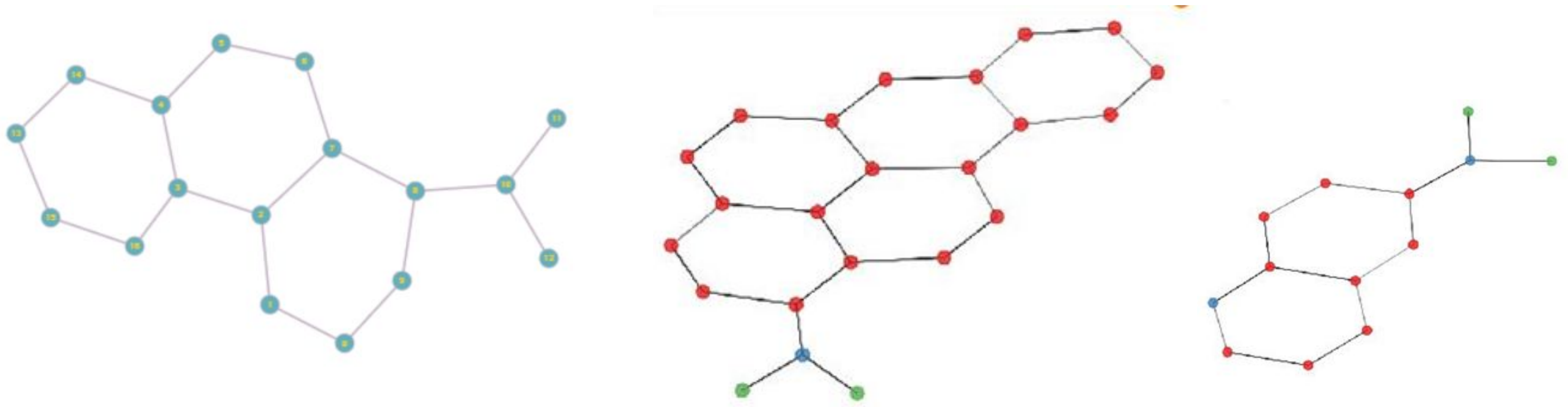
NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

School of Computer
Science and Engineering
College of Engineering

Explainable Graph Classification with Deep Learning Models

Comparing efficacy of existing methods using Subgraph Isomorphism

Student: Rajiv Balamurugan   Supervisor: Asst Prof Arijit Khan



From left to right: the original input graph, the derived computation graph, and the resultant evidence subgraph produced by the method we want to investigate.

Project Objectives:

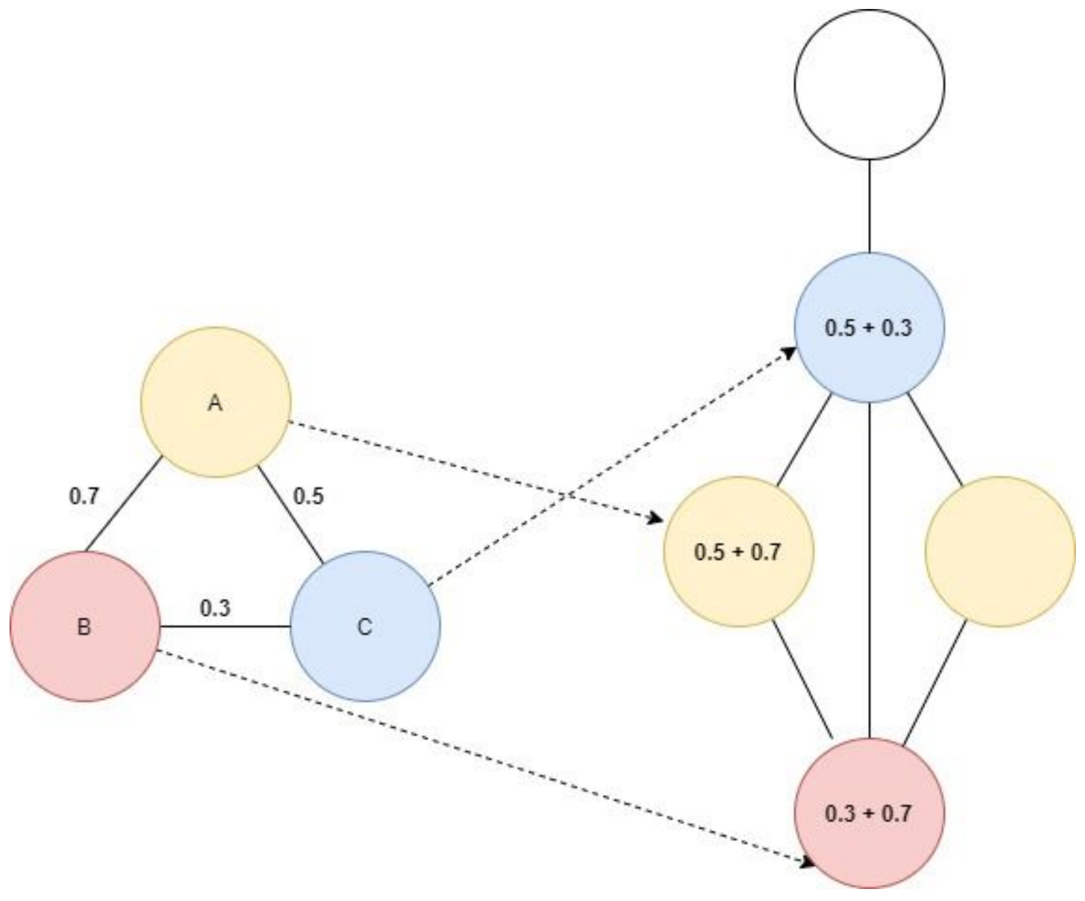
We want to find the explanations behind the classification results of Graph Neural Networks(GNN). GNNs are generally black boxes so we do not understand the inner workings that explains a certain predicted label. To interpret the results of classification, we find important nodes, also known as salient nodes, of the input graphs that contribute the most to the predicted label. These nodes are thus used to justify the classification results to stakeholders. The new method we want to investigate, GNNExplainer, does not have the same approach. Instead of finding important nodes, it produces a subgraph which is a composite of important structural features of the input graph. We thus find the salient nodes through subgraph isomorphism. After identifying the salient nodes, we find the metrics for the explainability method: fidelity, contrastivity and sparsity.

Fidelity: The degree to which the occlusion of salient nodes change the classification result.

Contrastivity: The degree to which the salient nodes contribute to a specific predicted class label.

Sparsity: The degree to which the salient nodes are rare.

Subgraph Isomorphism to find salient nodes


A
0.7
0.5
B
0.3
C
0.5 + 0.3
0.5 + 0.7
0.3 + 0.7