Data Mining and Analytics Group
Department of Informatics
Technical University of Munich

Overlapping Community Detection with Graph Neural Networks

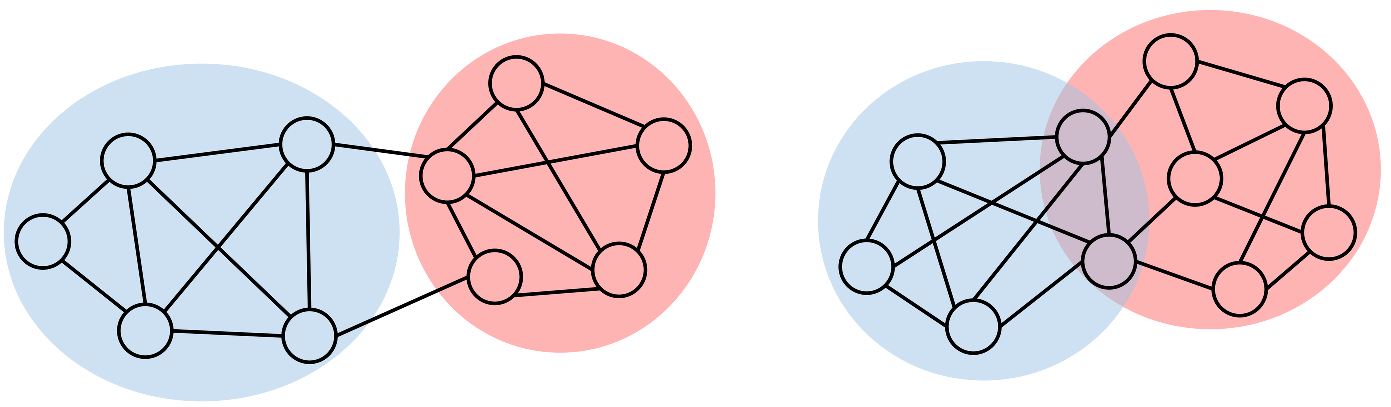
Oleksandr Shchur, Stephan Günnemann

Motivation

- Community detection in graphs is a fundamental problem
- Applications: protein function prediction, social network analysis, fraud detection, neuroscience

Overlapping vs. disjoint communities

Existing deep learning methods can only find disjoint communities



But communities in real networks are overlapping!

Are deep learning methods suitable for overlapping community detection in graphs?

Background: Bernoulli-Poisson model

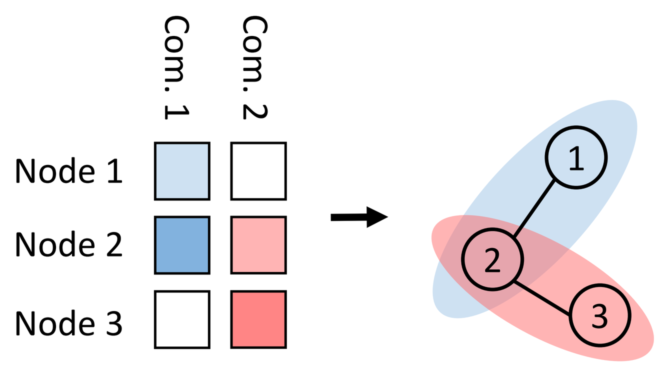
Non-negative community affiliation matrix $F \in \mathbb{R}_{\geq 0}^{N \times C}$

F_{uc} = strength of node u's membership in community c

Generative model for the graph

$$A_{uv} \sim Bernoulli(1 - \exp(-F_u F_v^T))$$

Intuition: Connection probability is proportional to the number of shared communities



