Adversarial Graph Unlearning

Thesis proposal

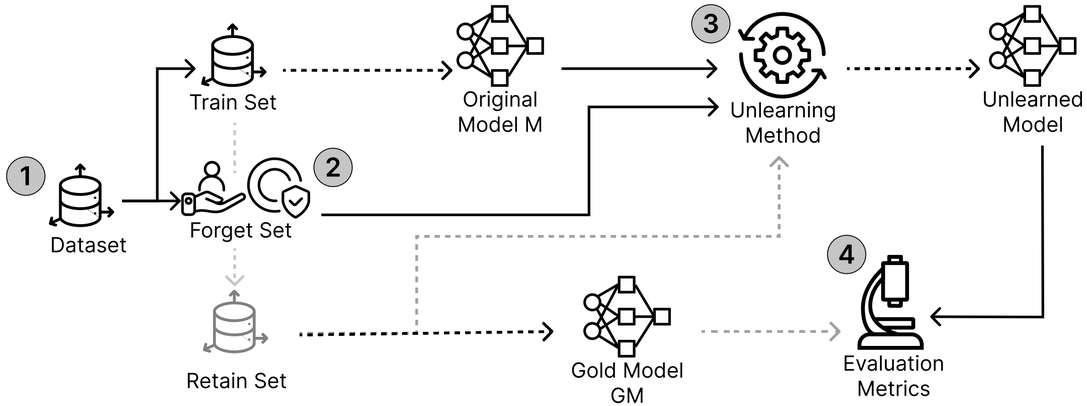
giovanni.stilo@univaq.it, andrea.dangelo6@graduate.univaq.it

Machine Unlearning

Basics

- Machine Unlearning is the task of removing undesired data from the training set of a Machine Learning model.

- Read more:
- https://arxiv.org/pdf/2306.03558
- https://arxiv.org/pdf/2209.02299

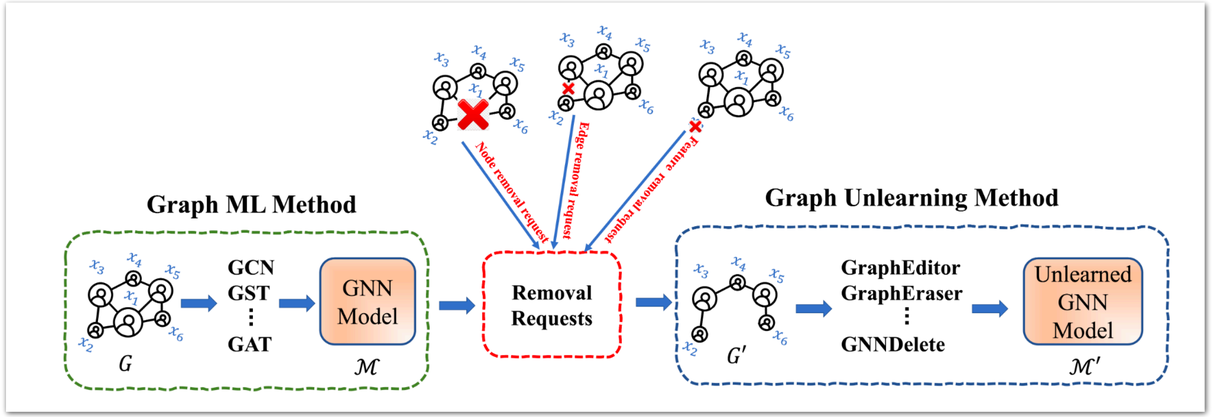


The process starts with a dataset (1), which is split into a Train Set and a Forget Set (2). The Train Set is used to train the Original Model (M), while the Forget Set contains samples to be unlearned. An Unlearning Model (3) is applied to remove the influence of the Forget Set, resulting in an Unlearned Model. A Retain Set is used to train a Gold Model (GM) for comparison. Finally, Evaluation Metrics (4) assess the effectiveness of unlearning by comparing the Unlearned Model to the Gold Model.

Graph Unlearning

3 types of Graph Unlearning:

- Node removal
- Edge removal
- Feature removal

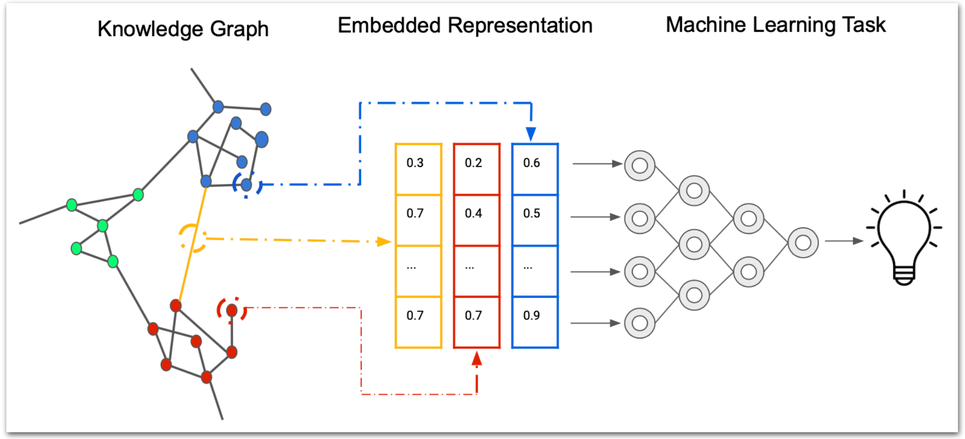


- No method in the literature, as of November 2024, has used an Adversarial model to tackle any of these

Graph Embedding

GNN can embed information on graphs

- Train an Edge predictor as a downstream task on a GNN
- Train GNN’ on the graphs without the edges to be removed, but keep the same edge predictor
- If the edge predictor predicts the removed edges with high probability, then the unlearning was not Robust



Robust Edge Unlearning

Adversarial Approach

- Consider a GAE (Graph AutoEncoder) with encoder and decoder.
- After the encoder, add an adversarial discriminator that is a link predictor
- We want to generate graphs that are as close as possible to the original ones but fool the predictor into not reconstructing the removed edges
- The, the obtained graphs will be the real Forget Set.

Robust Edge Unlearning

Adversarial Approach