GOttack: Universal Adversarial Attacks on Graph Neural Networks via Graph Orbits Learning

Zulfikar Alom, Tran Gia Bao Ngo, Murat Kantarcioglu, Cuneyt Gurcan Akcora

ICLR | 2025
The Thirteenth International Conference on Learning Representations

Learning on Graphs

- Graphs are omnipresent.
- Adversaries are very common in graph application scenarios.
- Graph Neural Networks (GNNs) have demonstrated superior performance in node classification tasks across diverse applications.
- GNNs have vulnerability to adversarial attacks, where minor perturbations can mislead model predictions.

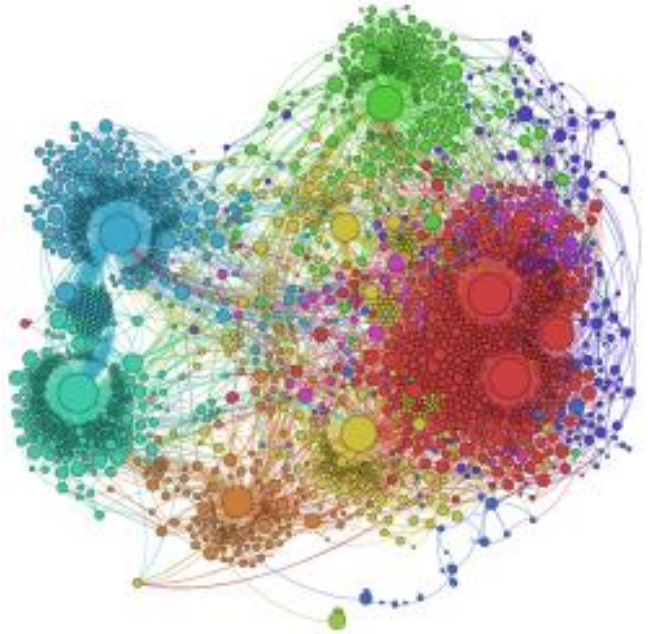
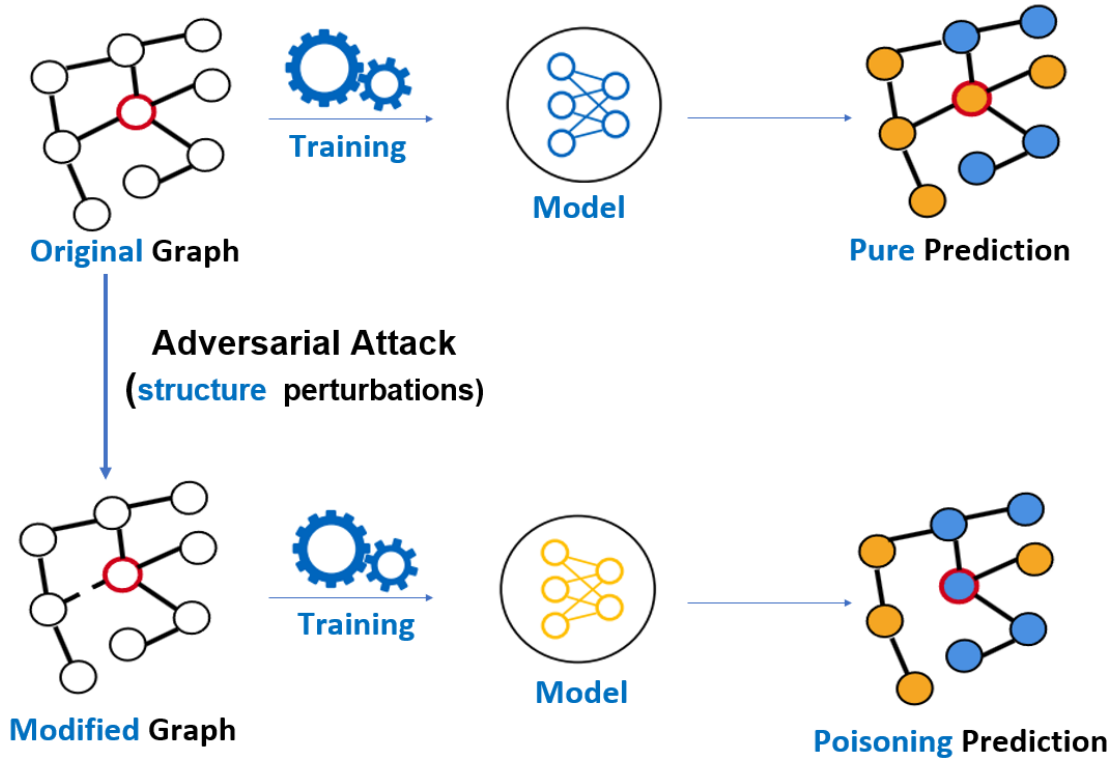


Image: https://www.mkbergman.com/968/a-new-best-friend-gephi-for-large-scale-networks/

Adversarial Attacks – Graphs

- Adversarial attacks are a real threat.
- Adversarial attacks on graphs can modify the graph structure.
- Poisoning Attack: Alters the training data in a transductive setting.