Learning: Maximum likelihood

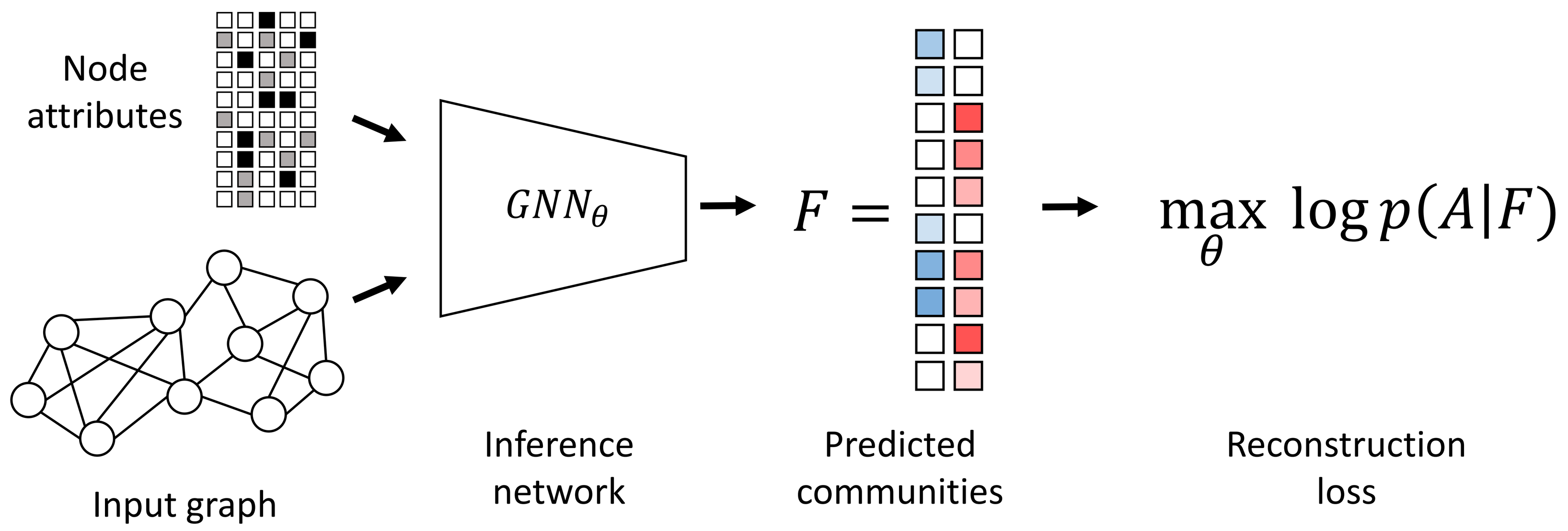
$$\max_{F \geq 0} \log p(A|F)$$

F is treated as a free variable in optimization

Idea: Train a GNN to predict the community affiliations

Neural Overlapping Community Detection (NOCD)

Predict communities using a Graph Neural Network (GNN)



Challenges

Real graphs are extremely sparse
→ Balance loss for edges and non-edges

Naïve loss computation is $O(N^2)$
→ Use stochastic optimization – leads to $O(batch\_size)$

Node attributes might be uninformative / unavailable
→ Use adjacency matrix as input

Attractive properties

- Accurate: State-of-the-art results in community recovery
- Scalable: Train on a graph with 800K edges in 3 minutes on a single consumer GPU
- Works out of the box: Single hyperparameter configuration works across graphs with extremely different properties

Inductive community detection

- Fit the model on the training graph G_{train}
- Predict communities for the test graph G_{test} with a single forward pass
- Applications
  1. Time evolving graphs: Infer communities of new nodes without retraining
  2. Scalability: Train using a small subset of the original graph

