Supplementary Material: On the Prediction Instability of Graph Neural Networks

Max Klabunde[0000−0002−7805−4725] (✉) and
Florian Lemmerich[0000−0001−7620−1376]

Faculty of Computer Science and Mathematics, University of Passau, Germany
{max.klabunde,florian.lemmerich}@uni-passau.de

With this supplementary material, we provide basic dataset statistics, mathematical detail, and additional experimental results.

A Dataset Statistics

The statistics are given in Table 1.

Table 1: Dataset statistics

Dataset	#Nodes	#Edges	#Classes
Citeseer	3327	9104	6
Pubmed	19717	88684	3
CS	18333	163788	15
Physics	34493	495924	5
Computers	13752	491722	10
Photo	7650	238162	8
WikiCS	11701	297110	10

B Mathematical Definitions

Here, we present the mathematical definitions of PageRank, clustering coefficient, and centered kernel alignment. For the k-core, we directly refer to Seidman [4, Section 2].

PageRank. The PageRank vector π is the solution to the eigenvector problem [3,2]:

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \bar{\mathbf{P}}, \ \boldsymbol{\pi}^T \mathbf{e} = 1, \tag{1}$$

where $\bar{\mathbf{P}}$ is the modified stochastic transition matrix on a graph and $\mathbf{e}$ is a vector of ones. $\bar{\mathbf{P}}$ is defined as follows:

$$\bar{\mathbf{P}} = \alpha \mathbf{P} + (1 - \alpha)\mathbf{E}, \tag{2}$$

where $\mathbf{P}$ is the original stochastic transition matrix (with zero rows replaced $\mathbf{e}^T/n$, with n being the number of nodes) and $\mathbf{E} = \mathbf{e}\mathbf{e}^T/n$. We use networkx 2.6.2 to compute the PageRank in our experiments with the default value $\alpha = 0.85$.

Clustering Coefficient. The clustering coefficient $CC(u)$ of a node u is defined as the number of triangles $T(u)$ that the node is part of divided by the theoretically possible number of triangles the node could be part of[1]:

$$CC(u) = \frac{T(u)}{deg(u)(deg(u) - 1)}, \tag{3}$$

where $deg(u)$ is the degree of node u [5].

Centered Kernel Alignment. Let the matrices $\mathbf{X} \in \mathbb{R}^{n \times p_1}, \mathbf{Y} \in \mathbb{R}^{n \times p_2}$ be activations on n instances of p_1 and p_2 neurons, respectively, with zero-centered columns. Then linear CKA is defined as follows, where $\| \cdot \|_F$ denotes the Frobenius norm [1]:

$$\text{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\|\mathbf{Y}^T\mathbf{X}\|_F^2}{\|\mathbf{X}^T\mathbf{X}\|_F \|\mathbf{X}^T\mathbf{X}\|_F}. \tag{4}$$

C Detailed Results

Additional to the results presented in the main paper, we disclose in this supplementary material the experimental outcomes for experiments trained with fixed seeds and CPU-training as well as the results for experiments according to other instability measures.

C.1 Overall Prediction Instability of GNNs

We provide the comparison between GPU without deterministic CUDA operations and CPU training, and between fixed and varying initialization in Table 2 and Table 3. While differences based on the compute platform exist, they are small compared to varying the initialization.

C.2 The Effect of Node Properties

For each of the properties, we show the prediction stability for the different septiles in the following figures.

- PageRank
  - Absolute disagreement: Figure 1
  - True disagreement: Figure 2
  - False disagreement: Figure 3

[1] As defined in the networkx 2.6.2 documentation.

Table 2: Prediction disagreement (in %) and standard deviation on different datasets. "Fixed" denotes that the random seed of the algorithms had been fixed, "CPU" describes that the model was trained on a CPU instead of GPU.

Dataset	Model	Accuracy	d	d_{norm}	d_{True}
CiteSeer	GAT (fixed)	70.10 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GAT	69.04 ± 0.88	10.31 ± 1.70	15.30 ± 2.61	5.11 ± 1.28
	GCN (fixed)	69.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	69.11 ± 0.61	7.32 ± 1.04	10.85 ± 1.77	3.58 ± 0.86
Pubmed	GAT (fixed)	76.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GAT	75.69 ± 0.60	3.75 ± 1.30	6.41 ± 2.62	2.38 ± 1.00
	GCN (fixed)	76.10 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	76.78 ± 0.55	2.57 ± 0.78	4.25 ± 1.50	1.57 ± 0.70
CS	GAT (CPU)	90.62 ± 0.38	3.62 ± 0.45	17.40 ± 2.10	1.74 ± 0.37
	GAT (fixed)	91.25 ± 0.04	0.44 ± 0.15	2.27 ± 0.79	0.22 ± 0.08
	GAT (fixed, CPU)	91.36 ± 0.00	0.03 ± 0.06	0.17 ± 0.33	0.01 ± 0.03
	GAT	90.70 ± 0.43	3.64 ± 0.46	17.52 ± 1.91	1.74 ± 0.40
	GCN (fixed)	91.21 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	90.80 ± 0.44	3.28 ± 0.59	15.51 ± 2.64	1.57 ± 0.46
Physics	GAT (CPU)	91.77 ± 0.65	3.89 ± 0.76	19.87 ± 3.93	1.88 ± 0.61
	GAT (fixed)	91.99 ± 0.03	0.13 ± 0.08	0.59 ± 0.46	0.07 ± 0.05
	GAT (fixed, CPU)	92.16 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GAT	91.89 ± 0.67	3.84 ± 0.78	19.87 ± 4.07	1.86 ± 0.64
	GCN (fixed)	92.55 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	92.67 ± 0.34	1.64 ± 0.42	8.76 ± 2.69	0.79 ± 0.33
Computers	GAT (CPU)	81.03 ± 1.50	9.22 ± 2.21	20.75 ± 5.52	4.61 ± 1.79
	GAT (fixed)	81.60 ± 0.61	6.00 ± 2.37	14.73 ± 5.99	3.06 ± 1.35
	GAT (fixed, CPU)	79.83 ± 0.01	0.10 ± 0.20	0.24 ± 0.47	0.05 ± 0.11
	GAT	80.85 ± 1.39	8.99 ± 2.27	20.31 ± 5.48	4.47 ± 1.75
	GCN (fixed)	80.08 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	80.96 ± 1.21	10.50 ± 2.34	24.94 ± 5.28	5.08 ± 1.61
Photo	GAT (CPU)	90.45 ± 0.67	4.01 ± 0.97	17.59 ± 4.43	1.83 ± 0.71
	GAT (fixed)	90.87 ± 0.46	2.57 ± 0.72	11.49 ± 3.28	1.16 ± 0.49
	GAT (fixed, CPU)	89.57 ± 0.03	0.68 ± 1.35	3.16 ± 6.32	0.31 ± 0.63
	GAT	90.39 ± 0.62	3.93 ± 0.86	17.40 ± 4.02	1.81 ± 0.64
	GCN (fixed)	90.34 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	90.76 ± 0.50	3.84 ± 0.82	18.29 ± 4.04	1.70 ± 0.54
WikiCS	GAT (CPU)	79.58 ± 0.23	3.73 ± 0.54	8.54 ± 1.23	1.71 ± 0.33
	GAT (fixed)	79.89 ± 0.18	1.84 ± 1.02	4.10 ± 2.23	0.84 ± 0.50
	GAT (fixed, CPU)	79.88 ± 0.07	0.55 ± 1.10	1.26 ± 2.52	0.26 ± 0.51
	GAT	79.58 ± 0.23	3.72 ± 0.52	8.54 ± 1.24	1.72 ± 0.32
	GCN (fixed)	79.27 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	79.42 ± 0.20	3.27 ± 0.42	7.43 ± 1.01	1.54 ± 0.27

Table 3: Continuation of Table 2. "Fixed" denotes that the random seed of the algorithms had been fixed, "CPU" describes that the model was trained on a CPU instead of GPU.

Dataset	Model	Accuracy	d_{False}	MAE
CiteSeer	GAT (fixed)	70.10 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GAT	69.04 ± 0.88	21.83 ± 3.76	3.32 ± 0.53
	GCN (fixed)	69.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	69.11 ± 0.61	15.66 ± 2.50	2.97 ± 0.36
Pubmed	GAT (fixed)	76.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GAT	75.69 ± 0.60	7.95 ± 3.18	2.25 ± 0.69
	GCN (fixed)	76.10 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	76.78 ± 0.55	5.84 ± 2.32	2.49 ± 0.94
CS	GAT (CPU)	90.62 ± 0.38	21.78 ± 3.26	0.69 ± 0.12
	GAT (fixed)	91.25 ± 0.04	2.74 ± 0.97	0.07 ± 0.02
	GAT (fixed, CPU)	91.36 ± 0.00	0.17 ± 0.34	0.00 ± 0.01
	GAT	90.70 ± 0.43	22.04 ± 3.33	0.68 ± 0.10
	GCN (fixed)	91.21 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	90.80 ± 0.44	20.02 ± 4.08	0.67 ± 0.21
Physics	GAT (CPU)	91.77 ± 0.65	26.01 ± 6.34	1.96 ± 0.39
	GAT (fixed)	91.99 ± 0.03	0.84 ± 0.61	0.06 ± 0.04
	GAT (fixed, CPU)	92.16 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GAT	91.89 ± 0.67	25.93 ± 6.21	2.00 ± 0.52
	GCN (fixed)	92.55 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	92.67 ± 0.34	12.28 ± 4.22	1.29 ± 0.45
Computers	GAT (CPU)	81.03 ± 1.50	28.75 ± 7.44	2.18 ± 0.42
	GAT (fixed)	81.60 ± 0.61	18.99 ± 7.46	1.50 ± 0.66
	GAT (fixed, CPU)	79.83 ± 0.01	0.28 ± 0.56	0.03 ± 0.05
	GAT	80.85 ± 1.39	27.92 ± 7.37	2.21 ± 0.50
	GCN (fixed)	80.08 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	80.96 ± 1.21	33.34 ± 6.93	2.43 ± 0.52
Photo	GAT (CPU)	90.45 ± 0.67	24.38 ± 6.37	1.48 ± 0.38
	GAT (fixed)	90.87 ± 0.46	16.48 ± 5.36	0.97 ± 0.25
	GAT (fixed, CPU)	89.57 ± 0.03	3.80 ± 7.60	0.25 ± 0.51
	GAT	90.39 ± 0.62	23.68 ± 5.80	1.42 ± 0.30
	GCN (fixed)	90.34 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	90.76 ± 0.50	24.69 ± 5.49	1.47 ± 0.26
WikiCS	GAT (CPU)	79.58 ± 0.23	11.56 ± 1.81	0.87 ± 0.12
	GAT (fixed)	79.89 ± 0.18	5.76 ± 3.18	0.46 ± 0.28
	GAT (fixed, CPU)	79.88 ± 0.07	1.72 ± 3.44	0.12 ± 0.25
	GAT	79.58 ± 0.23	11.54 ± 1.75	0.90 ± 0.14
	GCN (fixed)	79.27 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	GCN	79.42 ± 0.20	9.91 ± 1.40	0.73 ± 0.09

  - Normalized disagreement: Figure 4
  - MAE: Figure 5
- Clustering Coefficient
  - Absolute disagreement: Figure 6
  - True disagreement: Figure 7
  - False disagreement: Figure 8
  - Normalized disagreement: Figure 9
  - MAE: Figure 10
- K-core
  - Absolute disagreement: Figure 11
  - True disagreement: Figure 12
  - False disagreement: Figure 13
  - Normalized disagreement: Figure 14
  - MAE: Figure 15

For each of the properties, we show the relation to subgroup performance for GAT and GCN in the following figures. Each figure includes plots for all used stability measures.