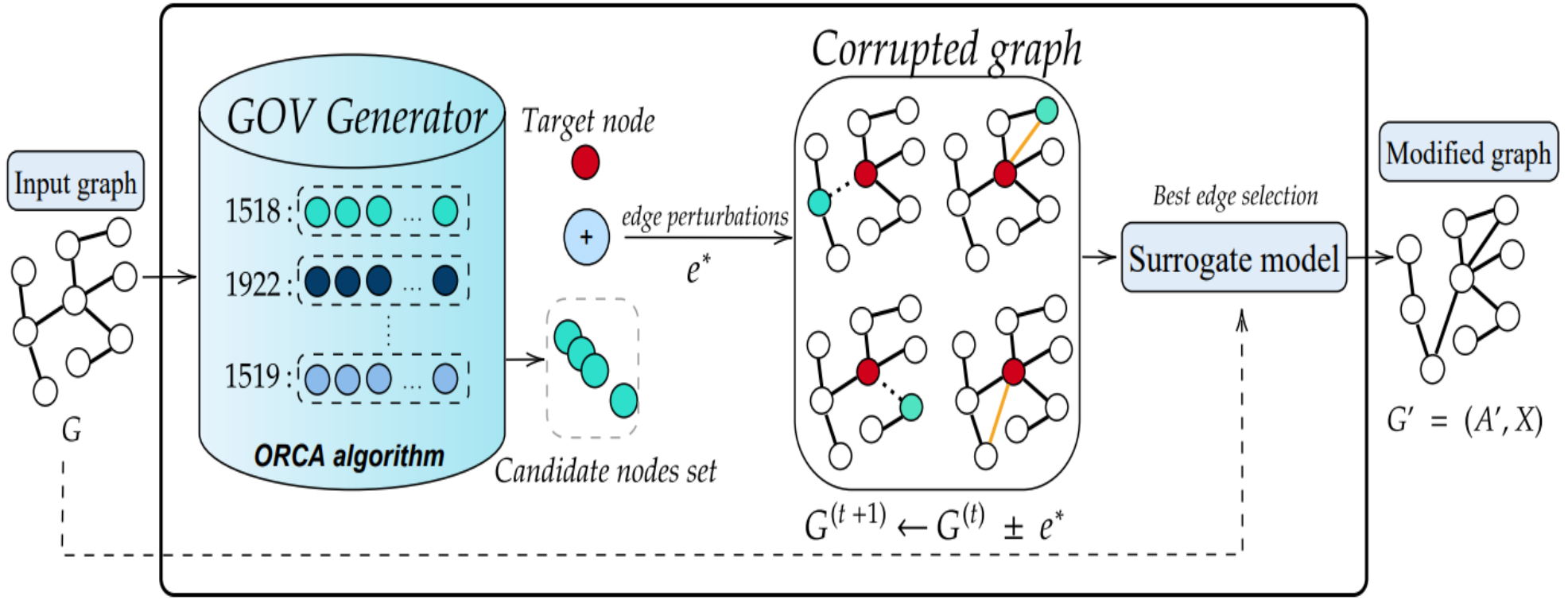


Proposed Approach

- **GOttack**, draws inspiration from the Mapper philosophy of Topological Data Analysis, to undermine the integrity of GNN predictions.
- Groups nodes according to their positions on the graph in potential attack strategies.
- Enhancing both the precision and time-efficiency of adversarial interventions.

