# A Formal Methods Approach Towards Deep Learning Interpretability

Kriten Kessel, Christopher Lazarus, Javier Sagastuy
Stanford University

## Summary

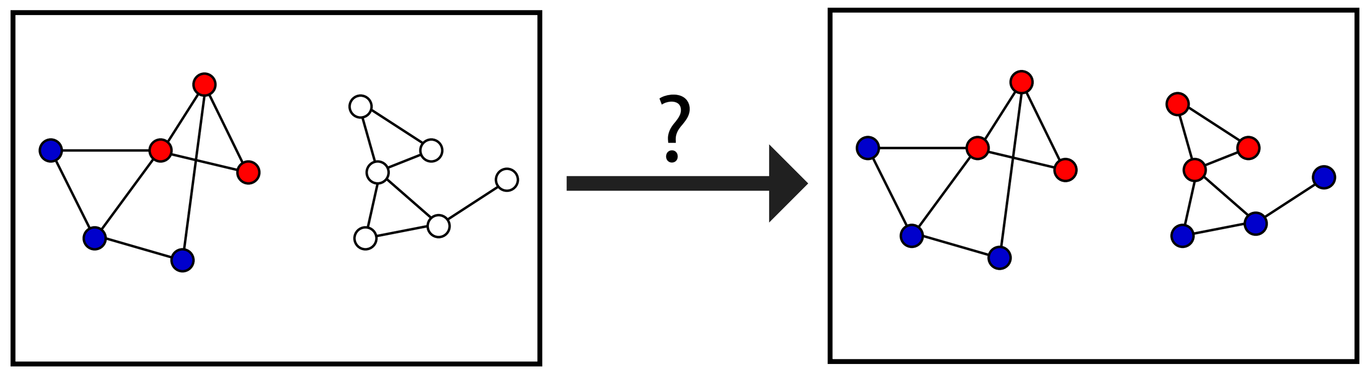
We tackle the problem of **node classification** where the goal is to classify nodes in a graph by leveraging nodes' features and graph structure. We focus on **semi-supervised** settings where only a subset of nodes is labelled and we aim at **transferring knowledge** across labelled and unlabelled nodes.

- **Problem:** node classification on a new population, not connected to the labelled nodes
- **Solution:** hallucinate edges between the labelled nodes and the unlabelled nodes, to reinforce the information flow
- **Results:** our method achieves +3.6% and +3.4% gain in accuracy over standard baselines on `cora` and `citeseer` datasets.

## Problem Setup

Input:

- adjacency matrices $A_L$ and $A_U$ for the two sets of nodes;
- features vectors $X_L$ and $X_U$;
- labels $y_L$

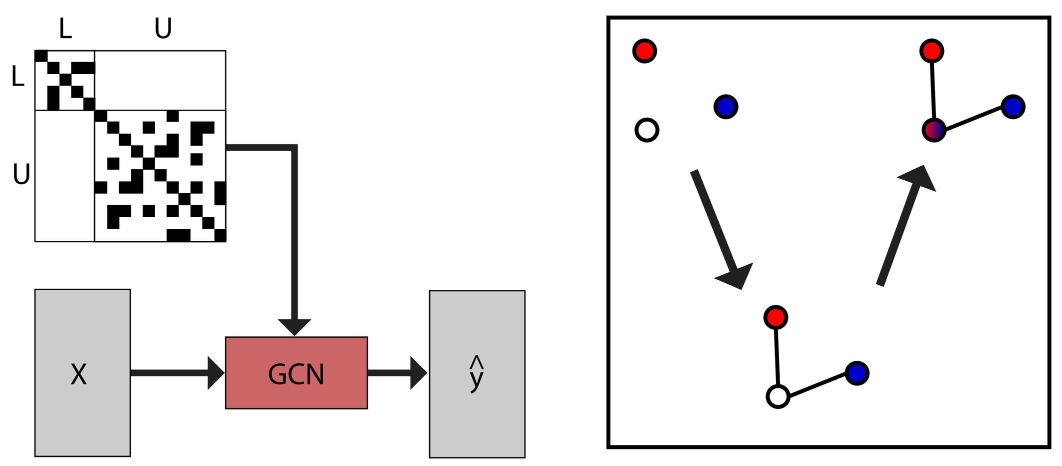



Output:

- predictions for the unlabelled nodes $\hat{y}_L$

## Baseline model

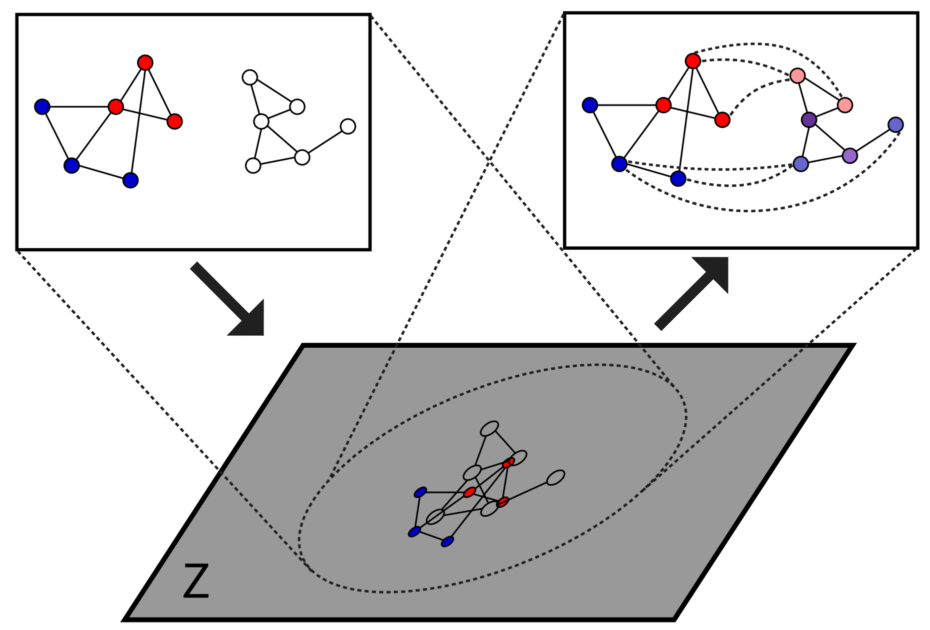
`GCN` [1] for node classification: $\hat{y} = h(Ah(AXW_1)W_2)$