- PageRank:
  - GCN: Figure 16
  - GAT: Figure 17
- Clustering Coefficient
  - GCN: Figure 18
  - GAT: Figure 19
- K-core
  - GCN: Figure 20
  - GAT: Figure 21
- Class label
  - GCN: Figure 22
  - GAT: Figure 23

C.3 The Effect of Model Design and Training Setup

For each of the experiments, we show the results for GAT and GCN in the following figures. Since the standard deviations make the plots difficult to read in some cases, we provide versions with and without information about the standard deviation (SD). Each figure includes plots for all used stability measures.

- Training Data
  - GCN: Figure 24, Figure 26 with SD
  - GAT: Figure 25, Figure 27 with SD
- Optimizer
  - GCN: Figure 28, Figure 30 with SD
  - GAT: Figure 29, Figure 31 with SD
- L2 Regularization
  - GCN: Figure 32, Figure 34 with SD

  - GAT: Figure 33, Figure 35 with SD
- Dropout
  - GCN: Figure 36, Figure 38 with SD
  - GAT: Figure 37, Figure 39 with SD
- Width
  - GCN: Figure 40, Figure 42 with SD
  - GAT: Figure 41, Figure 43 with SD
- Depth
  - GCN: Figure 44, Figure 46 with SD
  - GAT: Figure 45, Figure 47 with SD
- Combining Optimal Hyperparameters
  - GCN: Figure 48, Figure 50 with SD
  - GAT: Figure 49, Figure 51 with SD

C.4 Layer-wise Model Introspection

We provide plots for all datasets and both models in Figure 52 and Figure 53.

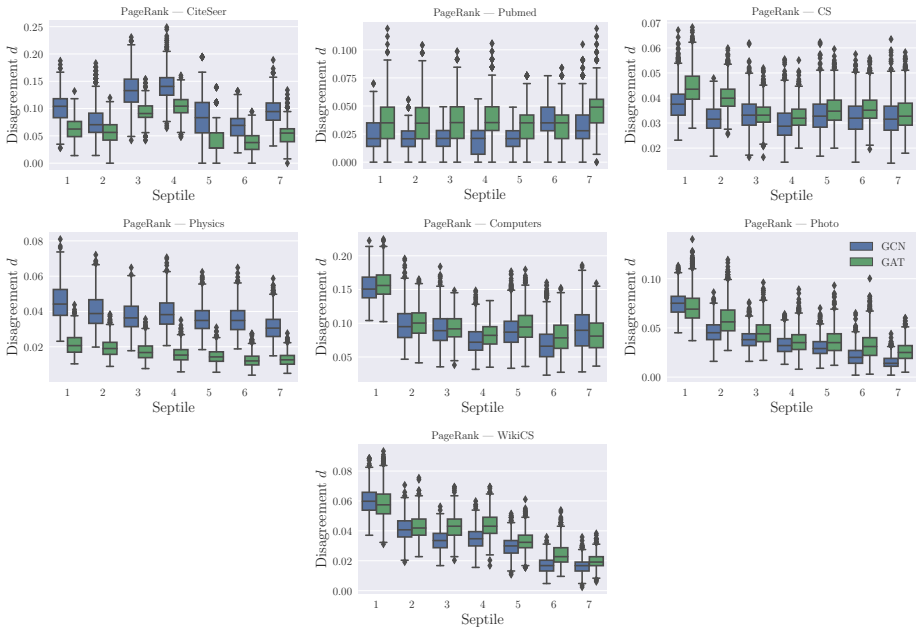

Fig. 1: Prediction disagreement for PageRank septiles.

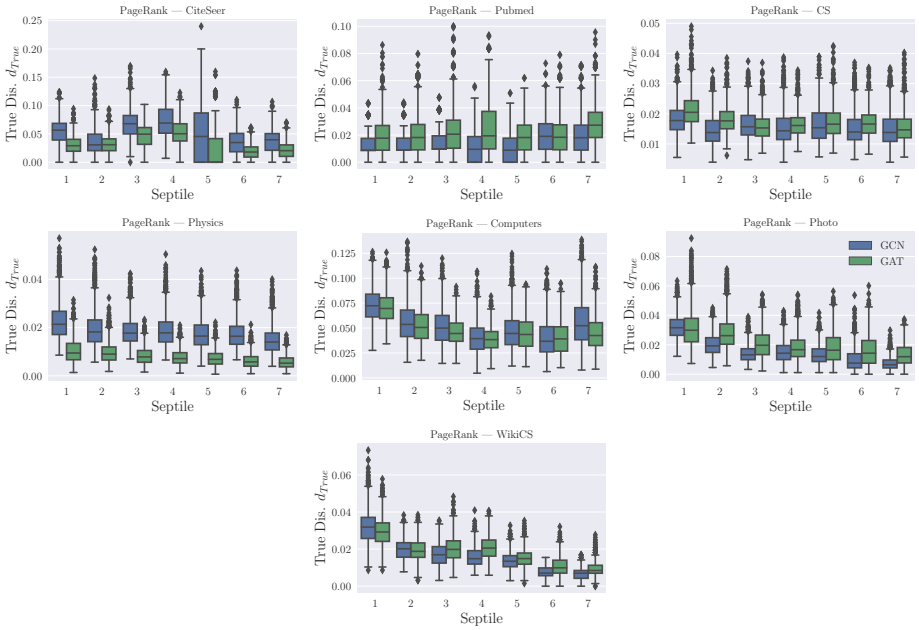

Fig. 2: True disagreement for PageRank septiles.

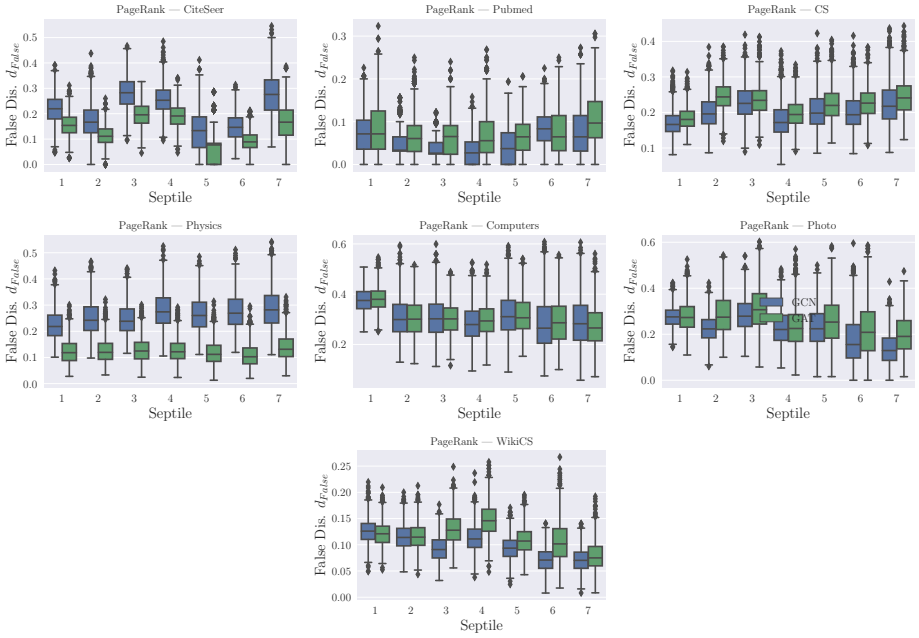

Fig. 3: False disagreement for PageRank septiles.

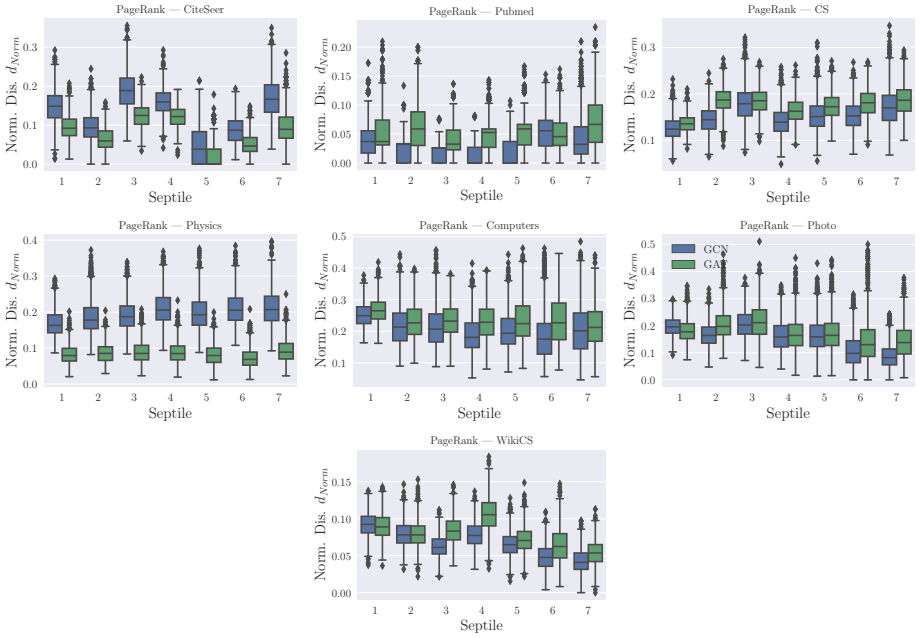

Fig. 4: Normalized disagreement for PageRank septiles.

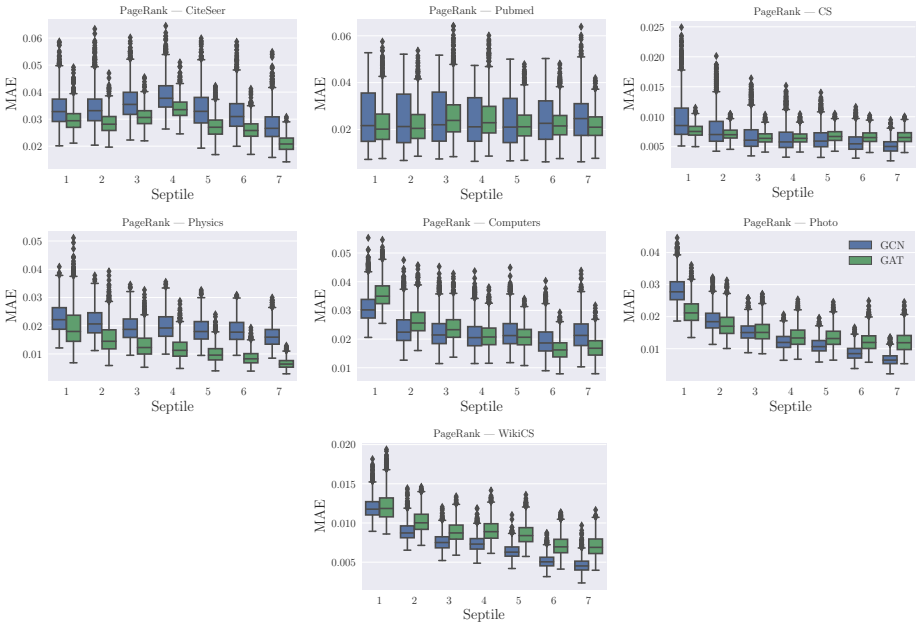

Fig. 5: MAE for PageRank septiles.

