Active Learning on Graphs

Aneesh Shetty
Arnab Jana
Riya Baviskar

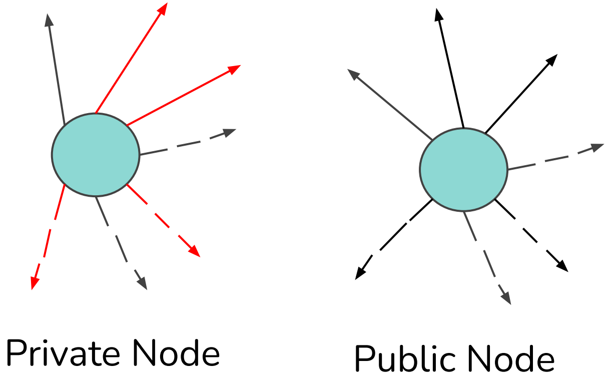
Problem

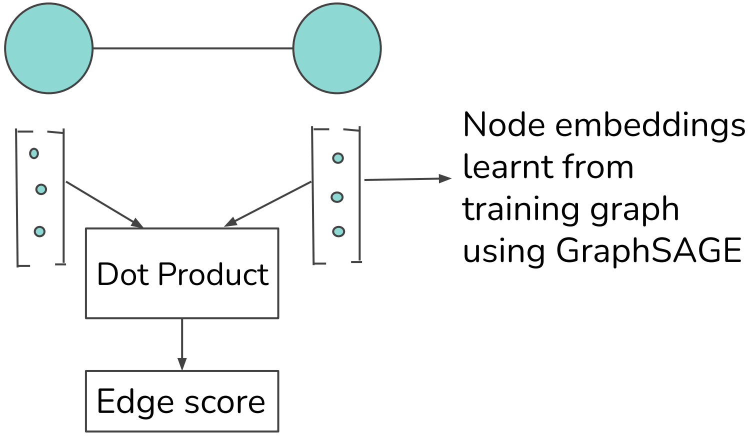
- Link prediction on graph, and retrieval of graph structure.
- We assume, we know the pool set, which contains node-pairs whose status is unknown.
- We have a fixed budget and we can query a node-pair to the oracle for knowing whether it is edge/non-edge, the cost of querying each node being 1 unit.

Standard LP

- We use a transductive model, GraphSAGE to learn the node-embeddings starting with initial domain-knowledge based node-features.
- The score of a node-pair is simply the dot product of the embeddings of its corresponding nodes.
- Evaluation metric- MAP across nodes with large degree, MAP across private nodes
- Loss function- Pairwise hinge-loss

Solid black lines are edges, dotted black lines are non-edges and red lines are private node-pairs