## Approach - Hallucigraph

Three-step process

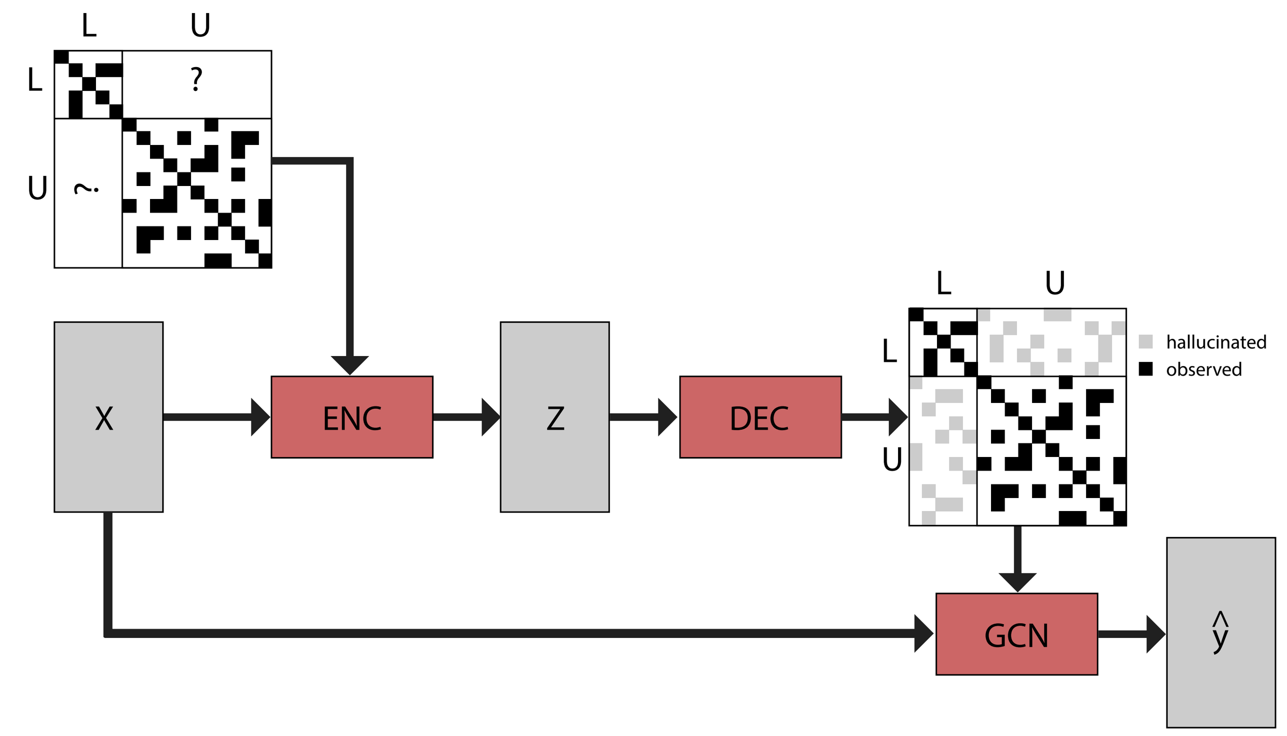
- learn low-dimensional node embeddings to encode node similarity (`VGAE` [2])
- hallucinate edges and complete the adjacency matrix (edges)
- run a `GCN` with the completed adjacency to predict node labels




1. Link prediction - Variational Graph Auto-Encoder (`VGAE`):

- $Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ and $\tilde{A} = \sigma(ZZ^T)$.
- with $\mu_Z, \sigma_Z = \mathrm{GCN}(A, X)$

$$\mathcal{L}_{LP} = -\mathbb{E}_{Z\sim q(Z|A,X)}[A_{ij}\log\tilde{A}_{ij} + (1 - A_{ij})\log(1 - \tilde{A}_{i,j})] + \mathrm{KL}(q(Z|A,X)||p(Z)).$$




