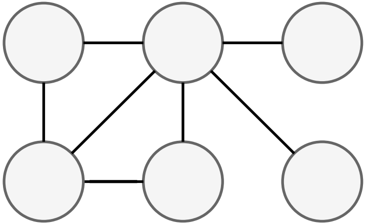
On the Prediction Instability of Graph Neural Networks

Max Klabunde and Florian Lemmerich

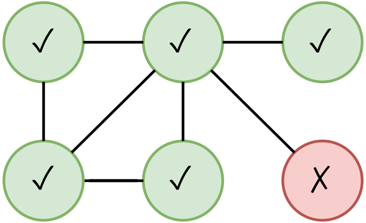
UNIVERSITY OF PASSAU

Faculty of Computer Science and Mathematics

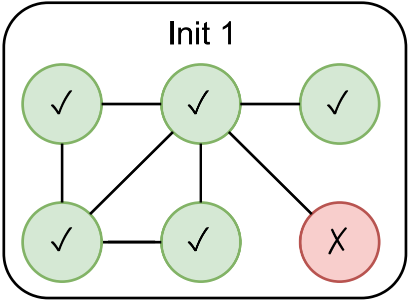
Prediction (in-)stability



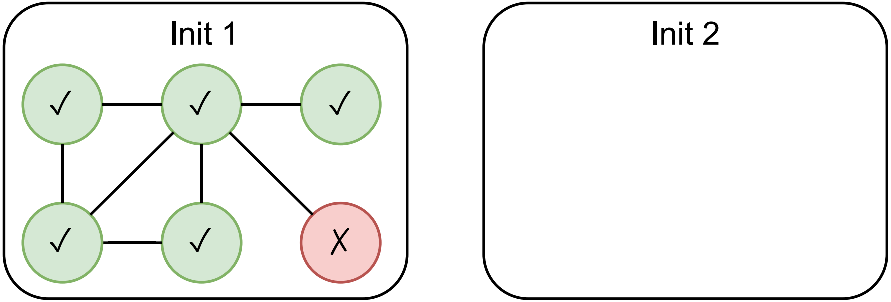
Prediction (in-)stability



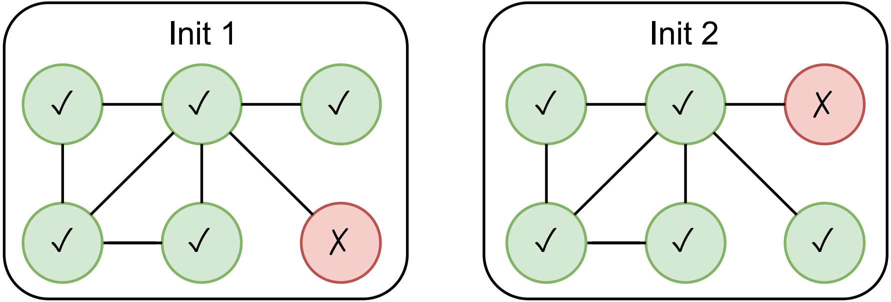
Prediction (in-)stability



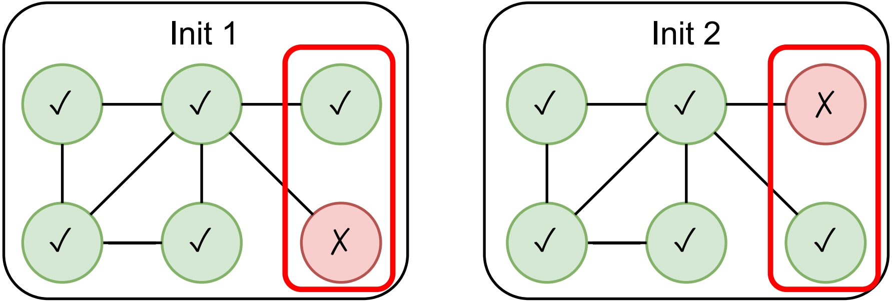
Prediction (in-)stability



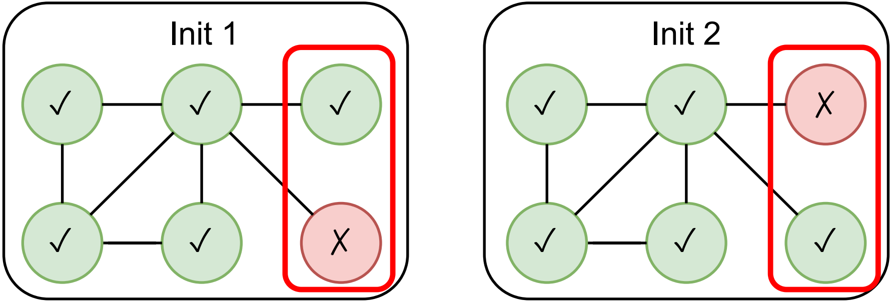
Prediction (in-)stability



Prediction (in-)stability

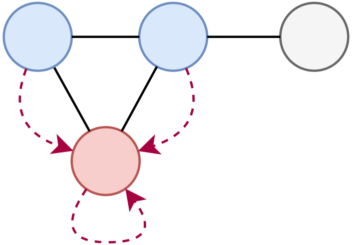


Prediction (in-)stability



- Individual predictions rely on random factors

Graph Neural Networks (GNNs)



- Focus: Graph Convolutional Networks (GCN)[1] and Graph Attention Networks (GAT)[2]

[1] Kipf and Welling 2017.

[2] Veličković et al. 2018.

Context

- Fully reproducible training may be infeasible[3]
