GNNGuard: Defending Graph Neural Networks against Adversarial Attacks

1. Take-Home Message

GNNGuard is a model-agnostic approach that can defend any Graph Neural Network against a variety of poisoning adversarial attacks.

2. Featured Properties

- Defense against a variety of attacks: e.g., directly targeted, influence targeted, and non-targeted attacks
- Integrates with any GNNs
- State-of-the-art performance on clean graphs
- Homophily and heterophily graphs: the first technique defending GNNs against attacks on both homophily and heterophily graphs

3. Motivation

- GNNs are highly vulnerable to adversarial attacks
  - Adversarial attacks: inject carefully-designed perturbations (e.g., fake edges) to graph to degrade GNN classifier
- The vulnerability significantly prevent GNNs from real-world applications

4. Method

GNNGuard detects fake edges and alleviates the negative impact on prediction by removing them or assigning them lower weights in neural message passing.

Compared to a typical GNN (panel A), GNNGuard (panel B) controls the message stream, such as blocking the message from irrelevant neighbors while strengthening messages from highly-related ones.

GNNGuard contains two key components:

- **Neighbor Importance Estimation:** 1) estimate the importance of each edge in neighborhood; 2) prune fake edges and assign lower weights to likely-fake edges
- **Layer-Wise Graph Memory:** 1) keeps partial memory of the pruned graph structure from the previous layer; 2) smooth the evolution of edge pruning

5. Experiments

GNNGuard outperforms existing defense approaches by 15.3% on average across five GNNs, three cutting-edge defense baselines, and three adversarial attackers.

### Results in Graphs with Homophily

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset Type</th>
<th>No Attack</th>
<th>Attack 1</th>
<th>Attack 2</th>
<th>Attack 3</th>
<th>Attack 4</th>
<th>Attack 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCN</td>
<td>Homophily</td>
<td>0.837</td>
<td>0.862</td>
<td>0.830</td>
<td>0.861</td>
<td>0.836</td>
<td>0.843</td>
</tr>
<tr>
<td>GAT</td>
<td>Homophily</td>
<td>0.827</td>
<td>0.866</td>
<td>0.820</td>
<td>0.852</td>
<td>0.826</td>
<td>0.843</td>
</tr>
<tr>
<td>GIN</td>
<td>Homophily</td>
<td>0.831</td>
<td>0.870</td>
<td>0.831</td>
<td>0.859</td>
<td>0.834</td>
<td>0.850</td>
</tr>
</tbody>
</table>

### Results in Graphs with Heterophily

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset Type</th>
<th>No Attack</th>
<th>Attack 1</th>
<th>Attack 2</th>
<th>Attack 3</th>
<th>Attack 4</th>
<th>Attack 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCN</td>
<td>Heterophily</td>
<td>0.834</td>
<td>0.860</td>
<td>0.834</td>
<td>0.859</td>
<td>0.835</td>
<td>0.848</td>
</tr>
<tr>
<td>GAT</td>
<td>Heterophily</td>
<td>0.826</td>
<td>0.863</td>
<td>0.822</td>
<td>0.849</td>
<td>0.828</td>
<td>0.845</td>
</tr>
<tr>
<td>GIN</td>
<td>Heterophily</td>
<td>0.832</td>
<td>0.871</td>
<td>0.834</td>
<td>0.859</td>
<td>0.836</td>
<td>0.851</td>
</tr>
</tbody>
</table>

Xiang Zhang, xiang_zhang@hms.harvard.edu
Marinka Zitnik, marinka@hms.harvard.edu