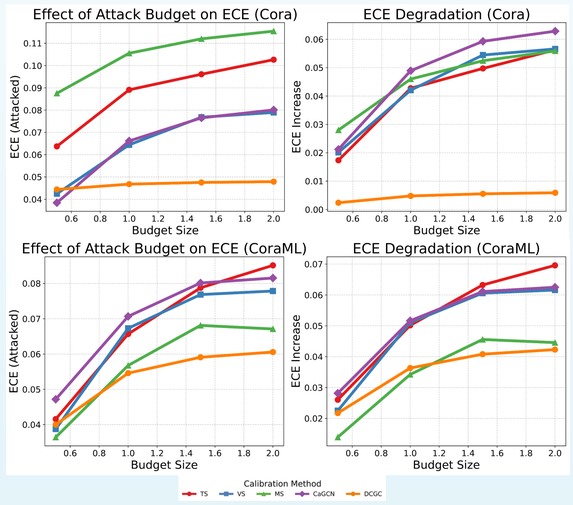
Results Overview

100/100	-0.3643	9.54 s / 100 Nodes
Attack Success Rate (ASR)	Avg. Confidence Reduction	Attack Runtime

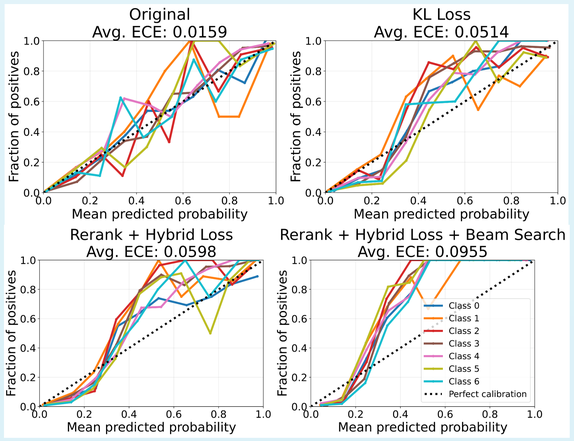
Performance Summary (Cora)

Method	Avg. Runtime (s)	ASR	Avg. Δ Confidence
Random	2.10	100/100	-0.237
IGA	15921	0/100	0
FGA	4.91	76/100	-0.2853
UGCA (ours)	**9.54**	**100/100**	**-0.3643**

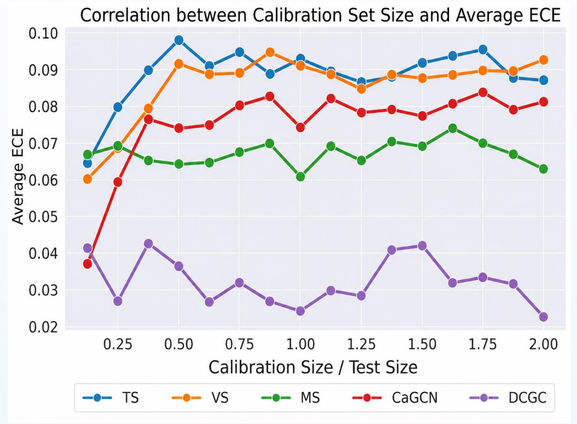
Robustness Under Attack Budget



Component Effect on Cora (ECE)



Correlation between Calibration Set Size and Average ECE



Conclusion / Takeaways

- UGCA is the first dedicated framework for calibration attacks on Graph Neural Networks.
- GNN confidence can be severely manipulated without changing the predicted label.
- Improving predictive accuracy alone is insufficient for trustworthy deployment.
- Graph-aware and data-centric calibration strategies show greater robustness under adversarial structural perturbations.
- Trustworthy GNN systems require robust confidence estimates, not only accurate class predictions.

Accurate GNNs are NOT necessarily trustworthy.

Acknowledgements

Wang is supported by the Army Research Office (ARO) under Grant No. W911NF-2110198 and a Cisco Faculty Research Award. Cheng is supported by National Science Foundation (NSF) Grants #2312862, CAREER #2440542, and #2533996, the NSF-Simons SkAI Institute, National Institutes of Health (NIH) Grant #R01AG091762, NSF ACCESS Computing Resources, a Google Research Scholar Award, and a Cisco gift grant. Dung and Hieu are supported by Grant VUNI.2526.CAIR.03, Center for AI Research, VinUniversity.