GOttack Framework

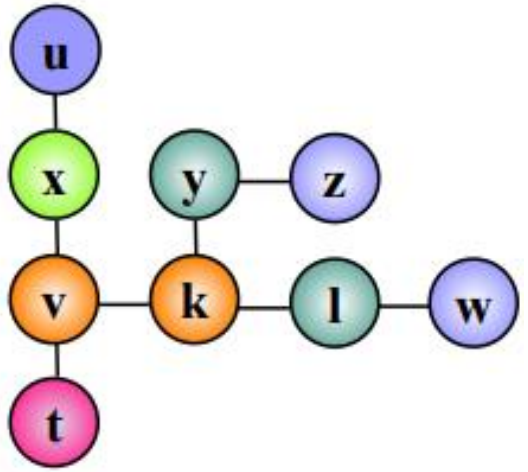
- Extract connected induced subgraphs in graph.
- Identify node orbits, defined by the automorphisms of each graphlet.
- Utilize the ORCA algorithm to compute orbit counts from all graphlets.
- Discover unique structural patterns associated with orbits 15 and 18 nodes.
- Nodes in orbits 15 and 18 to strategically add or remove edges.



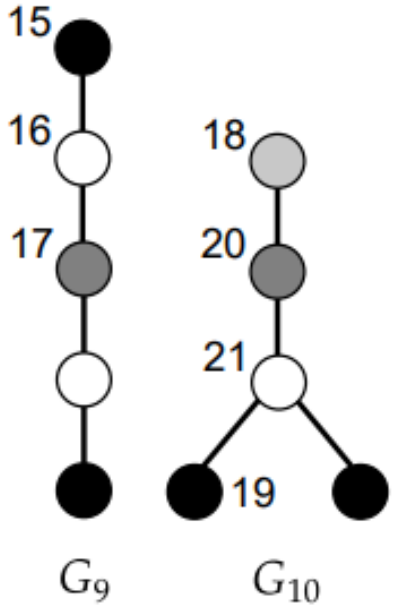
Graph Structure Poisoning Via Orbit Learning

Topological observation influenced by two Mapper philosophy:

1. Orbit Proxy: under the homophily assumption, graph neighbors are similar to a node in the label.
2. Periphery Orbits. Orbits 15 and 18 indicate the topological periphery in a graph and provide a useful proxy for identifying distant nodes that differ in labels.



- Periphery Orbits observation forms the backbone of GOttack strategy; identify nodes of periphery orbits (i.e., 15 and 18) and affect the adjacency matrix (i.e., add or remove edges to these nodes) to confuse a GNN to misclassify a target.



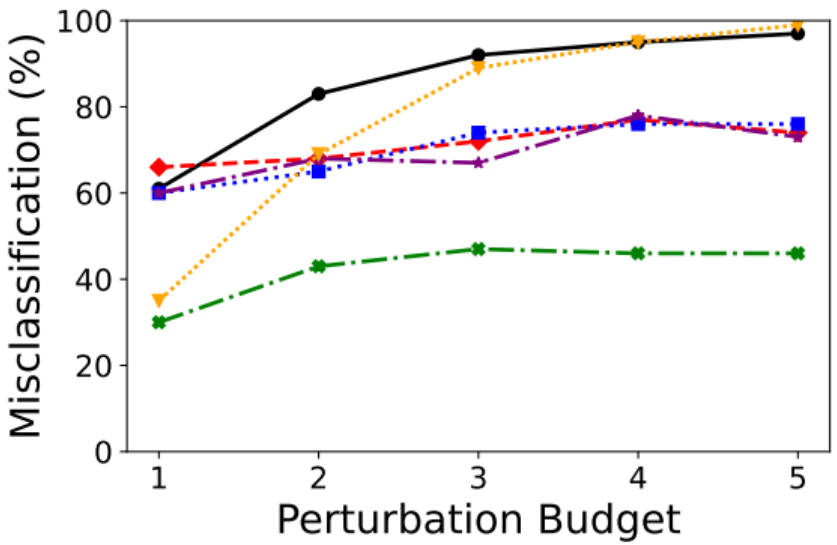
Orbit-based Node Selection in Nettack

Dataset	Orbits	% of nodes	% in 1^{st} Attack	% in 2^{nd} Attack
Cora ($h = 0.81$)	1518	24.00%	77.00%	71.10%
	1519	14.41%	5.70%	14.29%
	1819	11.59%	10.00%	12.50%
Citeseer ($h = 0.74$)	1518	21.99%	51.60%	61.29%
	1519	21.18%	12.50%	15.00%
	1819	11.56%	3.29%	0.00%
Polblogs ($h = 0.91$)	1518	9.41%	97.50%	60.00%
	1519	2.29%	0.00%	0.00%
	1819	12.93%	2.50%	27.50%
Pubmed ($h = 0.81$)	1518	20.14%	25.00%	10.00%
	1519	23.07%	32.50%	52.50%
	1618	2.24%	22.50%	10.00%
BlogCatalog ($h = 0.40$)	1518	3.25%	2.50%	22.50%
	1519	61.57%	62.50%	37.50%
	1922	19.77%	35.00%	40.00%