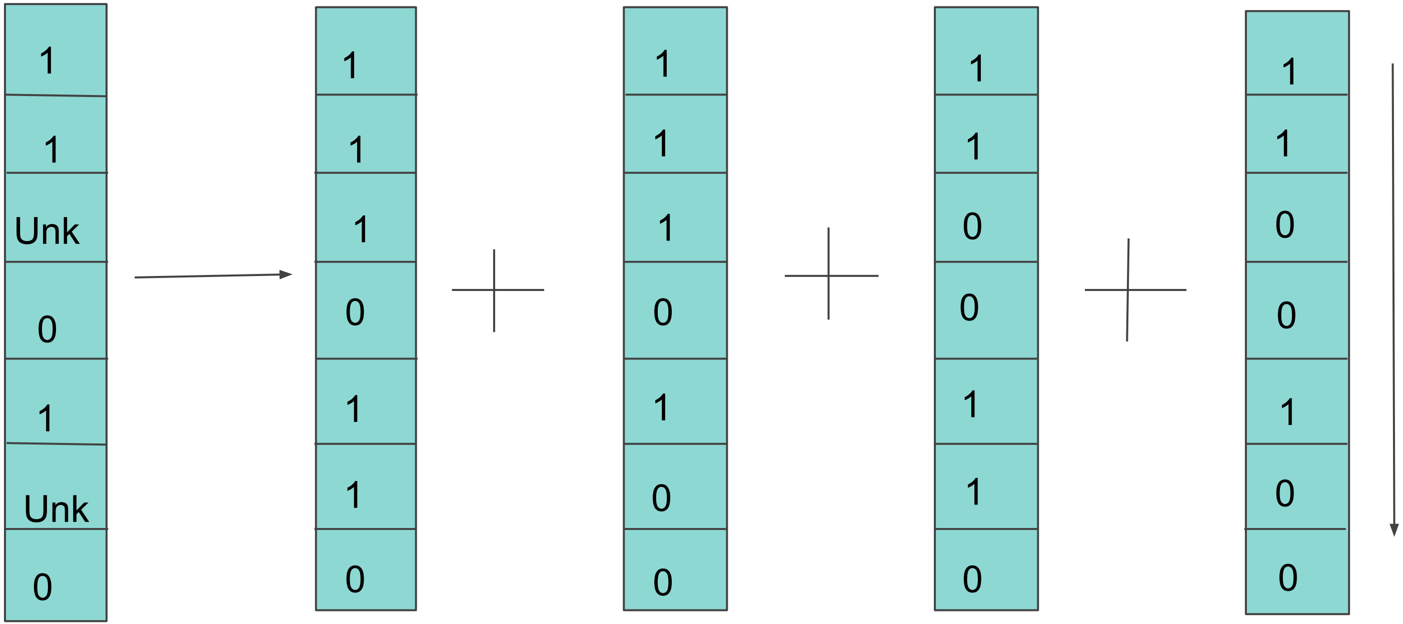




Active Learning

- Pool set is part of the training set, but since the label is unknown, we cannot calculate the actual training loss.
- We assume a uniform distribution across all possible assignments of labels to the pool set.
- The loss function to be minimized is thus the expected loss of our model output and the label assignment available from data.

Example



1	1	1	1	1
1	1	1	1	1
Unk	1	1	0	0
0	0	0	0	0
1	1	1	1	1
Unk	1	0	1	0
0	0	0	0	0

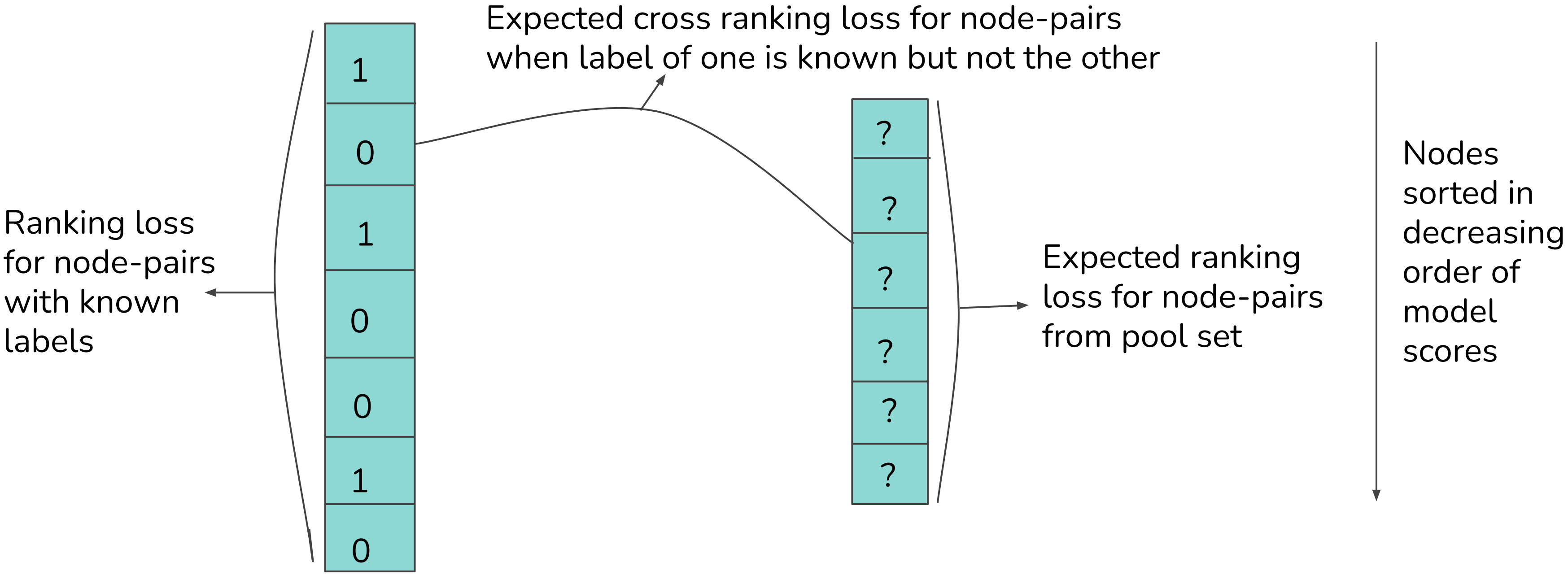
Ranklist for a node. Sorted in decreasing order of scores as given by our model

All these label assignments have equal probability.

Largest Expected Loss Pair

- Our loss function is decomposable to sum of loss functions due to all edge pairs.
- We can also compute the contribution to the loss by a single edge by marginalizing over all edges with common incident node.
- We sort the node-pairs in the pool set in the decreasing order of their contribution to the expected loss and query the top-K to the oracle.
- This should decrease the expected loss wrt to the current model by the maximum margin.

