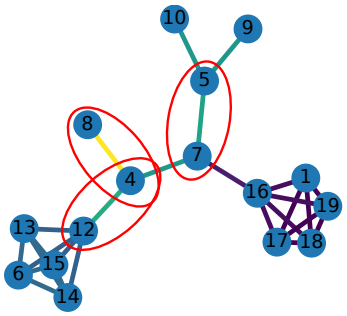
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Adversarial Attacks in Graph Neural Networks



Graph Neural Networks (GNNs) are bringing the power of deep neural networks to the domain of graph-structured problems. This allows prediction and classification tasks on a variety of interesting problems. Like other neural architectures, they are very vulnerable to adversarial attacks. Such attacks create innocent-looking modifications to the graph — but the implications are drastic classification errors.

Consider the example in the figure. The GNN detects that the graph contains cliques. The edges that make the GNN most confident in its prediction are highlighted. Changing these edges can cause the biggest change in prediction, even though all cliques are unaffected. Clearly, removing any highlighted edge is also a "small" change: after all, it is "only" one edge. But every marked edge has the consequence of making the graph disconnected. In this thesis, we want to explore less drastic changes that still tremendously impact the performance of a trained GNN.

Requirements: Knowledge in Machine Learning and or Deep Learning is advantageous. We will have weekly meetings to discuss the intermediate progress, think together about future ideas, and tackle open questions.

Interested? Please contact us for more details!

Contact

- Lukas Faber: lfaber@ethz.ch, ETZ G60.1
- Zhao Meng: zhmeng@ethz.ch, ETZ G61.3