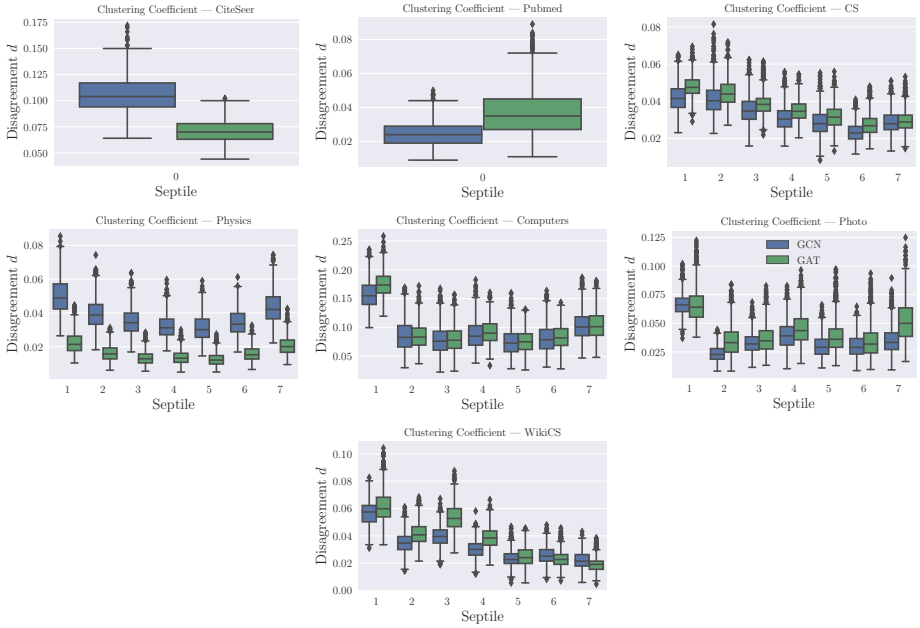

Fig. 6: Prediction disagreement for clustering coefficient septiles.

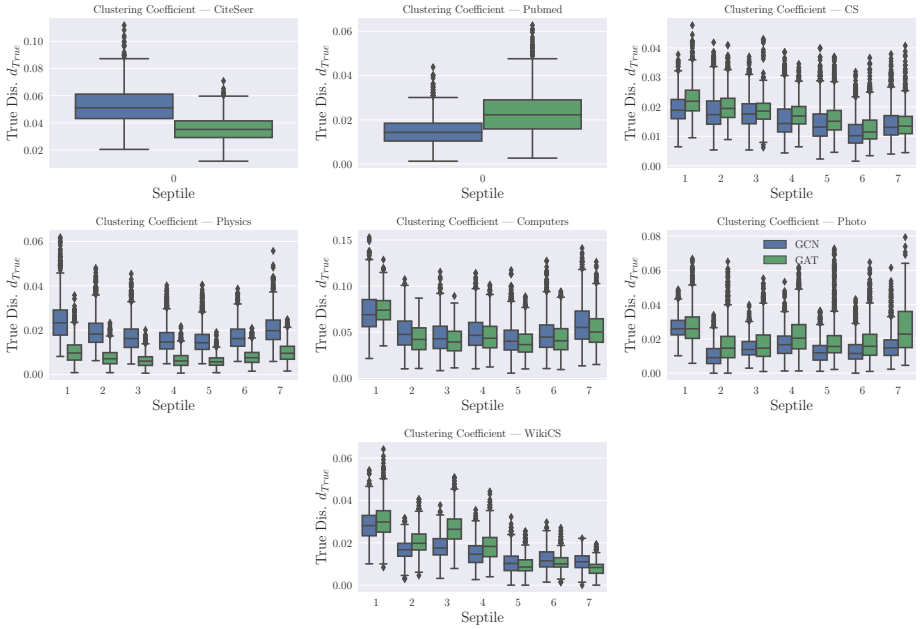

Fig. 7: True disagreement for clustering coefficient septiles.

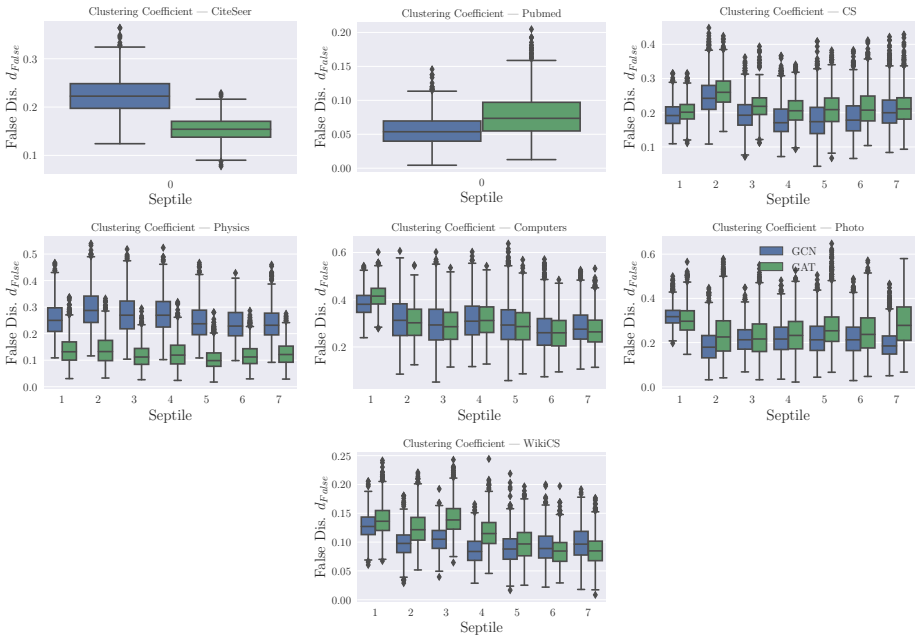

Fig. 8: False disagreement for clustering coefficient septiles.

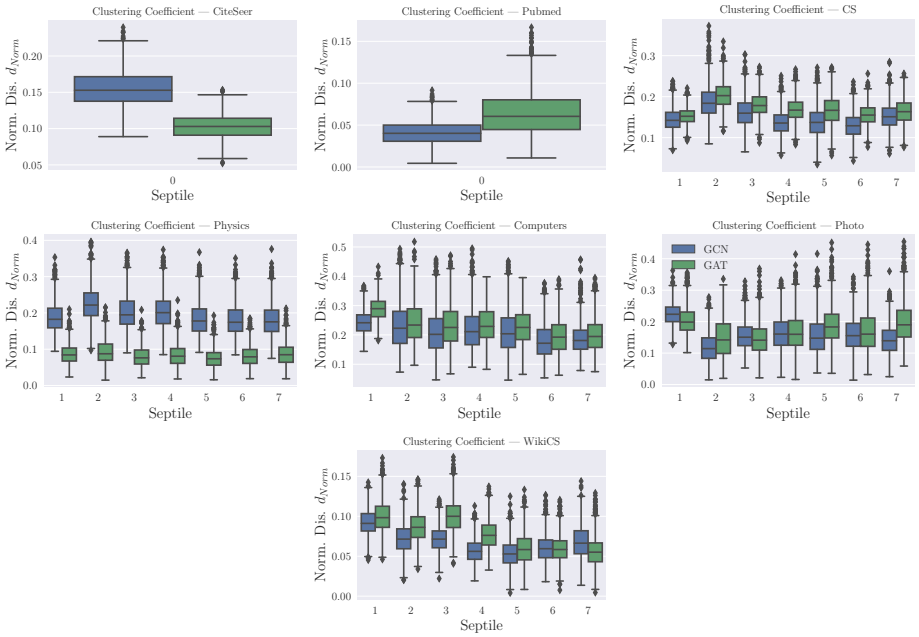

Fig. 9: Normalized disagreement for clustering coefficient septiles.

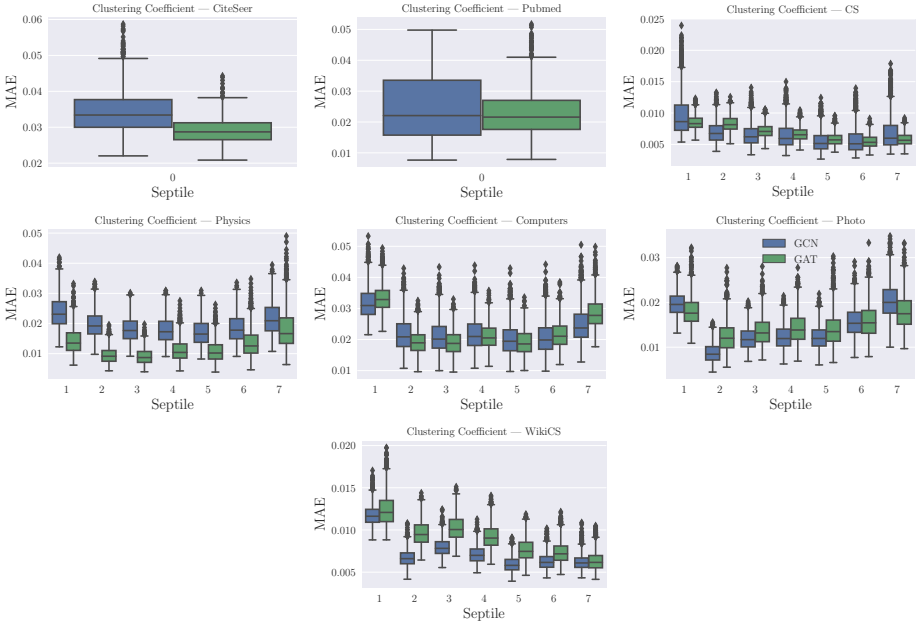

Fig. 10: MAE for clustering coefficient septiles.

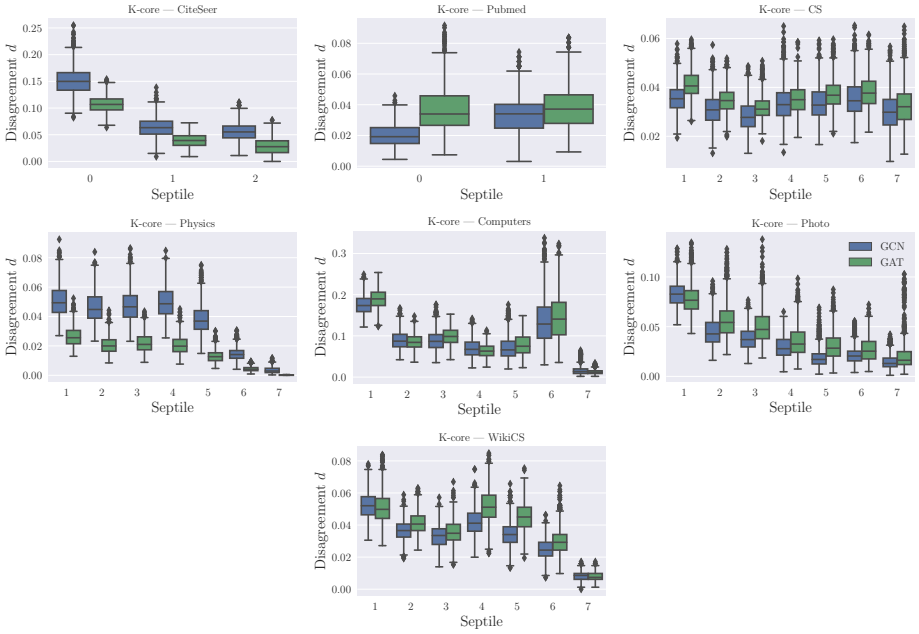

Fig. 11: Prediction disagreement for k-core septiles.

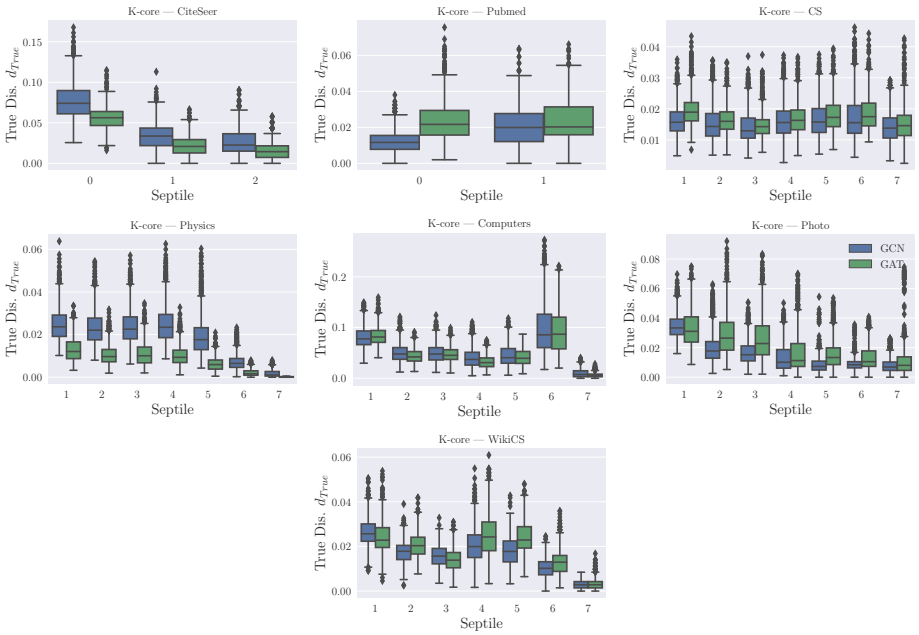

Fig. 12: True disagreement for k-core septiles.