- Previous works mainly focus on large non-graph models[4]
- Work on graph models focused on older unsupervised methods with geometrical perspective[5]

[3] Zhuang et al. 2022.

[4] Summers and Dinneen 2021.

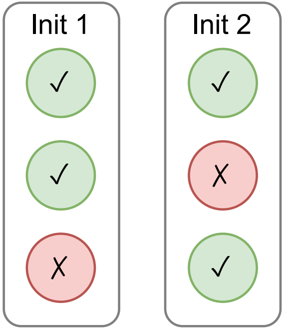
[5] Wang et al. 2022; Schumacher et al. 2021.

Outline

1. Introduction
2. Do GNNs exhibit significant prediction instability?
3. What is the influence of node properties?
4. What is the influence of hyperparameters?
5. Where do instabilities arise?
6. Summary

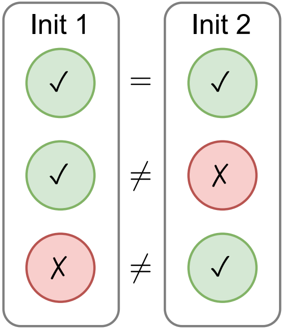
How to measure prediction instability

- Measure with disagreement: share of predictions that are different between two models



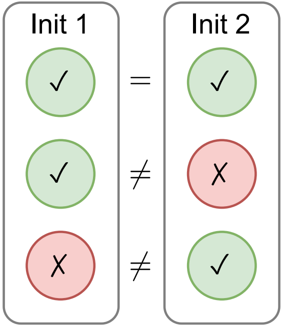
How to measure prediction instability

- Measure with disagreement: share of predictions that are different between two models



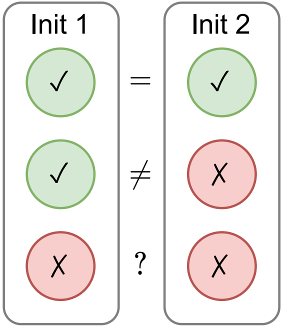
How to measure prediction instability

- Measure with disagreement: share of predictions that are different between two models