Contribution to expected ranking loss of a particular node



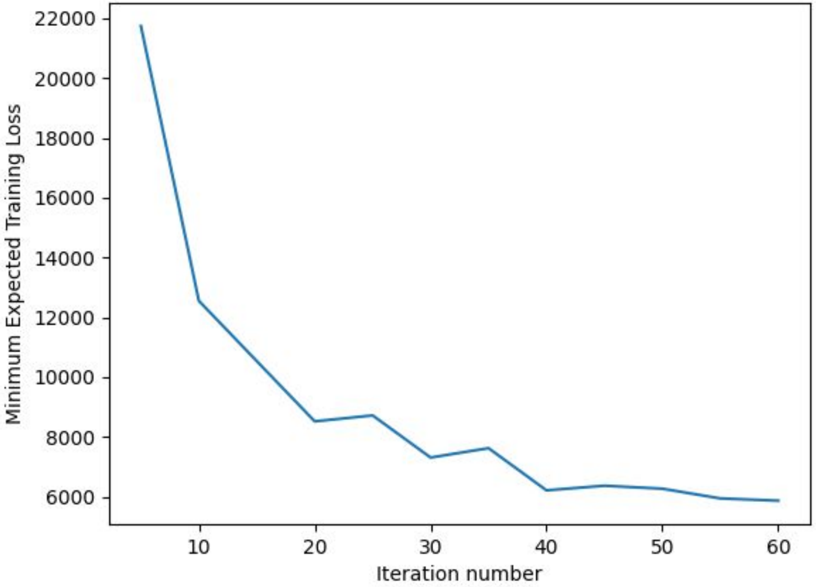
Expected cross ranking loss for node-pairs when label of one is known but not the other

Ranking loss for node-pairs with known labels

Expected ranking loss for node-pairs from pool set

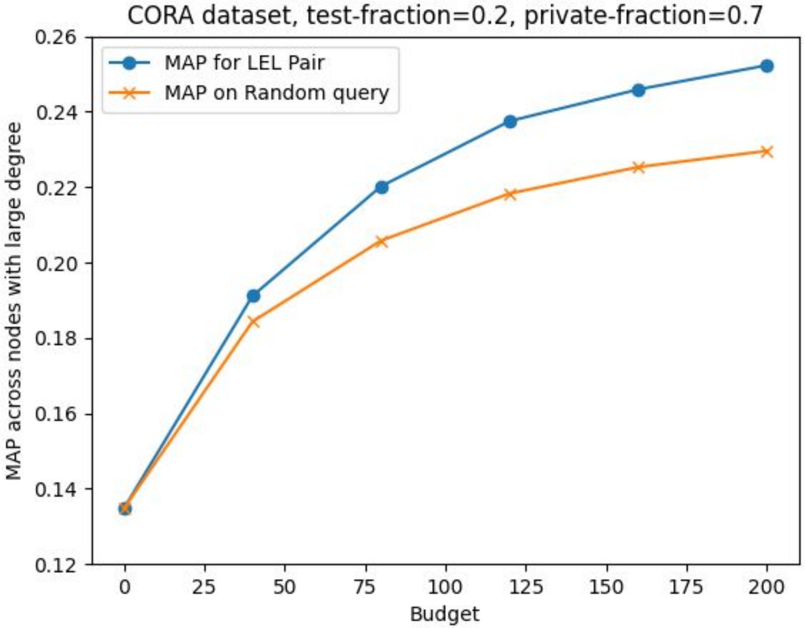
Nodes sorted in decreasing order of model scores

Setup and Results



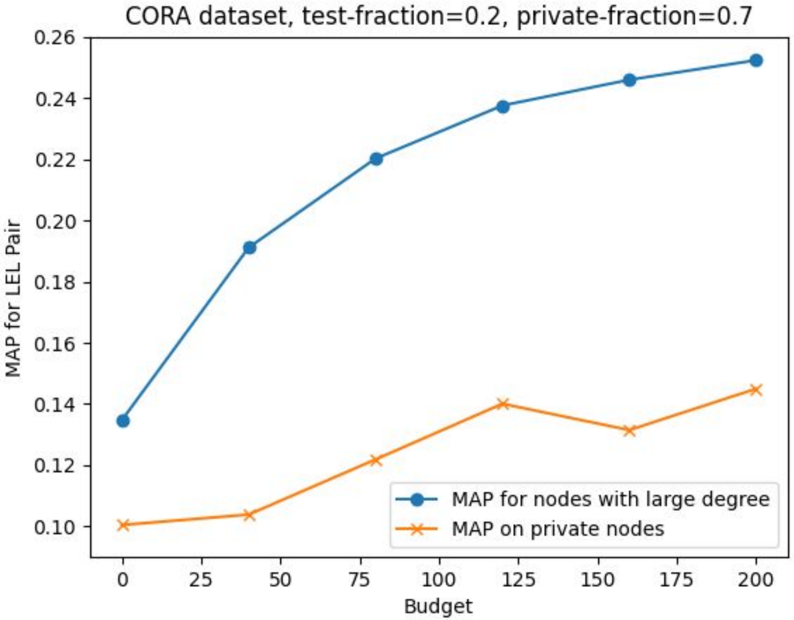
Expected Training Loss vs Iteration number

Setup and Results



Comparison of Random Query and LeL Pair

Setup and Results



MAP on test set for selected nodes and private nodes

Future Work

- Other formulations of Active Learning usually used for classification can be modified for this problem. Challenge would be how to make them Domain Agnostic
    - Entropy minimization over predicted probabilities of the node pairs in the pool set
    - Querying by selection of node pair with largest disagreement over an ensemble of learners
    - Variance Minimization Formulations

THANK YOU!