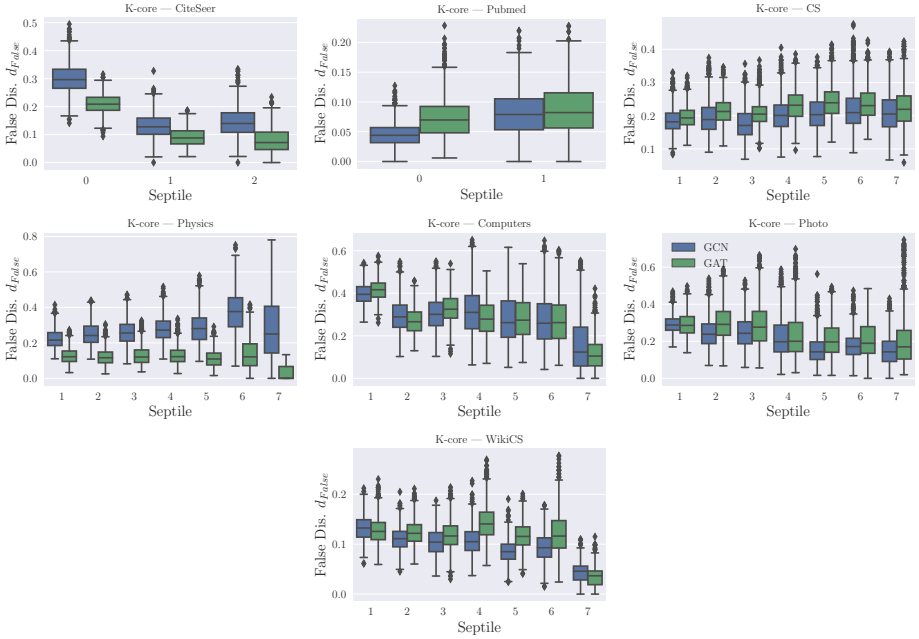

Fig. 13: False disagreement for k-core septiles.

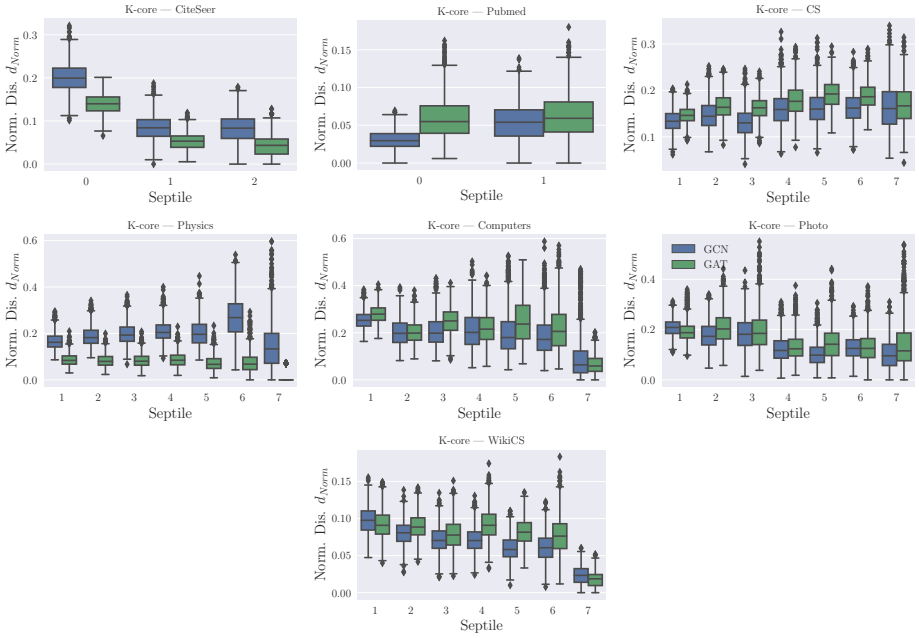

Fig. 14: Normalized disagreement for k-core septiles.

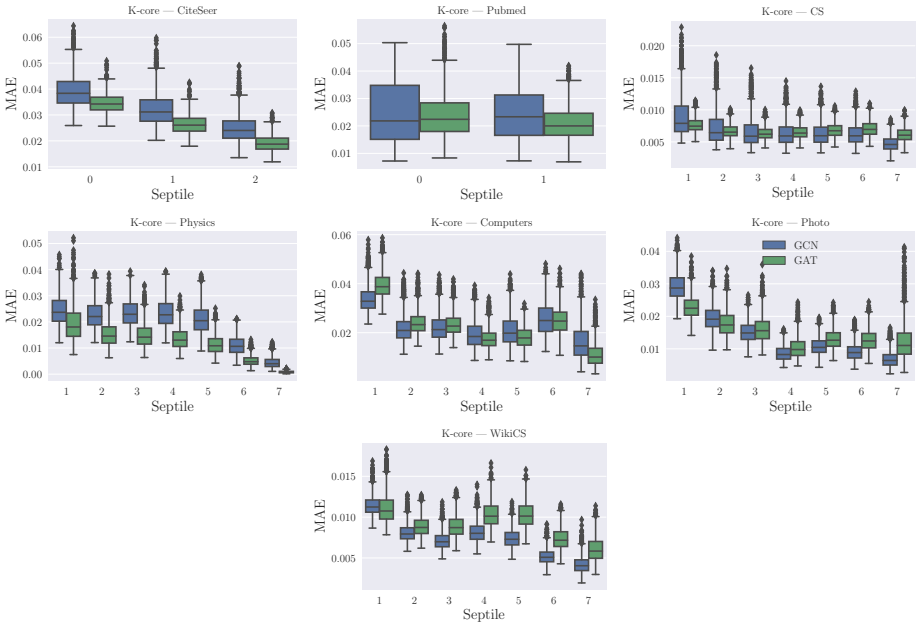

Fig. 15: MAE for k-core septiles.

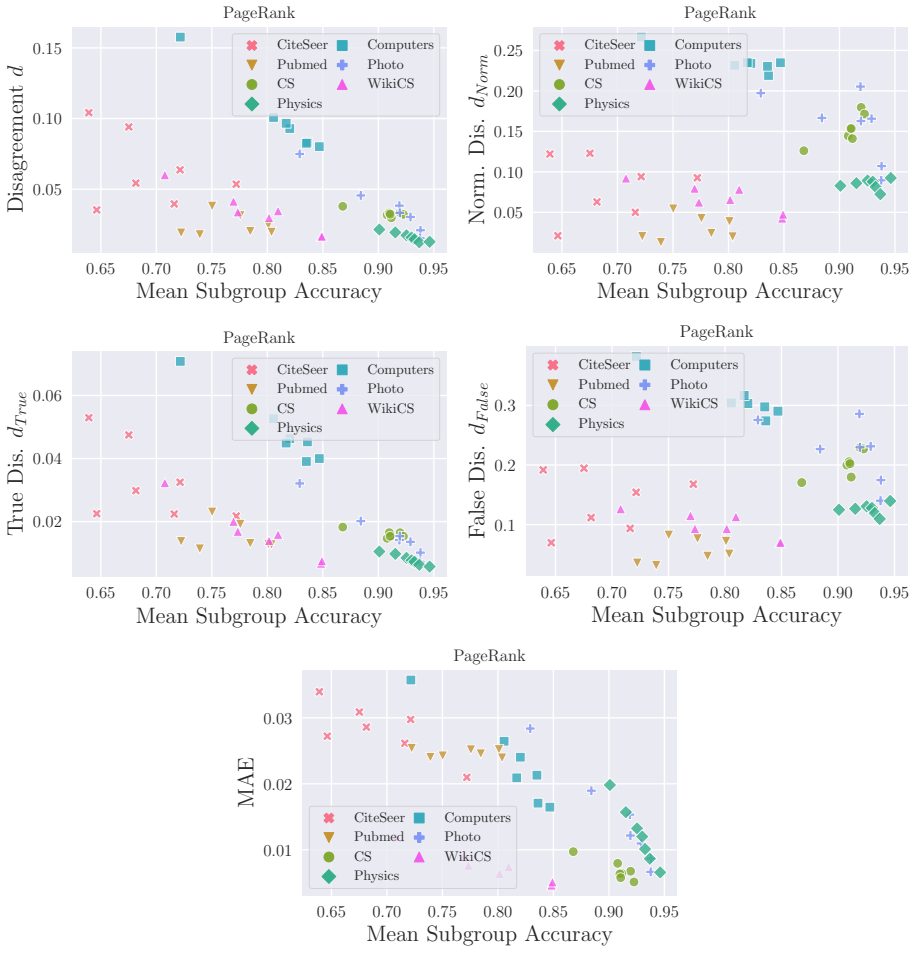

Fig. 16: Prediction disagreement in relation to performance of subgroups of PageRank for GCN

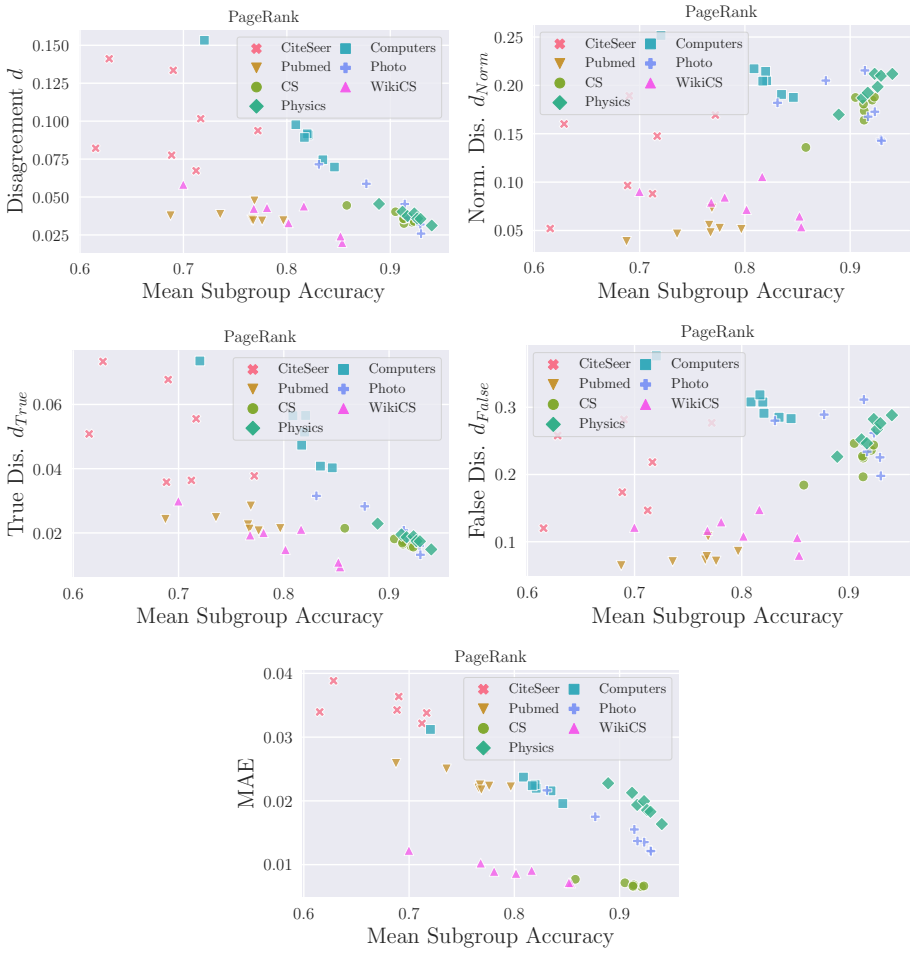

Fig. 17: Prediction disagreement in relation to performance of subgroups of PageRank for GAT