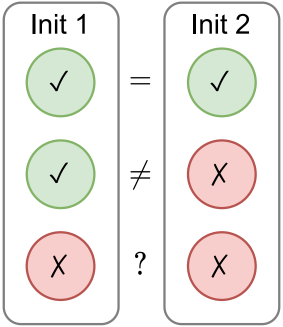
Disagreement $= \frac{2}{3}$

Error rate affects possible disagreement

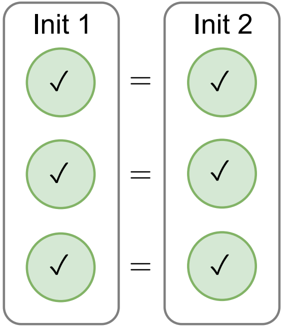


Disagreement $\geq \frac{1}{3}$

Error rate affects possible disagreement

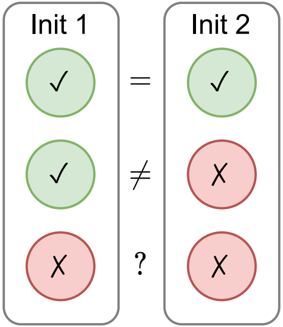


Disagreement $\geq \frac{1}{3}$

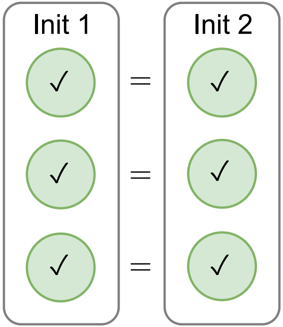


Disagreement $= \frac{0}{3}$

Error rate affects possible disagreement



Disagreement $\geq \frac{1}{3}$



Disagreement $= \frac{0}{3}$

- Normalization possible but similar results

GNN predictions are significantly unstable

- Train 50 different initializations and compare pairwise

Dataset	Model	Accuracy	d	d_{norm}	d_{False}
CiteSeer	GAT				
	GCN				
Pubmed	GAT				
	GCN				
CS	GAT				
	GCN				
Physics	GAT				
	GCN				
Computers	GAT				
	GCN				
Photo	GAT				
	GCN				
WikiCS	GAT				
	GCN				

GNN predictions are significantly unstable

- Train 50 different initializations and compare pairwise

Dataset	Model	Accuracy	d	d_{norm}	d_{False}
CiteSeer	GAT	69.0 ± 1.0			
	GCN	69.2 ± 0.7			
Pubmed	GAT	75.7 ± 0.6			
	GCN	76.8 ± 0.5			
CS	GAT	90.7 ± 0.5			
	GCN	90.7 ± 0.5			
Physics	GAT	92.0 ± 0.7			
	GCN	92.7 ± 0.3			
Computers	GAT	81.0 ± 1.5			
	GCN	81.2 ± 0.9			
Photo	GAT	90.3 ± 0.8			
	GCN	90.8 ± 0.5			
WikiCS	GAT	79.6 ± 0.3			
	GCN	79.4 ± 0.2			

GNN predictions are significantly unstable

- Train 50 different initializations and compare pairwise

Dataset	Model	Accuracy	d	d_{norm}	d_{False}
CiteSeer	GAT	69.0 ± 1.0	10.5 ± 1.7		
	GCN	69.2 ± 0.7	7.1 ± 1.0		
Pubmed	GAT	75.7 ± 0.6	3.7 ± 1.4		
	GCN	76.8 ± 0.5	2.4 ± 0.7		
CS	GAT	90.7 ± 0.5	3.7 ± 0.5		
	GCN	90.7 ± 0.5	3.3 ± 0.6		
Physics	GAT	92.0 ± 0.7	3.8 ± 0.8		
	GCN	92.7 ± 0.3	1.6 ± 0.4		
Computers	GAT	81.0 ± 1.5	9.5 ± 2.2		
	GCN	81.2 ± 0.9	9.9 ± 1.9		
Photo	GAT	90.3 ± 0.8	4.4 ± 1.1		
	GCN	90.8 ± 0.5	3.7 ± 0.8		
WikiCS	GAT	79.6 ± 0.3	3.8 ± 0.5		
	GCN	79.4 ± 0.2	3.3 ± 0.4		

