Adversarial Attacks on Neural Networks for Graph Data
Daniel Zügner, Amir Akbarnejad, Stephan Günnemann

**Motivation**
- Deep learning for graphs has brought great improvements on many tasks, e.g. node representation learning or node classification.
- Adversaries are now a real threat in the applications where these models are likely to be used, e.g. search engines, recommender systems.
- Before deploying these models, we have to answer the question:

  Are deep learning models for graphs robust with respect to adversarial attacks?

**Focus:** Semi-supervised Node Classification

**Possible Attacks on Graphs**
- In addition to feature vectors, adversarial attacks on graphs can modify the graph structure.
- Even more critical: a node can be attacked without direct access due to network effects.

**Surrogate Model**
- Linear surrogate model based on two-layer GCN.
- Enables computation of the exact impact of a perturbation efficiently and in closed form.
- Attacker chooses perturbation that maximizes loss on the surrogate model (one at a time).

**Unnoticeability Constraints**
- $(A', X') = (A, X)$: What are sensible measures of "closeness" for graphs?

**Unnoticeability Constraints**
- Fundamental property of graphs: degree distribution.
- Hypothesis test: Were the original and modified degree distributions $D$ and $D'$ generated by the same underlying powerlaw distribution?

<table>
<thead>
<tr>
<th>Class probabilities (clean data)</th>
<th>Class probabilities (5 modifications)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>Class 2</td>
</tr>
<tr>
<td>Clean</td>
<td>0.90</td>
</tr>
<tr>
<td>Ours</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Example Attack**
- Transfer experiments: Graph Convolutional Network (GCN), Column Network, DeepWalk.
- Perturbation budget is $d$, where $d$ is the target's degree.
- Evaluation on 5 different splits; 10-retraining per attack.

**Experimental Results**
- We propose an efficient algorithm for adversarial attacks on deep-learning for graphs.
- No access to the classifier is needed for the attack.
- Our attacks are successful even under restrictive attack scenarios, e.g. no access to target node or limited knowledge about the graph.