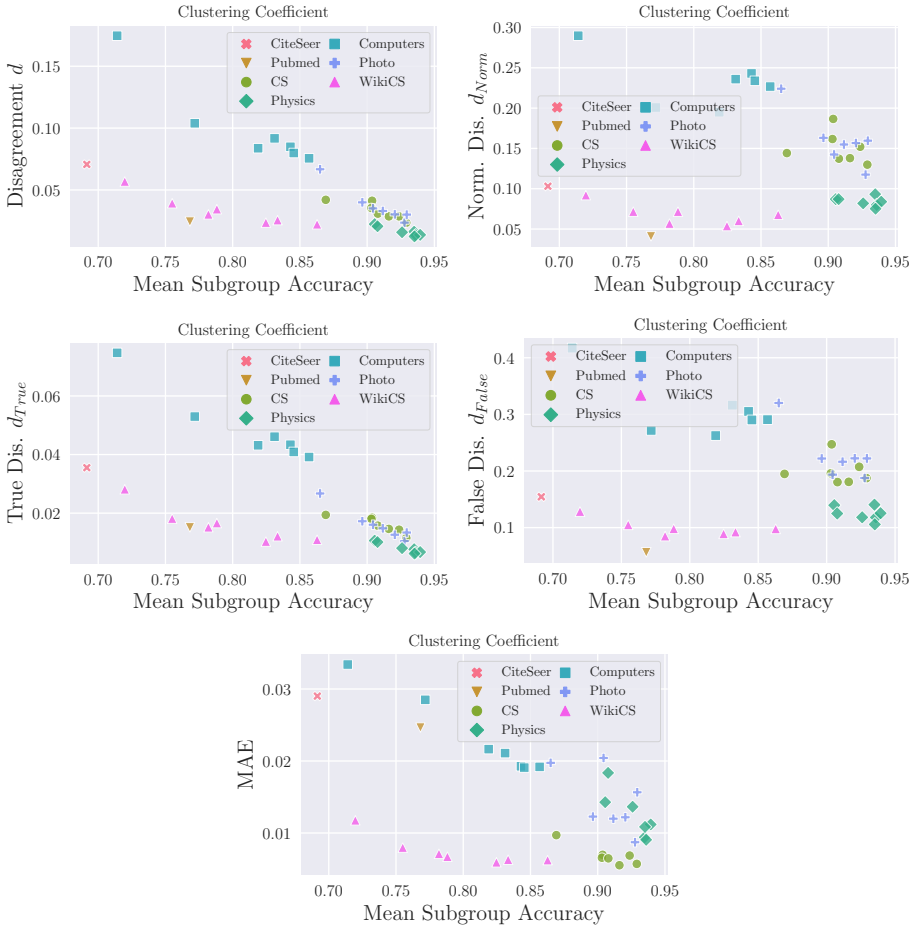

Fig. 18: Prediction disagreement in relation to performance of subgroups of clustering coefficient for GCN

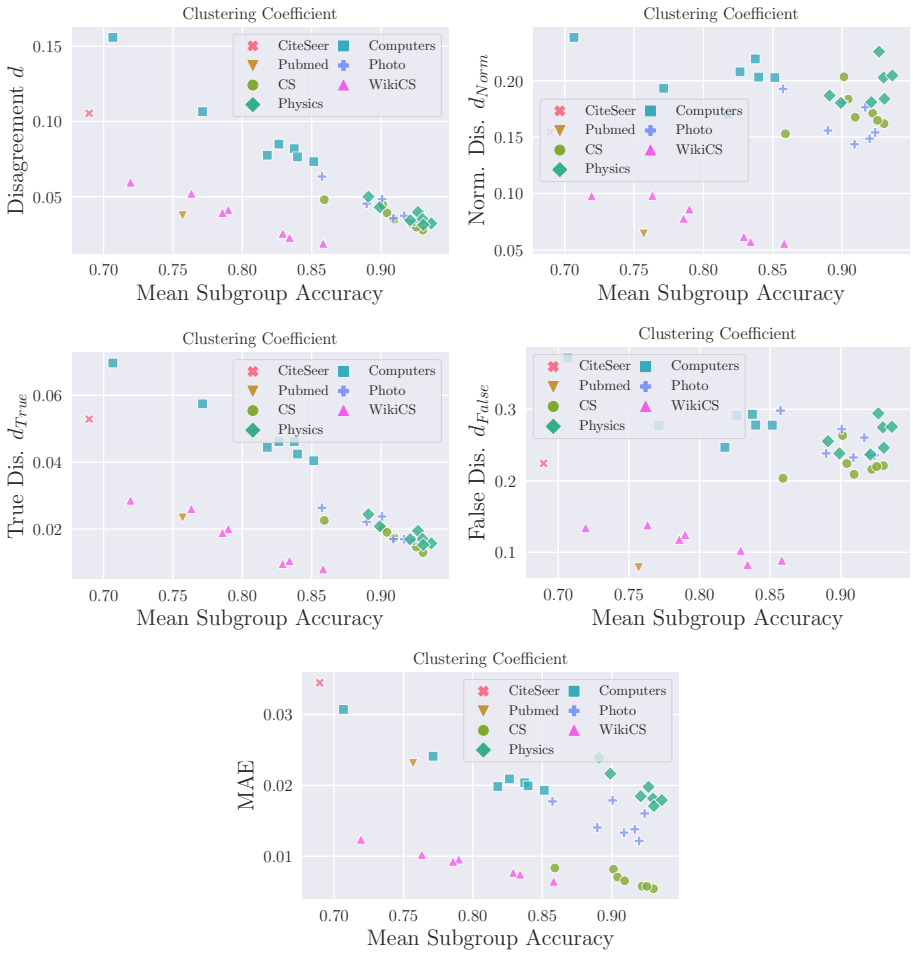

Fig. 19: Prediction disagreement in relation to performance of subgroups of clustering coefficient for GAT

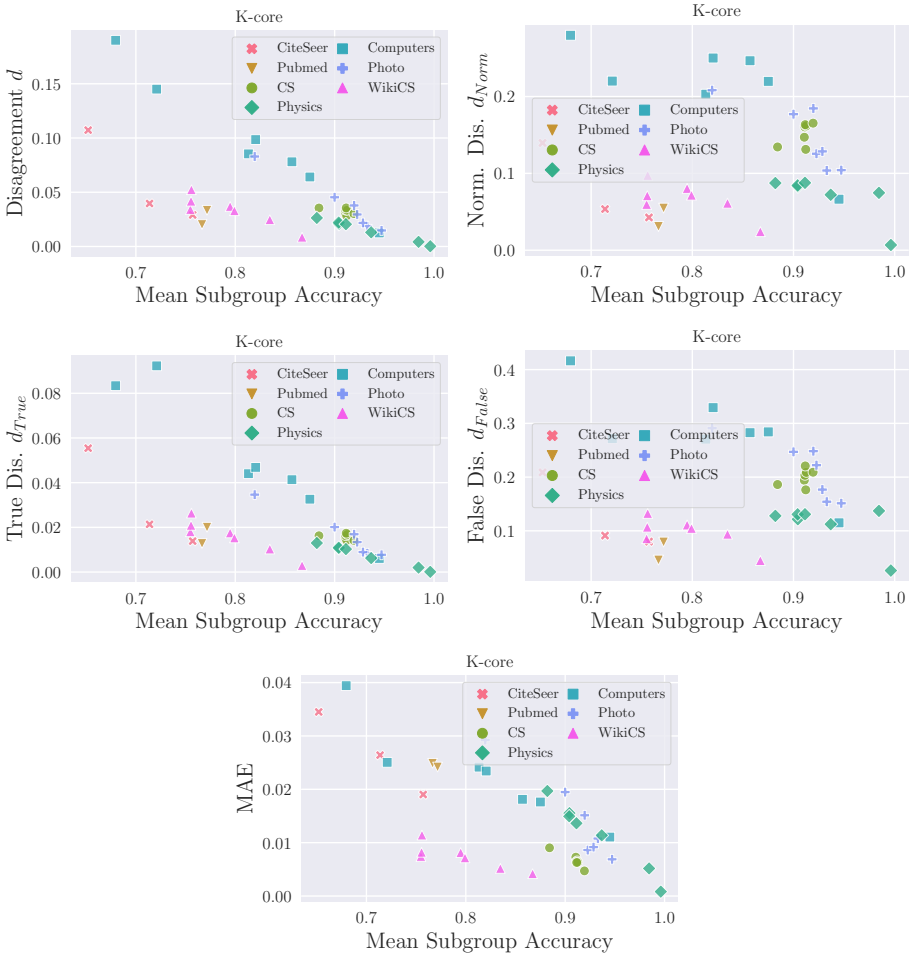

Fig. 20: Prediction disagreement in relation to performance of subgroups of k-core for GCN

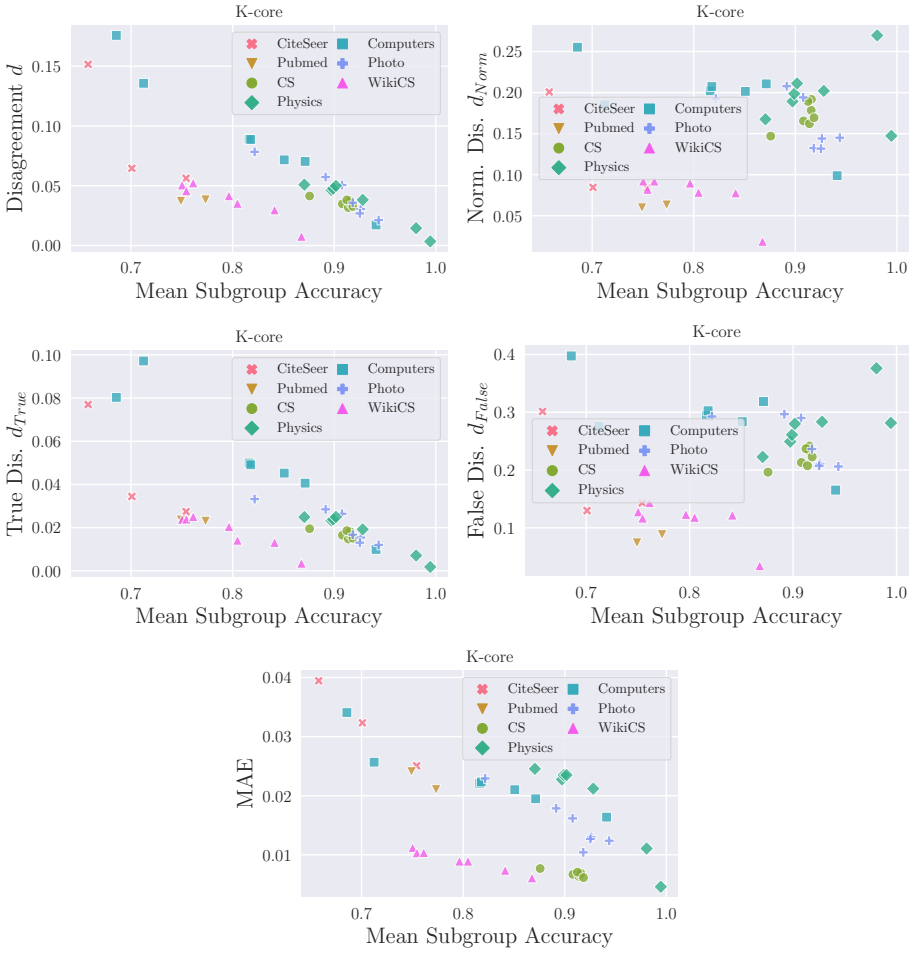

Fig. 21: Prediction disagreement in relation to performance of subgroups of k-core for GAT