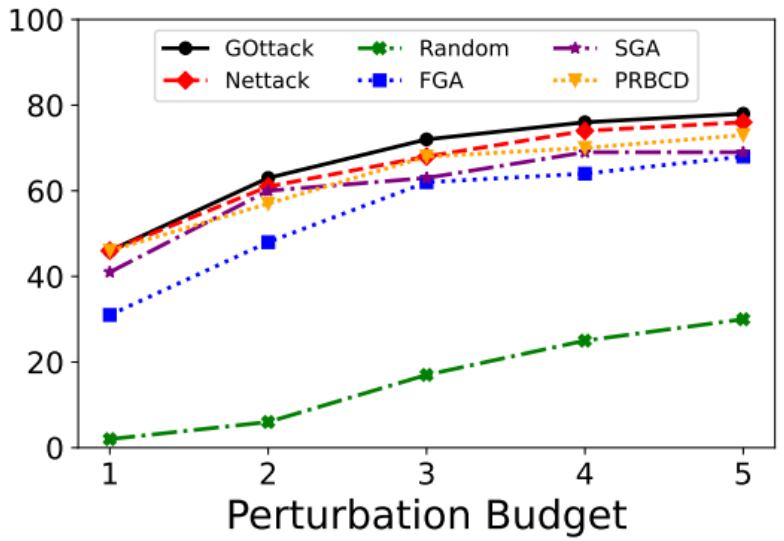
GOttack Evaluation Results I

	Cora			Citeseer			Polblogs			BlogCatalog			Pubmed		
Method	GSAGE	GCN	GIN	GSAGE	GCN	GIN	GSAGE	GCN	GIN	GSAGE	GCN	GIN	GSAGE	GCN	GIN
Random	19.11	2.07	17.30	30.01	2.01	10.03	15.13	12.04	17.04	3.09	12.09	4.19	14.02	20.05	20.05
Nettack	58.04	34.06	46.10	**66.09**	46.04	57.04	29.02	38.04	13.02	50.11	20.02	**65.07**	52.01	50.04	47.02
FGA	54.08	32.05	40.09	60.11	31.10	44.10	22.08	31.09	14.02	46.15	10.04	61.09	42.03	32.02	52.00
SGA	**61.06**	41.05	**57.05**	60.06	41.06	57.06	**35.05**	37.07	**35.08**	51.45	24.03	61.02	30.00	57.04	47.01
PRBCD	35.02	41.06	36.10	35.04	46.04	42.02	8.01	**42.07**	33.06	33.05	**33.05**	5.09	38.02	52.03	43.06
GOttack (ours)	59.05	**41.52**	37.03	61.09	**46.07**	**57.06**	29.03	41.08	15.08	**52.10**	22.04	63.08	**52.08**	**57.09**	**55.05**

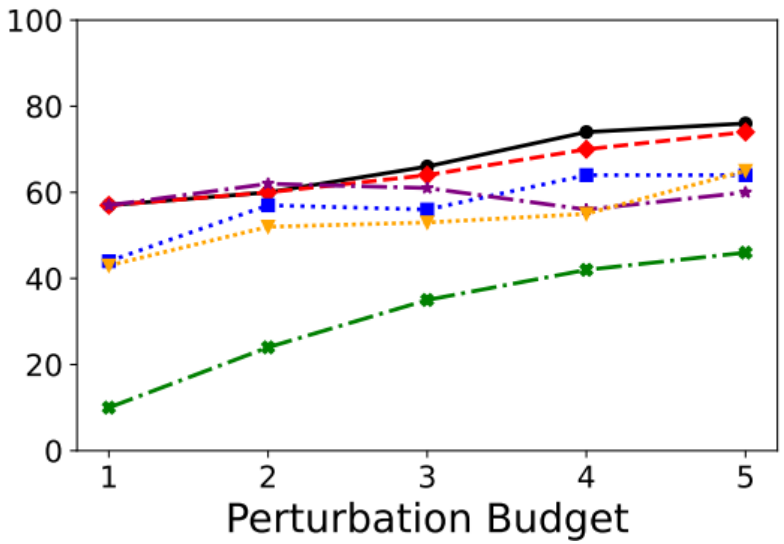
Misclassification rate (in %) (↑) of target nodes in five datasets where three backbone GNNs are attacked in node classification with budget Δ = 1.