GNN predictions are significantly unstable

UNIVERSITÄT PASSAU

- Train 50 different initializations and compare pairwise

Dataset	Model	Accuracy	d	d_{norm}	d_{False}
CiteSeer	GAT	69.0 ± 1.0	10.5 ± 1.7		
	GCN	69.2 ± 0.7	7.1 ± 1.0		
Pubmed	GAT	75.7 ± 0.6	3.7 ± 1.4		
	GCN	76.8 ± 0.5	2.4 ± 0.7		
CS	GAT	90.7 ± 0.5	3.7 ± 0.5		
	GCN	90.7 ± 0.5	3.3 ± 0.6		
Physics	GAT	92.0 ± 0.7	3.8 ± 0.8		
	GCN	92.7 ± 0.3	1.6 ± 0.4		
Computers	GAT	81.0 ± 1.5	9.5 ± 2.2		
	GCN	81.2 ± 0.9	9.9 ± 1.9		
Photo	GAT	90.3 ± 0.8	4.4 ± 1.1		
	GCN	90.8 ± 0.5	3.7 ± 0.8		
WikiCS	GAT	79.6 ± 0.3	3.8 ± 0.5		
	GCN	79.4 ± 0.2	3.3 ± 0.4		

GNN predictions are significantly unstable

- Train 50 different initializations and compare pairwise

Dataset	Model	Accuracy	d	d_{norm}	d_{False}
CiteSeer	GAT	69.0 ± 1.0	10.5 ± 1.7	15.4 ± 2.5	
	GCN	69.2 ± 0.7	7.1 ± 1.0	10.3 ± 1.6	
Pubmed	GAT	75.7 ± 0.6	3.7 ± 1.4	6.4 ± 2.7	
	GCN	76.8 ± 0.5	2.4 ± 0.7	4.1 ± 1.4	
CS	GAT	90.7 ± 0.5	3.7 ± 0.5	17.3 ± 2.0	
	GCN	90.7 ± 0.5	3.3 ± 0.6	15.4 ± 2.7	
Physics	GAT	92.0 ± 0.7	3.8 ± 0.8	19.7 ± 4.2	
	GCN	92.7 ± 0.3	1.6 ± 0.4	8.6 ± 2.7	
Computers	GAT	81.0 ± 1.5	9.5 ± 2.2	21.6 ± 5.6	
	GCN	81.2 ± 0.9	9.9 ± 1.9	24.2 ± 4.9	
Photo	GAT	90.3 ± 0.8	4.4 ± 1.1	18.9 ± 4.9	
	GCN	90.8 ± 0.5	3.7 ± 0.8	17.5 ± 3.7	
WikiCS	GAT	79.6 ± 0.3	3.8 ± 0.5	8.6 ± 1.3	
	GCN	79.4 ± 0.2	3.3 ± 0.4	7.6 ± 1.0	

GNN predictions are significantly unstable

UNIVERSITÄT PASSAU

- Train 50 different initializations and compare pairwise

Dataset	Model	Accuracy	d	d_{norm}	d_{False}
CiteSeer	GAT	69.0 ± 1.0	10.5 ± 1.7	15.4 ± 2.5	22.3 ± 3.8
	GCN	69.2 ± 0.7	7.1 ± 1.0	10.3 ± 1.6	15.1 ± 2.4
Pubmed	GAT	75.7 ± 0.6	3.7 ± 1.4	6.4 ± 2.7	8.0 ± 3.3
	GCN	76.8 ± 0.5	2.4 ± 0.7	4.1 ± 1.4	5.6 ± 2.2
CS	GAT	90.7 ± 0.5	3.7 ± 0.5	17.3 ± 2.0	22.0 ± 3.6
	GCN	90.7 ± 0.5	3.3 ± 0.6	15.4 ± 2.7	19.9 ± 4.1
Physics	GAT	92.0 ± 0.7	3.8 ± 0.8	19.7 ± 4.2	25.7 ± 6.4
	GCN	92.7 ± 0.3	1.6 ± 0.4	8.6 ± 2.7	12.2 ± 4.3
Computers	GAT	81.0 ± 1.5	9.5 ± 2.2	21.6 ± 5.6	29.6 ± 7.3
	GCN	81.2 ± 0.9	9.9 ± 1.9	24.2 ± 4.9	31.9 ± 6.0
Photo	GAT	90.3 ± 0.8	4.4 ± 1.1	18.9 ± 4.9	26.0 ± 6.9
	GCN	90.8 ± 0.5	3.7 ± 0.8	17.5 ± 3.7	24.1 ± 5.5
WikiCS	GAT	79.6 ± 0.3	3.8 ± 0.5	8.6 ± 1.3	11.7 ± 1.8
	GCN	79.4 ± 0.2	3.3 ± 0.4	7.6 ± 1.0	10.1 ± 1.4