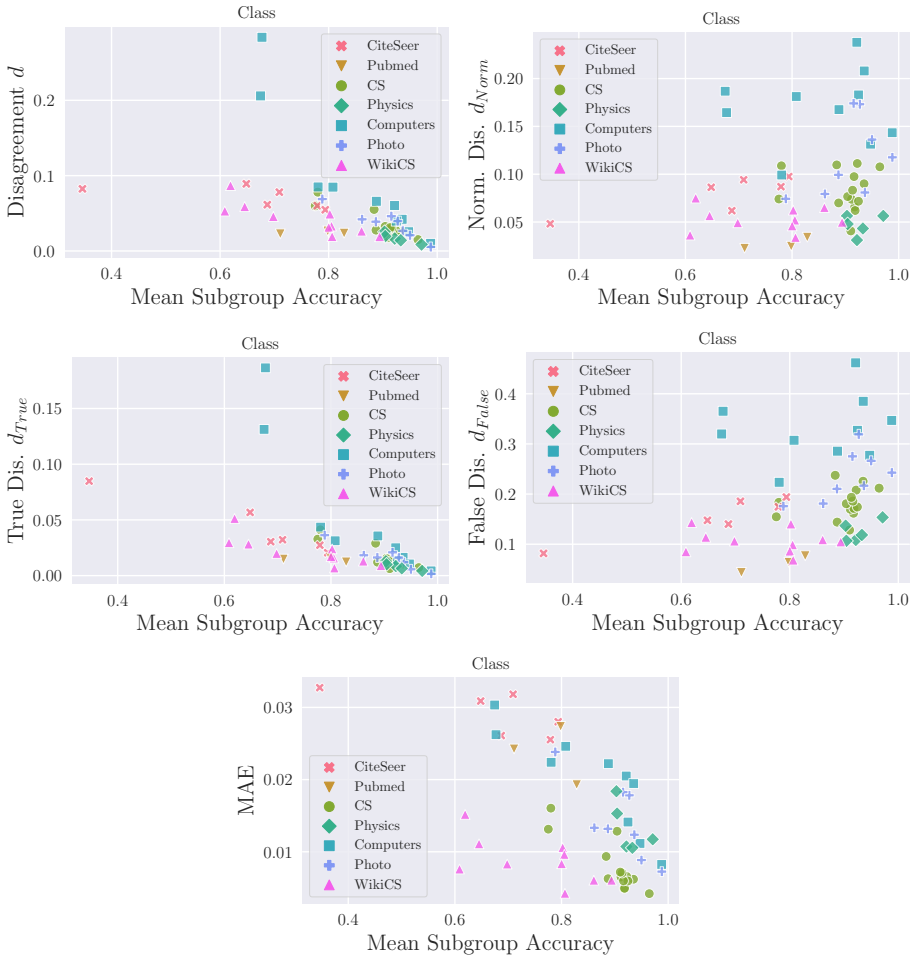

Fig. 22: Prediction disagreement in relation to performance of class label subgroups for GCN

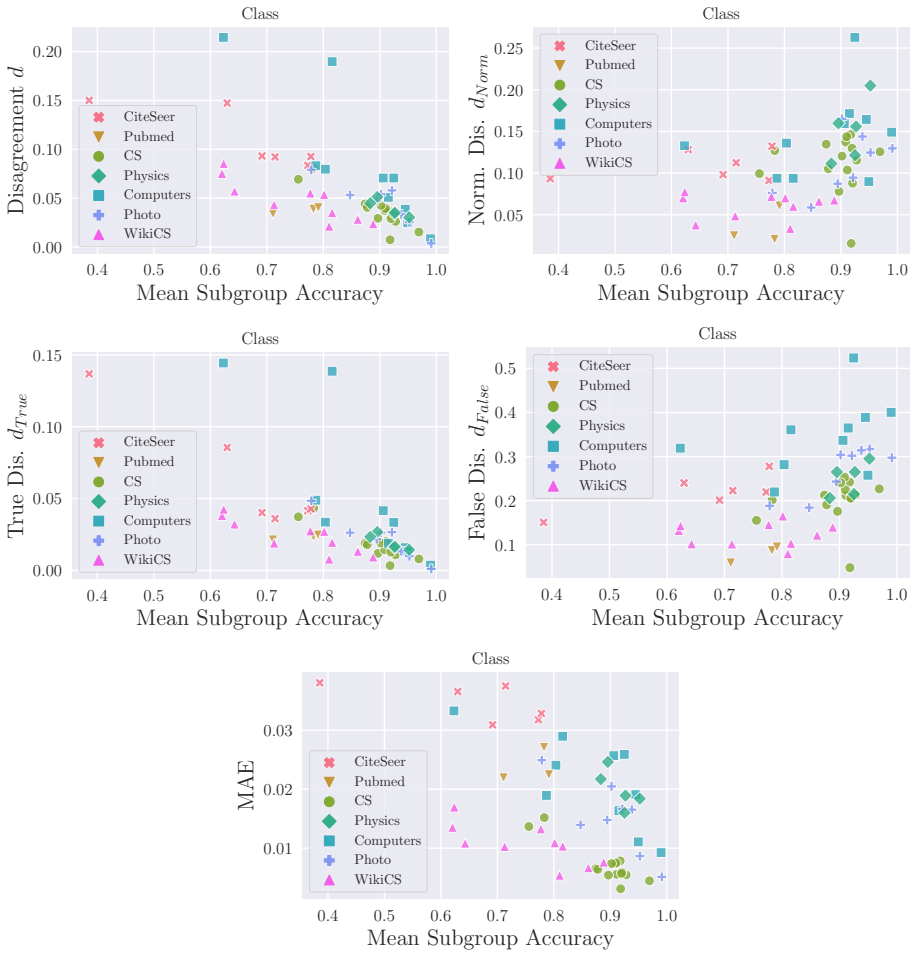

Fig. 23: Prediction disagreement in relation to performance of class label subgroups for GAT

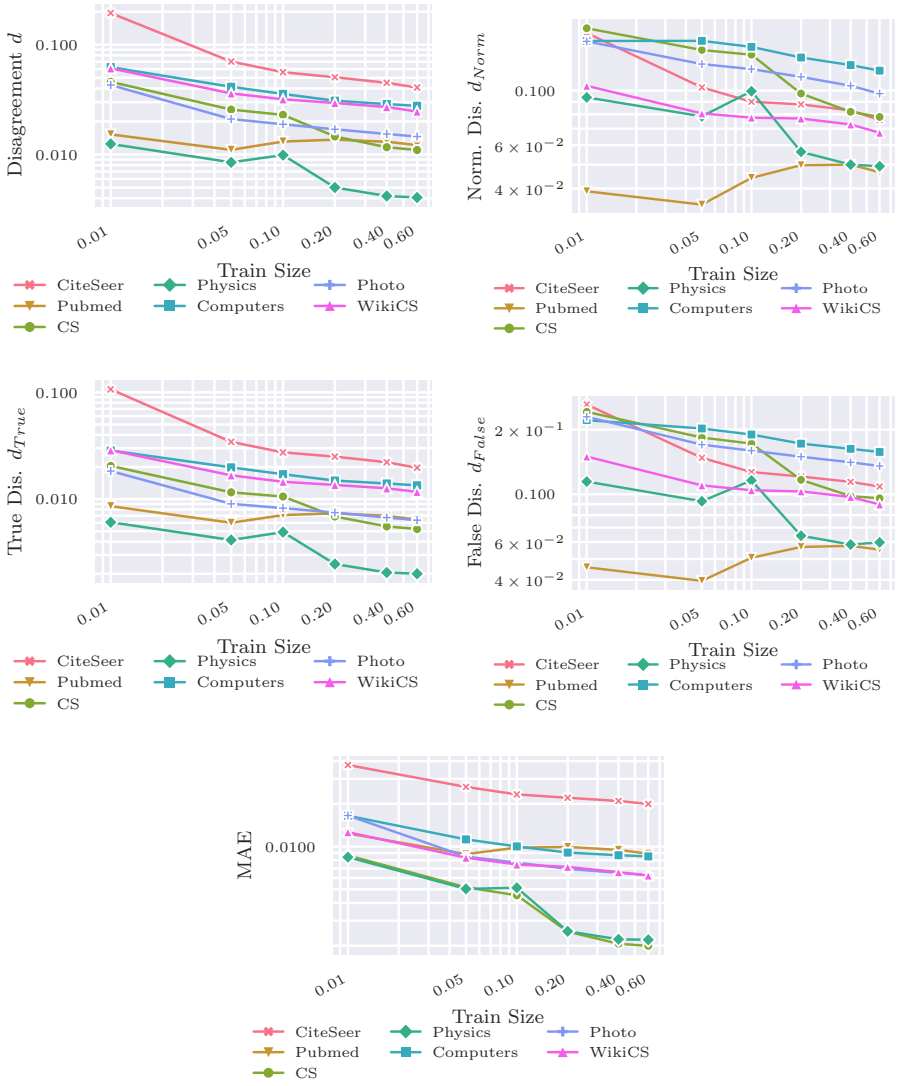

Fig. 24: Results for different amount of training labels for GCN.

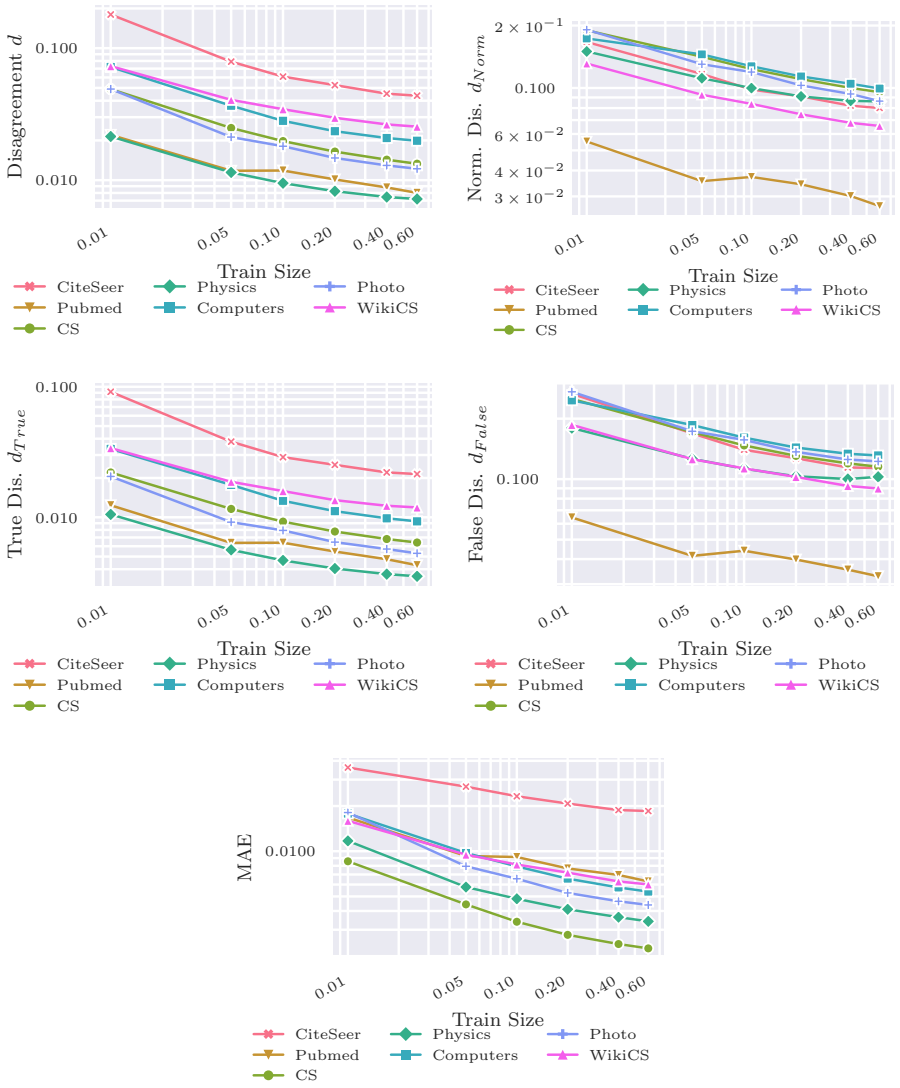

Fig. 25: Results for different amount of training labels for GAT.