2. Edge hallucination produces $\hat{A}$:

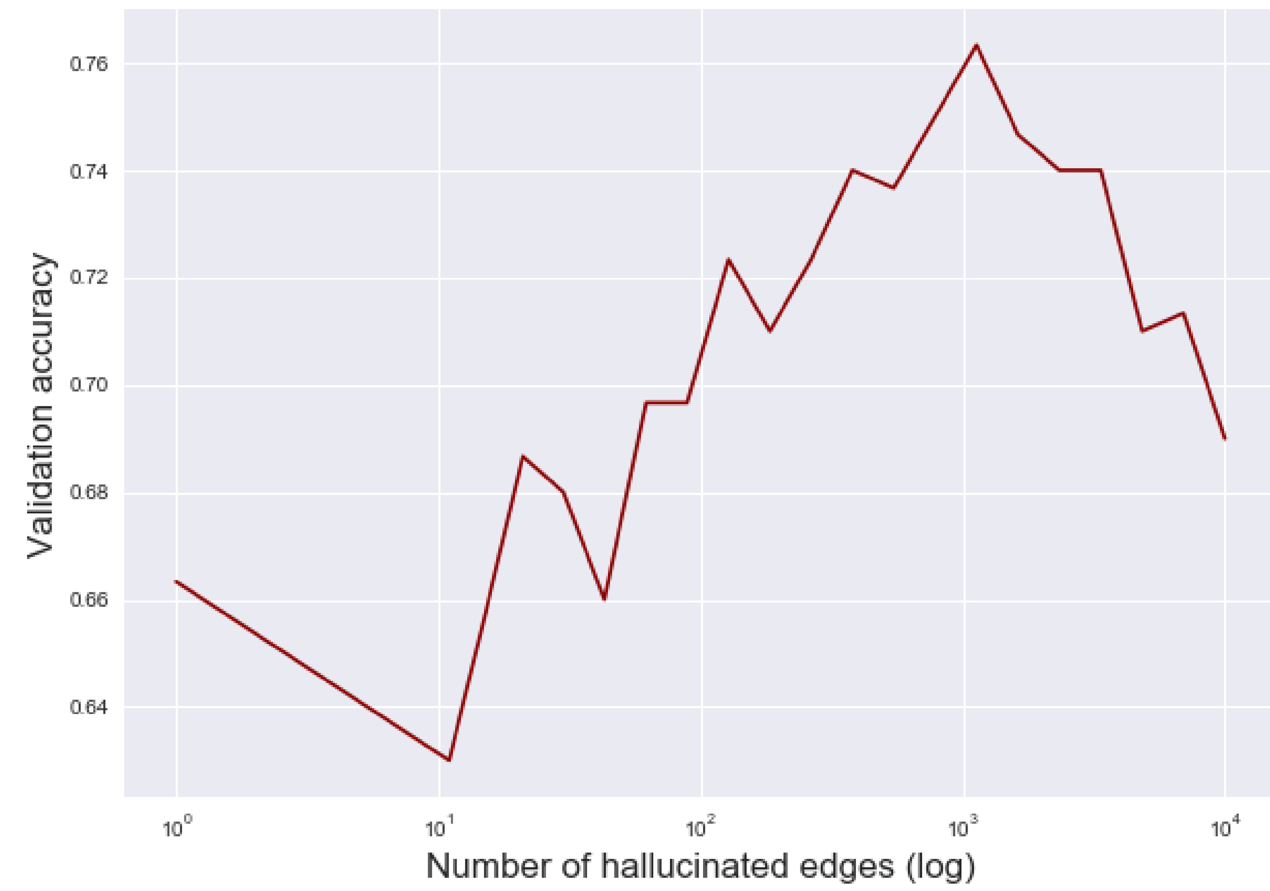
- top$K$ ($K$ hyper-parameter)
- sampling using gumbel softmax [4] trick (allows gradients to flow)

3. Node classification

- $\hat{y} = \mathrm{GCN}(\hat{A}, X)$

## Results

Classification performance per number of hallucinated edges (cora)




Node classification results

| Model | cora | citeseer |
|---|---|---|
| MLP | 56.4% | 56.8 % |
| GCN | 71.3% | 65.8% |
| Hallucigraph | **74.9%** | **69.2%** |

Table: Accuracy results for node classification task on three publication datasets where we removed all $LU$ edges; using plain nodes features without edges (MLP); using the edges within labelled nodes and within unlabelled nodes (GCN); using Hallucigraph.

- As shown in [1], the GCN improves on standalone MLP by leveraging the connections between nodes
- By adding "hallucinated" edges, we improve the connectivity structure, and we obtain more predictive power

## References

[1] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.
[2] Thomas N Kipf and Max Welling. Variational graph auto-encoders. arXiv preprint arXiv:1611.07308, 2016.
[3] Aditya Grover, Aaron Zweig, and Stefano Ermon. Graphite: Iterative generative modeling of graphs. arXiv preprint arXiv:1803.10459, 2018
[4] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax.arXiv preprint arXiv:1611.01144, 2016


{chami, clazarus, marcthib}@stanford.edu