GNN predictions are significantly unstable

- Train 50 different initializations and compare pairwise

Dataset	Model	Accuracy	d	d_{norm}	d_{False}
CiteSeer	GAT	69.0 ± 1.0	10.5 ± 1.7	15.4 ± 2.5	22.3 ± 3.8
	GCN	69.2 ± 0.7	7.1 ± 1.0	10.3 ± 1.6	15.1 ± 2.4
Pubmed	GAT	75.7 ± 0.6	3.7 ± 1.4	6.4 ± 2.7	8.0 ± 3.3
	GCN	76.8 ± 0.5	2.4 ± 0.7	4.1 ± 1.4	5.6 ± 2.2
CS	GAT	90.7 ± 0.5	3.7 ± 0.5	17.3 ± 2.0	22.0 ± 3.6
	GCN	90.7 ± 0.5	3.3 ± 0.6	15.4 ± 2.7	19.9 ± 4.1
Physics	GAT	92.0 ± 0.7	3.8 ± 0.8	19.7 ± 4.2	25.7 ± 6.4
	GCN	92.7 ± 0.3	1.6 ± 0.4	8.6 ± 2.7	12.2 ± 4.3
Computers	GAT	81.0 ± 1.5	9.5 ± 2.2	21.6 ± 5.6	29.6 ± 7.3
	GCN	81.2 ± 0.9	9.9 ± 1.9	24.2 ± 4.9	31.9 ± 6.0
Photo	GAT	90.3 ± 0.8	4.4 ± 1.1	18.9 ± 4.9	26.0 ± 6.9
	GCN	90.8 ± 0.5	3.7 ± 0.8	17.5 ± 3.7	24.1 ± 5.5
WikiCS	GAT	79.6 ± 0.3	3.8 ± 0.5	8.6 ± 1.3	11.7 ± 1.8
	GCN	79.4 ± 0.2	3.3 ± 0.4	7.6 ± 1.0	10.1 ± 1.4

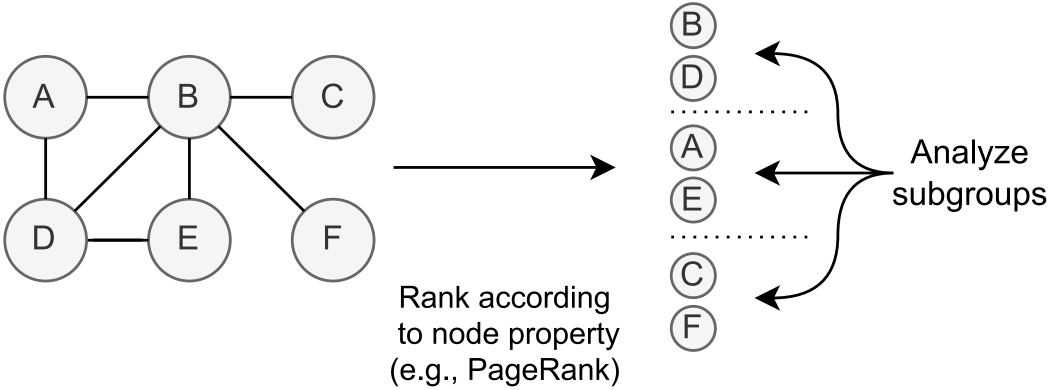
Outline

Setup

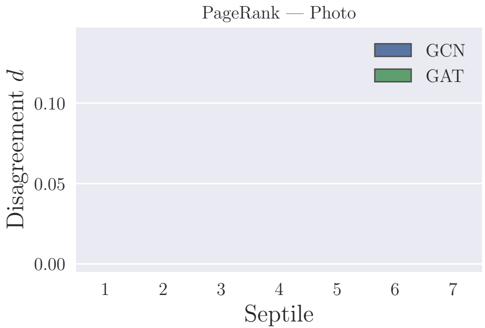
- Centrality, clustering etc. could influence stability
- PageRank, clustering coefficient, k-core, class label