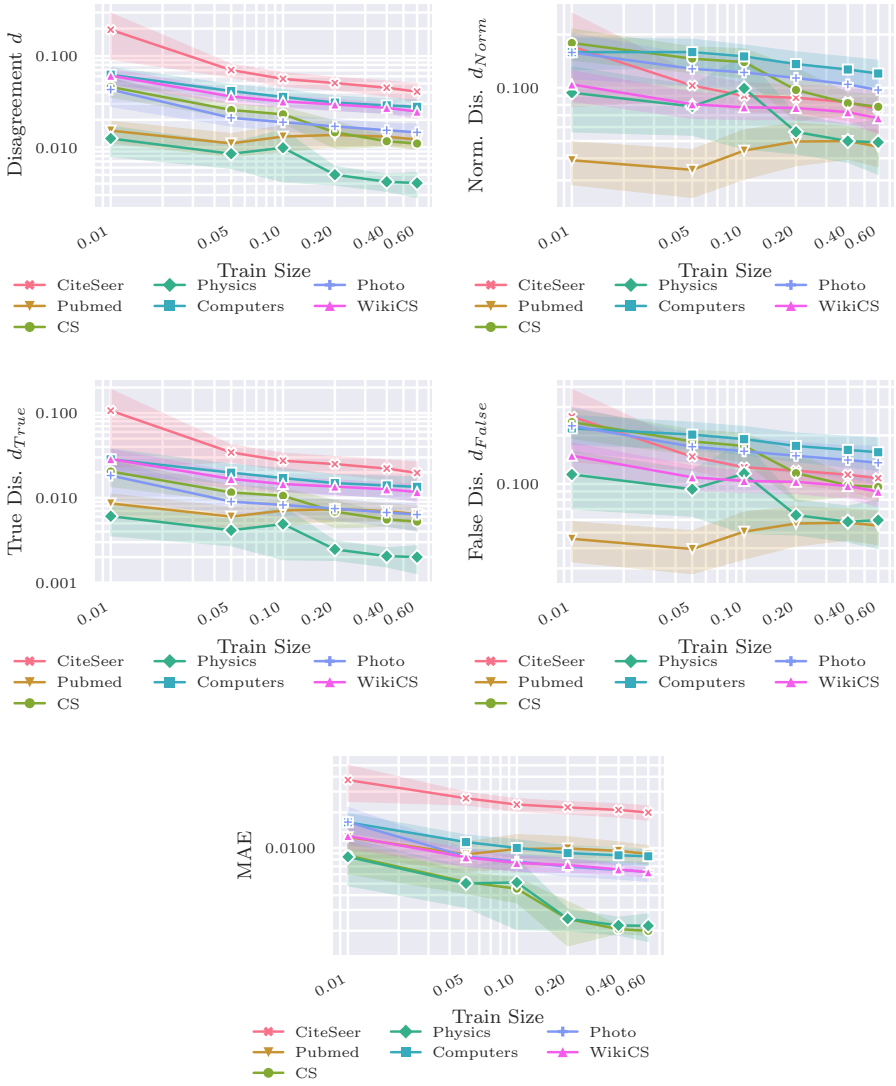

Fig. 26: Results for different amount of training labels for GCN.

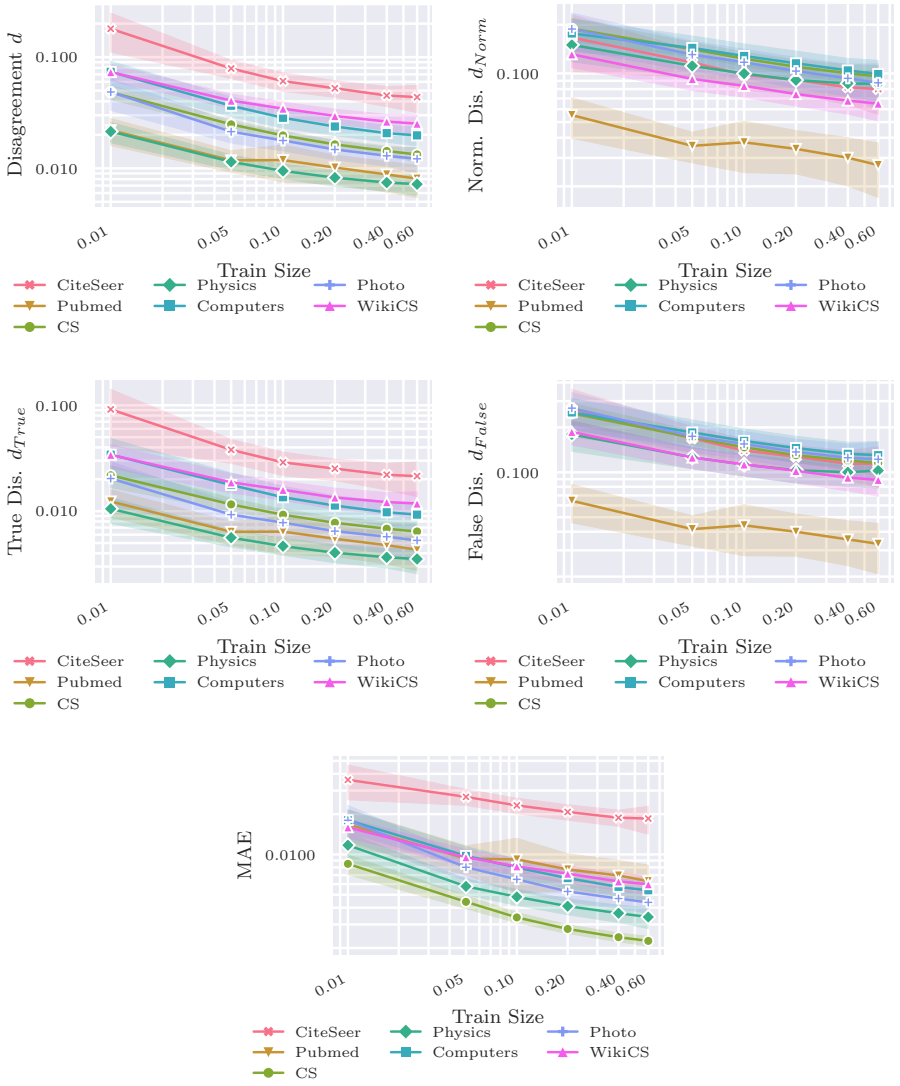

Fig. 27: Results for different amount of training labels for GAT.

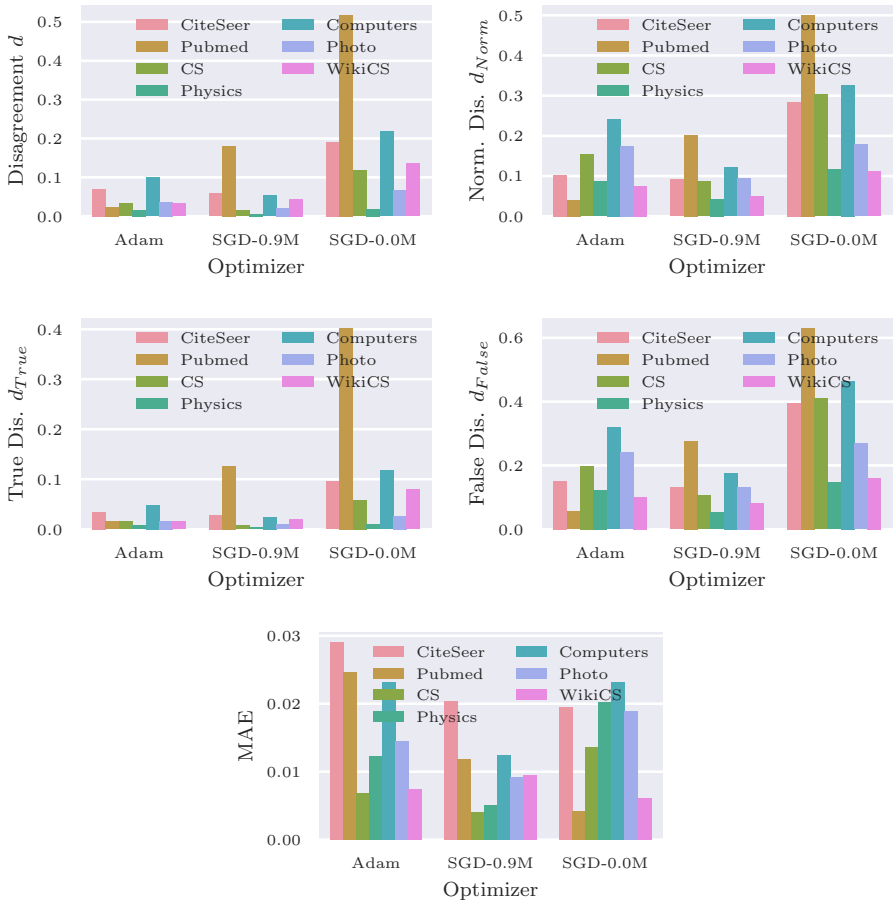

Fig. 28: Results for different optimizers for GCN.

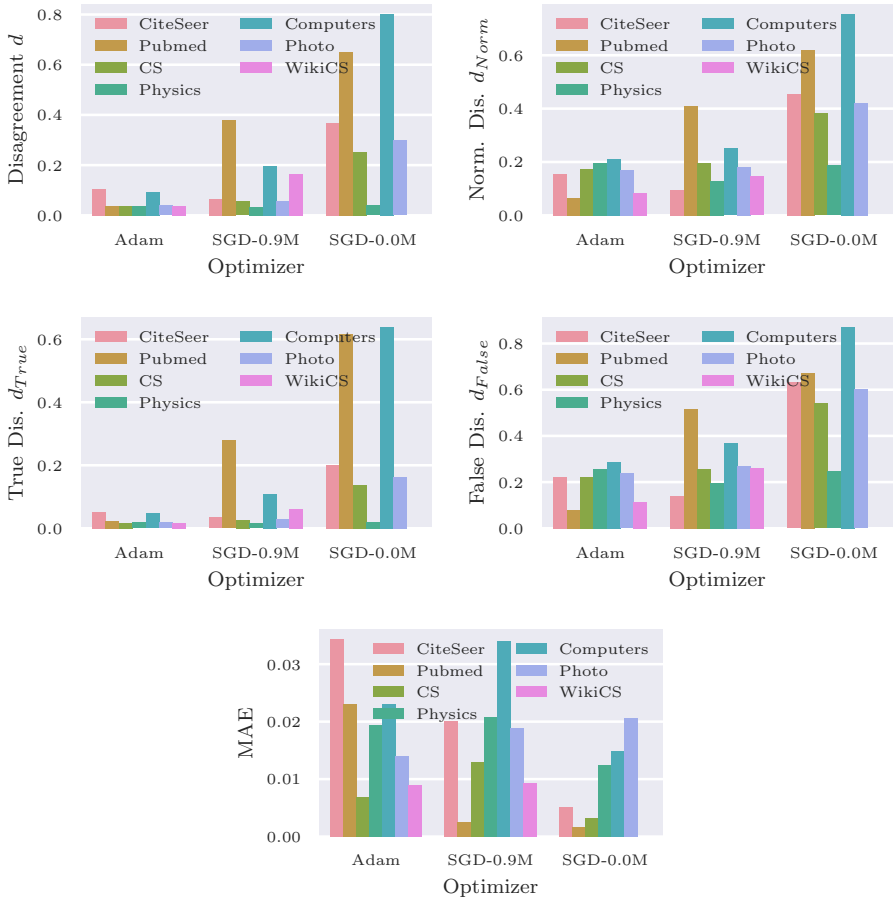

Fig. 29: Results for different optimizers for GAT.