GraphSAGE model



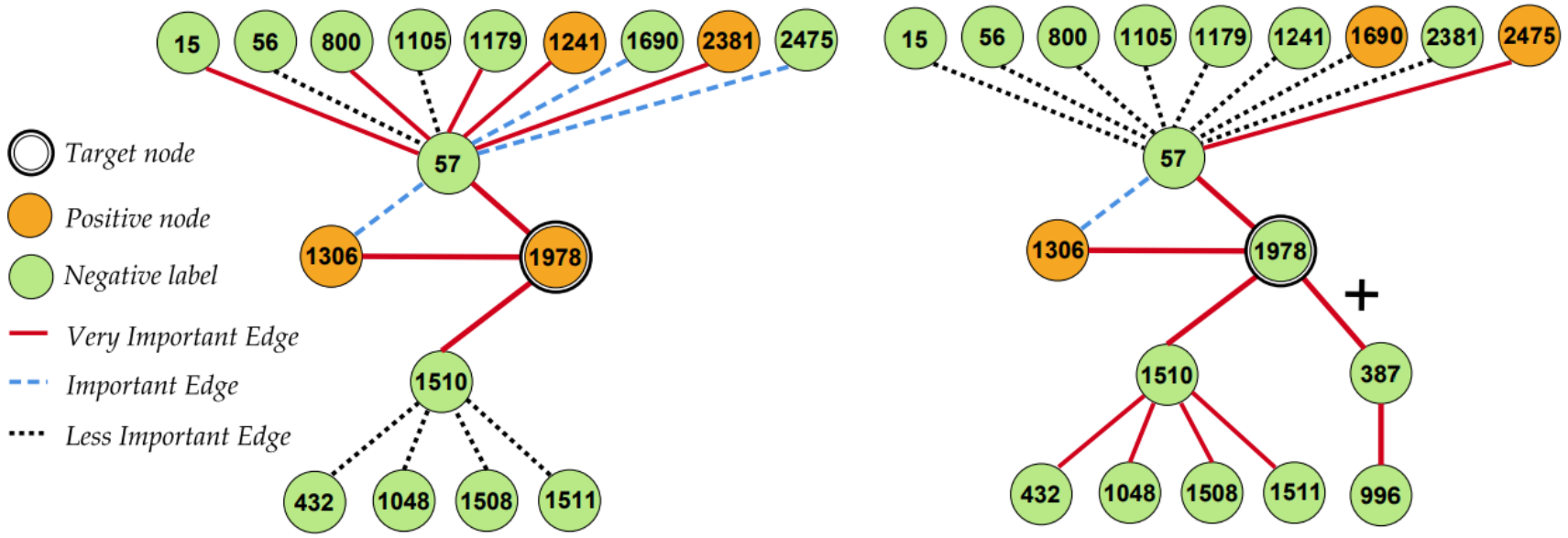
GCN model



GIN model

Budgeted (Δ = 1 to 5) attack results on the Citeseer dataset.

GOttack Evaluation Results II



The computation graph for the targeted node 1978 from the CORA datasets, as identified by GNNExplainer. The edge (1978, 387) is added during the successful attack. Edge importances change considerably after the attack and negative class gains importance due to the newly added nodes.

	Cora				Citeseer				Polblogs				BlogCatalog			
Method	RGCN	JAC	SVD	MDGCN	RGCN	JAC	SVD	MDGCN	RGCN	JAC	SVD	MDGCN	RGCN	JAC	SVD	MDGCN
SGA	44.01	33.02	28.03	32.05	**53.00**	36.01	24.04	**36.01**	**46.03**	43.05	**16.01**	**43.05**	25.02	**25.02**	15.05	**20.01**
Nettack	**48.08**	38.01	24.01	32.05	46.04	42.02	**28.04**	27.05	38.05	46.03	12.04	33.02	35.04	19.03	19.02	17.02
GOttack (ours)	43.04	**39.01**	**28.08**	**32.07**	48.07	**42.09**	25.03	28.02	40.02	**53.06**	10.02	34.07	**35.05**	20.02	**20.00**	19.03

Misclassification rate (in %) (↑) of target nodes in different datasets against four defense models are attacked in node classification.

GOTTACK

mdzulfikar.alom@utoledo.edu

GitHub: https://github.com/cakcora/GOttack