Setup

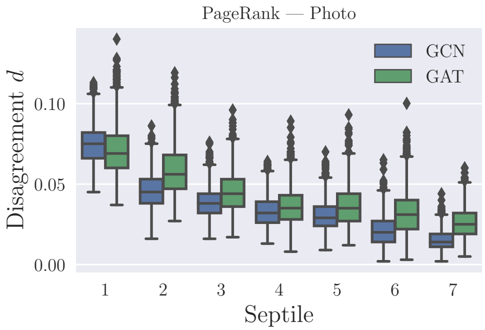
- Centrality, clustering etc. could influence stability
- PageRank, clustering coefficient, k-core, class label



Results: PageRank

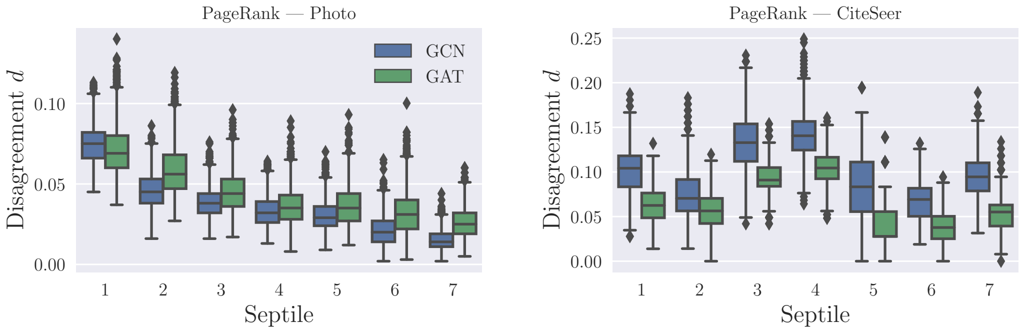


Results: PageRank



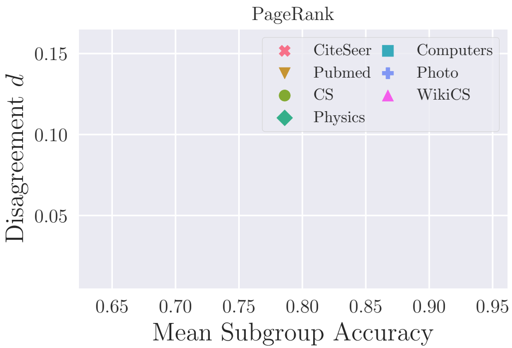
- Central nodes are more stable

Results: PageRank

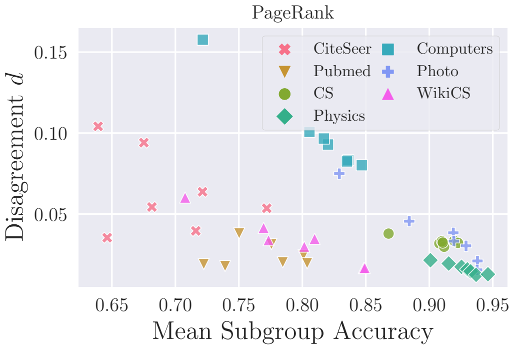


- Central nodes are more stable
- Less clear effects for other properties

Error Rate and Disagreement are correlated



Error Rate and Disagreement are correlated

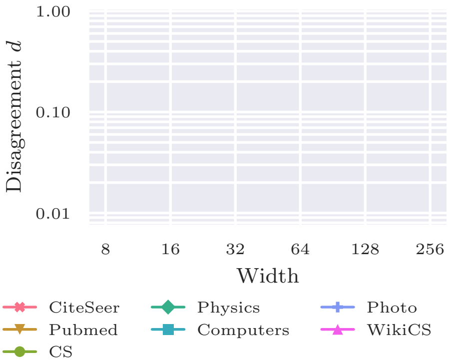


- Central nodes are more stably predicted
- Normalized disagreement explains this effect