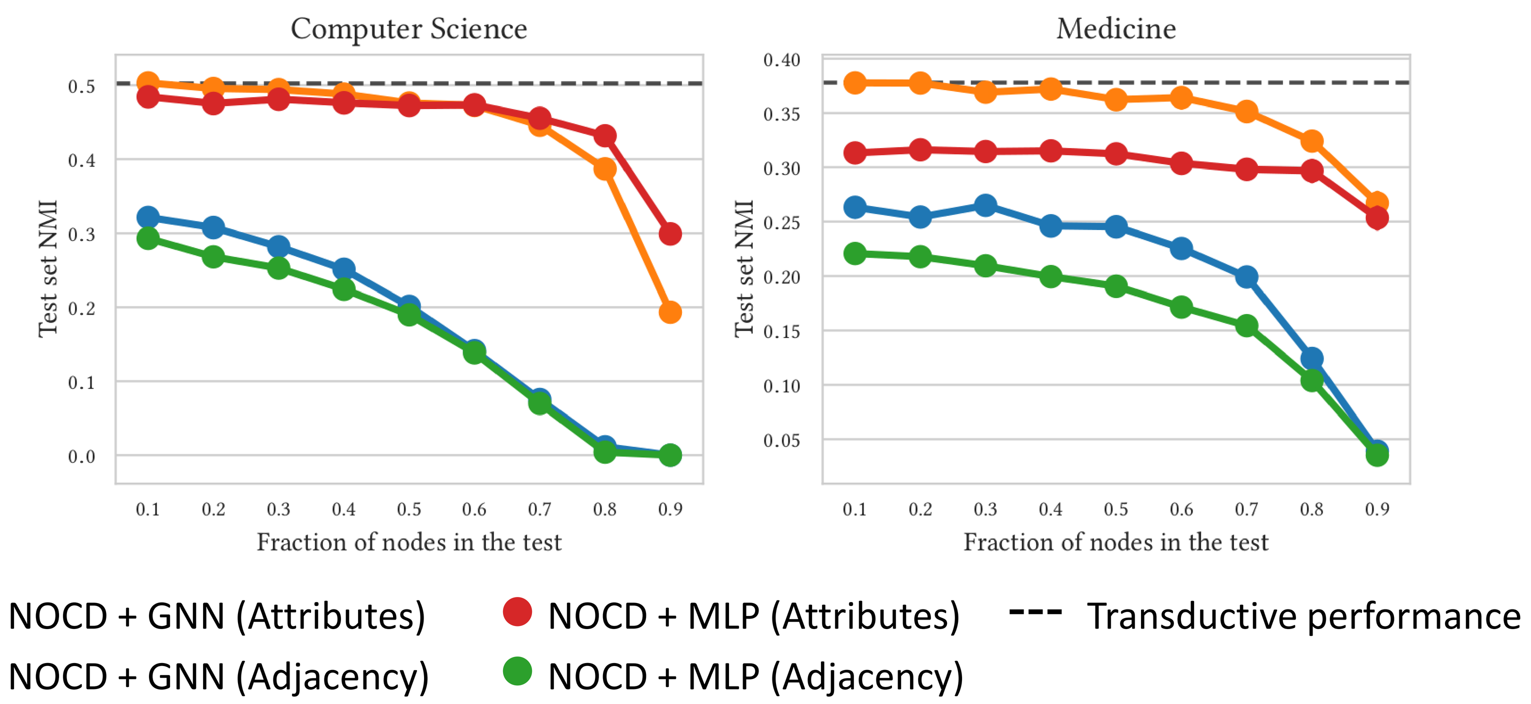
NOCD recovers ground truth communities

- Excellent results in community recovery
- Choose between attribute- and adjacency-based variants of the model by choosing one with better reconstruction loss

Dataset	BigCLAM	CESNA	SNetOC	DW/NEO	NOCD-Adj	NOCD-Attr
Facebook 414	48.3	50.3	52.0	40.9	56.3	59.8
Facebook 686	13.8	13.3	10.6	11.8	20.6	21.0
Facebook 698	45.6	39.4	44.9	40.1	49.3	41.7
Facebook 1684	32.7	28.0	26.1	37.2	34.7	26.1
Facebook 1912	21.4	21.2	21.4	20.8	36.8	35.6
Comp. Science	0.0	33.8	DNF	3.2	34.2	50.2
Engineering	7.9	24.3	DNF	4.7	18.4	39.1
Medicine	0.0	14.4	DNF	5.5	27.4	37.8

Inductive formulation speeds up training

Training using only 30% of the original graph leads to virtually no decrease in performance



GNN is the crucial component

Replacing GNN with an MLP degrades the performance

Dataset	NOCD + GNN	NOCD + MLP	Free variable
Facebook 414	59.8 ± 1.8	22.1 ± 3.1	49.2 ± 0.4
Facebook 686	21.0 ± 0.9	1.5 ± 0.7	13.5 ± 0.9
Facebook 698	41.7 ± 3.6	1.4 ± 1.3	41.5 ± 1.5
Facebook 1684	26.1 ± 1.3	17.1 ± 2.0	22.3 ± 1.4
Facebook 1912	35.6 ± 1.3	17.5 ± 1.9	18.3 ± 1.2
Computer Science	50.2 ± 2.0	49.2 ± 2.0	15.1 ± 2.2
Engineering	39.1 ± 4.5	44.5 ± 3.2	7.6 ± 2.2
Medicine	37.8 ± 2.8	31.8 ± 2.1	9.4 ± 2.3