Certifiable Robustness to Graph Perturbations
Aleksandar Bojchevski, Stephan Günnemann
www.kdd.in.tum.de/graph-cert/

TL;DR: First method for certifiable robustness to graph perturbations for a general class of models that includes label propagation and graph neural networks.

Research Questions
Certification: How to verify if a graph-based model is robust?
Robust Training: How can we improve (certified) robustness?

Semi-supervised Node Classification: Given a graph and a small number of labelled nodes predict the class of the rest

Flexible Threat Model
Attacker controls fragile edges they can turn on or off

Global Budget: perturb at most $B$ edges in total
Local Budget: at most $b_v$ edges for node $v$

Robustness Certificate
Guarantee: that the prediction does not change under any admissible perturbation of the input graph

Family of Models based on PageRank
Predictions are a linear function of (Personalized) PageRank
$log p_G(t,c) = h(c)^T \pi_G(t)$

Personalized PageRank $\pi(t)$: Stationary distribution of a random walker teleporting back to node $t$ with probability $\alpha$

EX 1 - Label Propagation: repeatedly diffuse initial beliefs $H^{(0)}$

$$ H^{(0)} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad H^{(t+1)} = (1-\alpha)D^{-1}AH^{(t)} + \alpha H^{(0)} $$

$$ log p_G(t,c) = H^{(0)}_{vc} \pi_G(t) $$

EX 2 - Graph Neural Network (PPNP): first map node features to initial beliefs with a NN $f_0$ then diffuse with $\pi_G(t)$

To increase ratio of certified nodes: decrease budget / lower $\alpha$

Certificate $\iff$ PageRank Optimization $\iff$ MDP
To compute a certificate solve a Markov Decision Process

Robust Training
Use worst-case margin during training to learn robust weights

Robust cross-entropy: worst-case instead of standard logits
Hinge-loss penalty: maximize the worst-case margin

$$ L_{RCE} = L_{CE}(-m^*) \quad L_{CEM} = L_{CE} + \sum_c \max(0, M - m^* \log(1+c/m^*)) $$

Certification Results
GNNs are more robust than Label/Feature Propagation

Two attacker scenarios: add and remove or only remove edges
Three models:
- (83%) GNN (PPNP)
- (82%) Feature Propagation
- (73%) Label Propagation

To increase ratio of certified nodes: decrease budget / lower $\alpha$

Local + Global budget: NP-Hard, so formulate as QP and relax

NeurIPS 2019
December 8 – 14
Vancouver, Canada