Outline

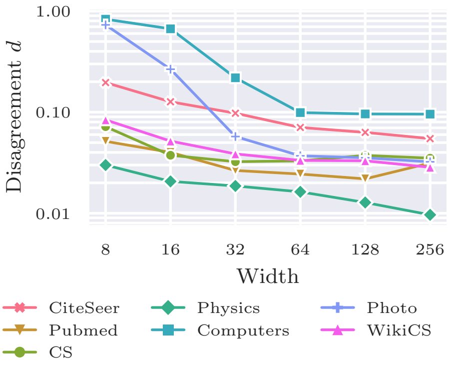
Influence of hyperparameters

- Change configuration of model along one axis, e.g., width



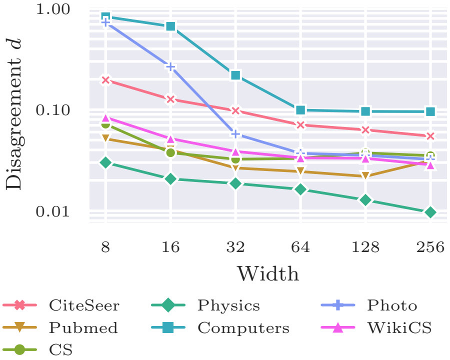
Influence of hyperparameters

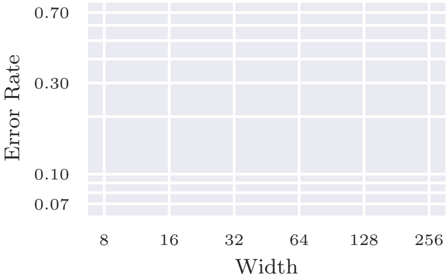
- Change configuration of model along one axis, e.g., width



Influence of hyperparameters

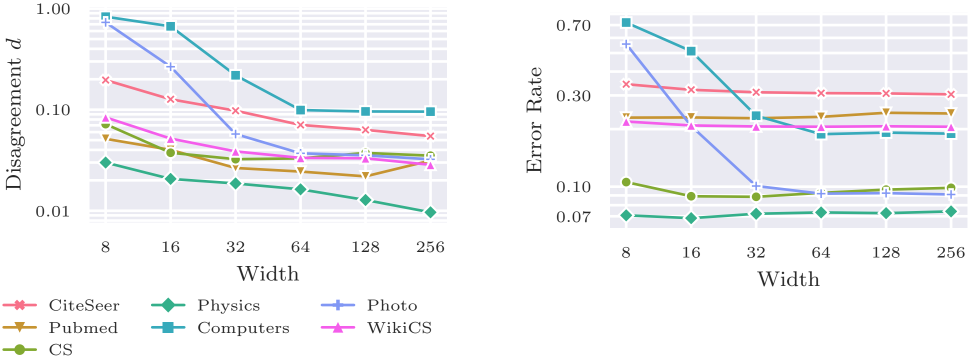
- Change configuration of model along one axis, e.g., width





