Cross-Domain Adversarial Training for Robust Message-Passing GNNs and Diffusion Models

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Does cross-domain adversarial training across graph datasets (e.g., citation networks to social networks) improve the robustness of message-passing GNNs and diffusion models, as measured by. 17 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adversarial Attack and Defense on Graph Data: A Survey. Research question: Does cross-domain adversarial training across graph datasets (e.g., citation networks to social networks) improve the robustness of message-passing GNNs and diffusion models, as measured by classification accuracy on unseen adversarial perturbations?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.7/10.

3 Results

14 papers retrieved. 17 claims extracted; 0 independently verified. Quality review score: 2.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Most existing adversarial attack works on graph data are poisoning attacks performed in the transductive learning settin	×	0.15
In poisoning attacks, the model is retrained after the attacker changes the data.	×	0.04
In evasion attacks, the parameters of the trained model are assumed to be fixed.	×	0.05
Evasion attacks only change the testing data and do not require retraining the model.	×	0.04
The goal of an integrity attack is to reduce the performance of target instances while keeping the total system performa	×	0.03
Availability attacks are easier to detect than integrity attacks under the poisoning attack setting.	×	0.04
Meaningful availability attack studies are generally conducted under the evasion attack setting.	×	0.04
Node embedding tasks use low dimensional representations of each node for adversarial attacks.	×	0.06
Link prediction tasks require input data where the target component represents a pair of nodes.	×	0.05
Graph classification tasks require graph representation instead of node representation.	×	0.07
Paper [64] designed an alternative operator based on graph powering to replace the classical Laplacian in GNN models.	×	0.04
The Citeseer dataset contains 3,327 nodes, 4,732 edges, 3,703 features, and 6 classes.	×	0.02
The Cora dataset contains 2,708 nodes, 5,429 edges, 1,433 features, and 7 classes.	×	0.02
The FGA algorithm is associated with paper [27] and is available at https://github.com/DSE-MSU/DeepRobust.	×	0.03
The Nettack algorithm is associated with paper [176] and is available at https://github.com/danielzuegner/nettack.	×	0.03
The RL-S2V and GraArgmax algorithms are associated with paper [33].	×	0.03
The Meta-self and Greedy algorithms are associated with paper [178] and are available at https://github.com/danielzuegne	×	0.02

References

- http://arxiv.org/abs/1905.11736v5
- http://arxiv.org/abs/1812.10528v4
- http://arxiv.org/abs/2104.09369v1