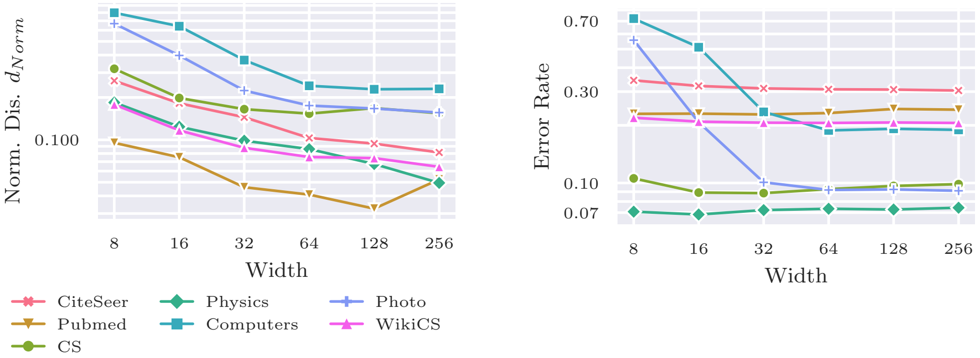
Influence of hyperparameters

- Change configuration of model along one axis, e.g., width



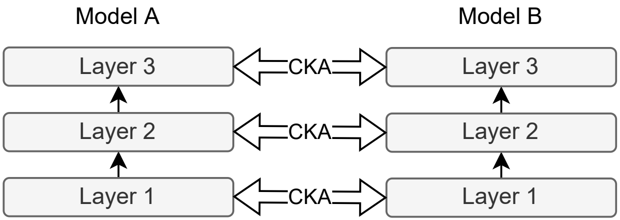
Influence of hyperparameters

- Change configuration of model along one axis, e.g., width



Outline

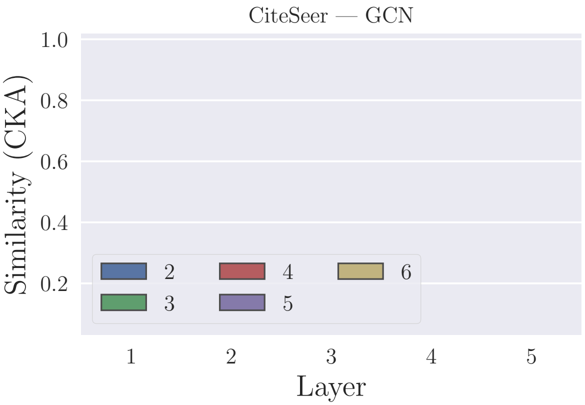
Where do instabilities arise?



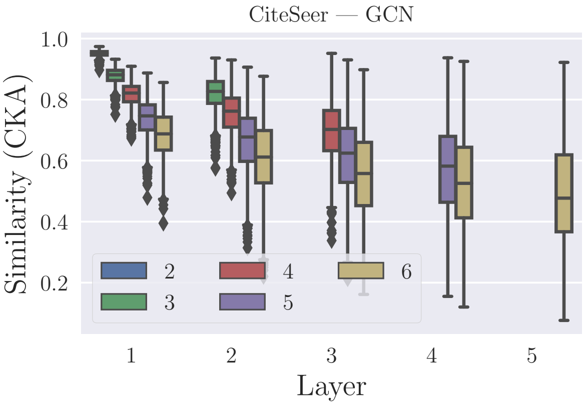
- Compare representations of layers over varying depth with Centered Kernel Alignment (CKA)[6]

[6] Kornblith et al. 2019.

Where do instabilities arise?



Where do instabilities arise?



- Deep layers and deep models are less stable

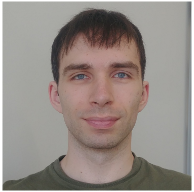
Outline

1. Introduction
2. Do GNNs exhibit significant prediction instability?
3. What is the influence of node properties?
4. What is the influence of hyperparameters?
5. Where do instabilities arise?
6. Summary

Summary

UNIVERSITÄT PASSAU

- Deep neural networks are not stable: GCN and GAT are no exception
- Data properties influence stability
- Selecting good hyperparameters can positively influence stability
- Internal representations mirror instability of predictions



Max Klabunde

Florian Lemmerich

Github: mklabunde/gnn-prediction-instability

References I

- Kipf, Thomas N. and Max Welling (2017). "Semi-Supervised Classification with Graph Convolutional Networks." In: *International Conference on Learning Representations.*
- Kornblith, Simon et al. (2019). "Similarity of Neural Network Representations Revisited." In: *Proceedings of the 36th International Conference on Machine Learning.*

References II

- Schumacher, Tobias et al. (2021). "The effects of randomness on the stability of node embeddings." In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.
- Summers, Cecilia and Michael J Dinneen (2021). "Nondeterminism and instability in neural network optimization." In: *International Conference on Machine Learning*.

References III

- Veličković, Petar et al. (2018). "Graph Attention Networks." In: *International Conference on Learning Representations*. accepted as poster.
- Wang, Chenxu et al. (2022). "Towards Understanding the Instability of Network Embedding." In: *IEEE Transactions on Knowledge and Data Engineering*.
- Zhuang, Donglin et al. (2022). "Randomness in neural network training: Characterizing the impact of tooling." In: *Proceedings of Machine Learning and Systems*.