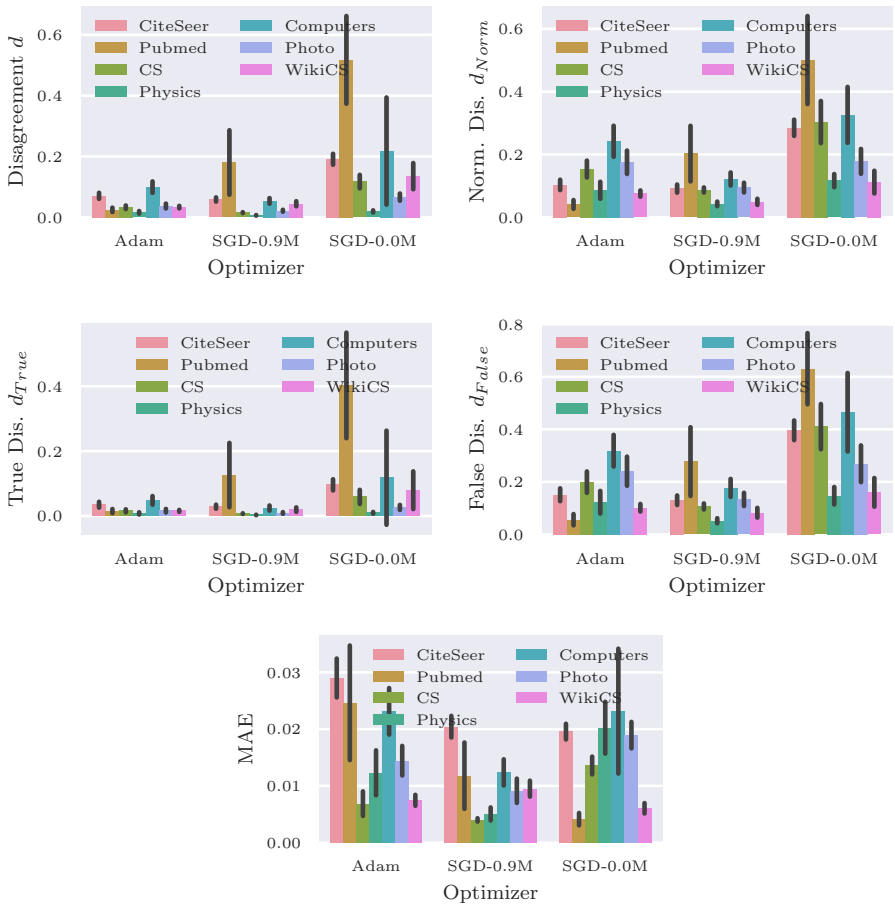

Fig. 30: Results for different optimizers for GCN.

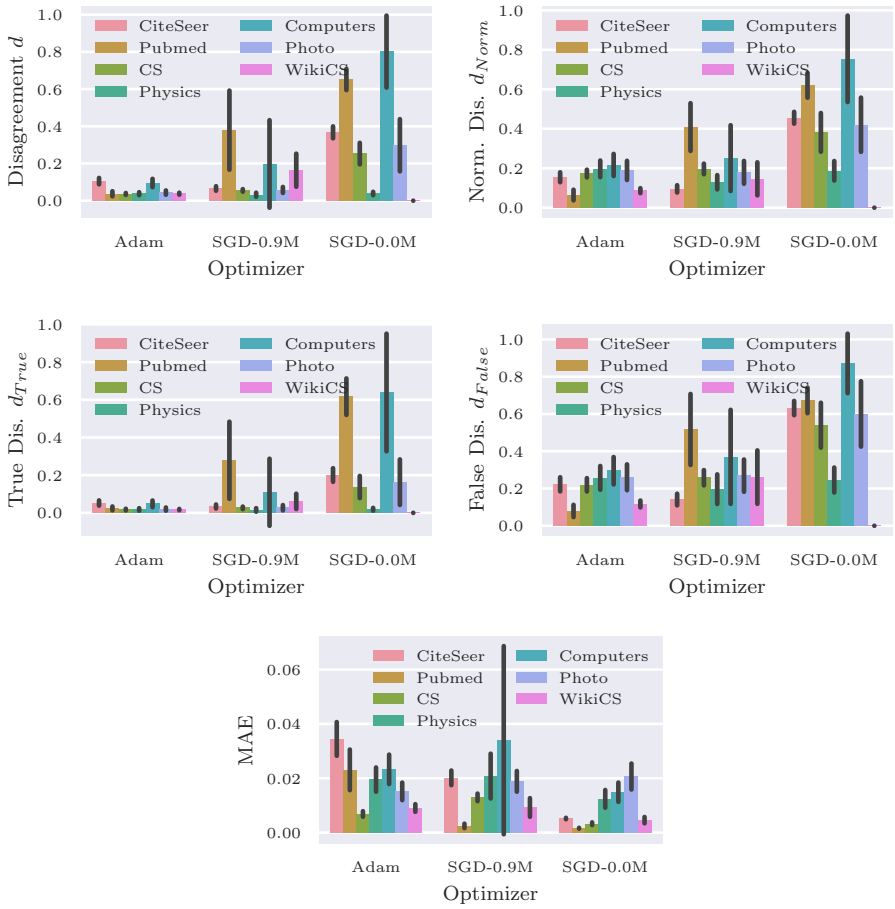

Fig. 31: Results for different optimizers for GAT.

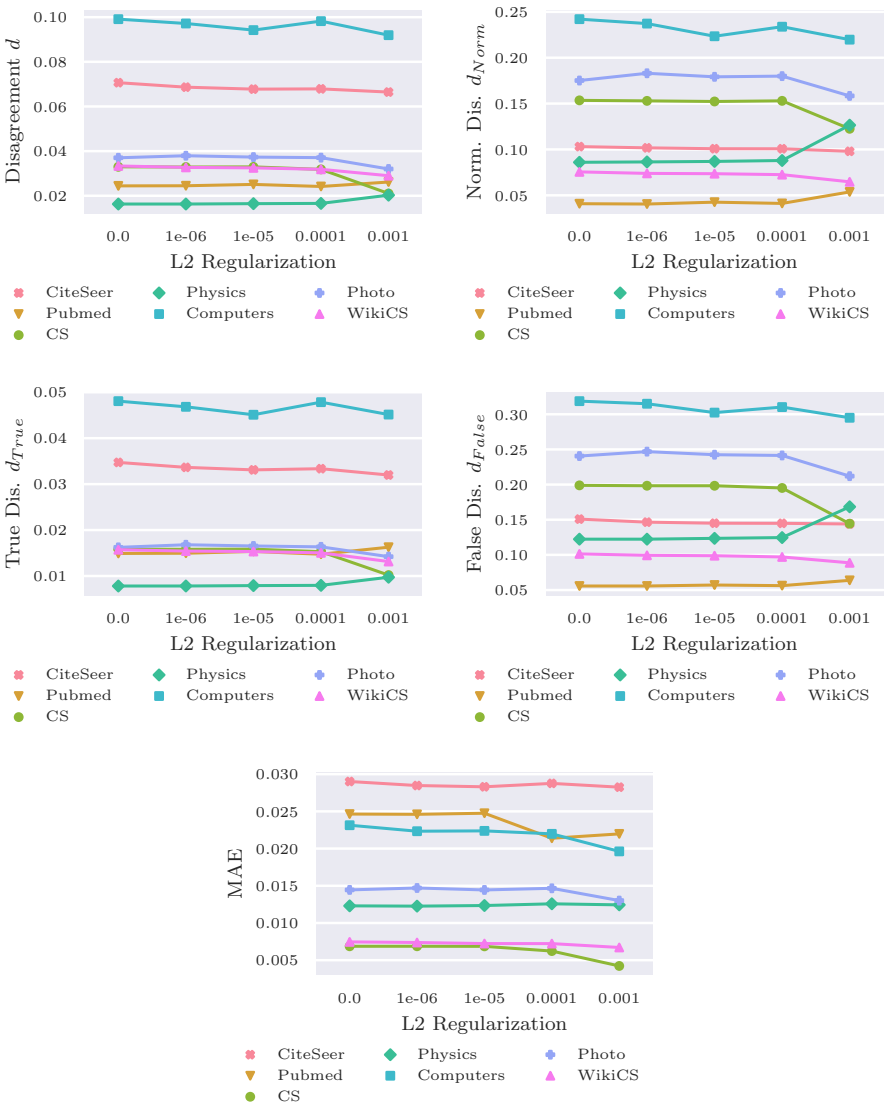

Fig. 32: Results for varying L2 regularization for GCN.

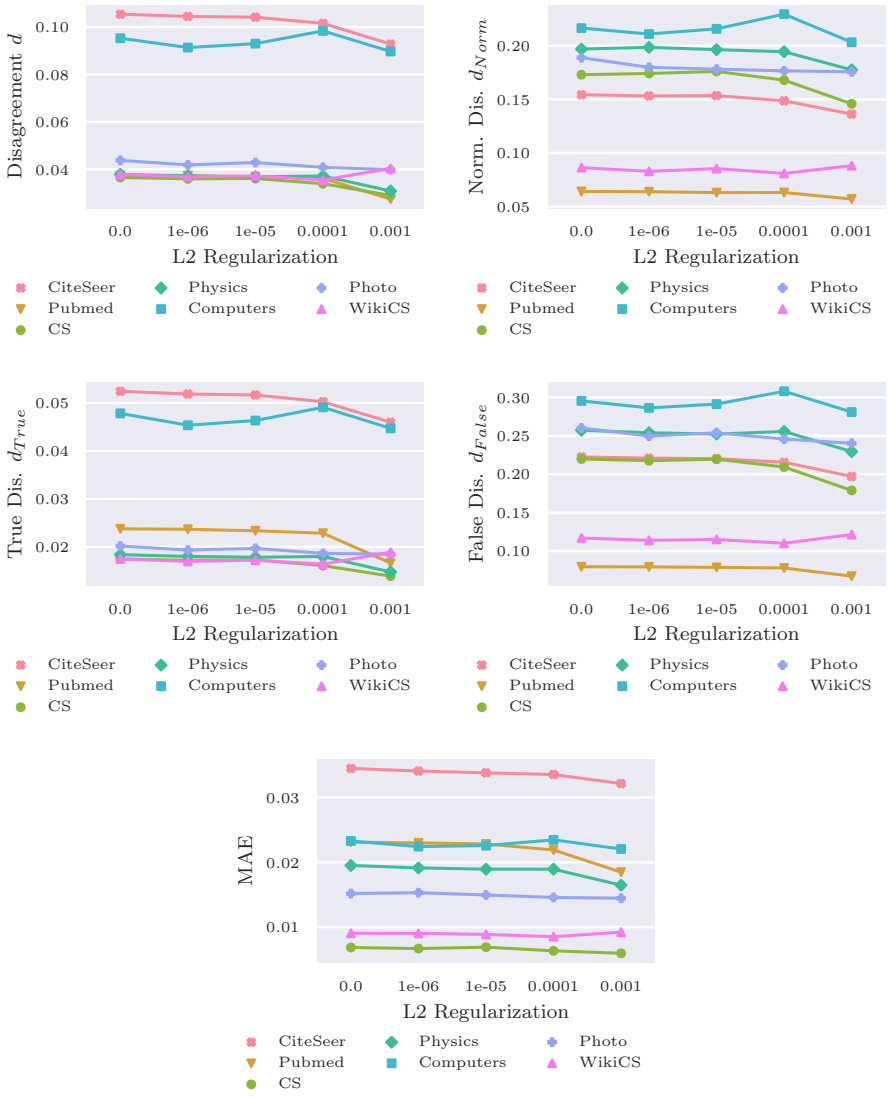

Fig. 33: Results for varying L2 regularization for GAT.

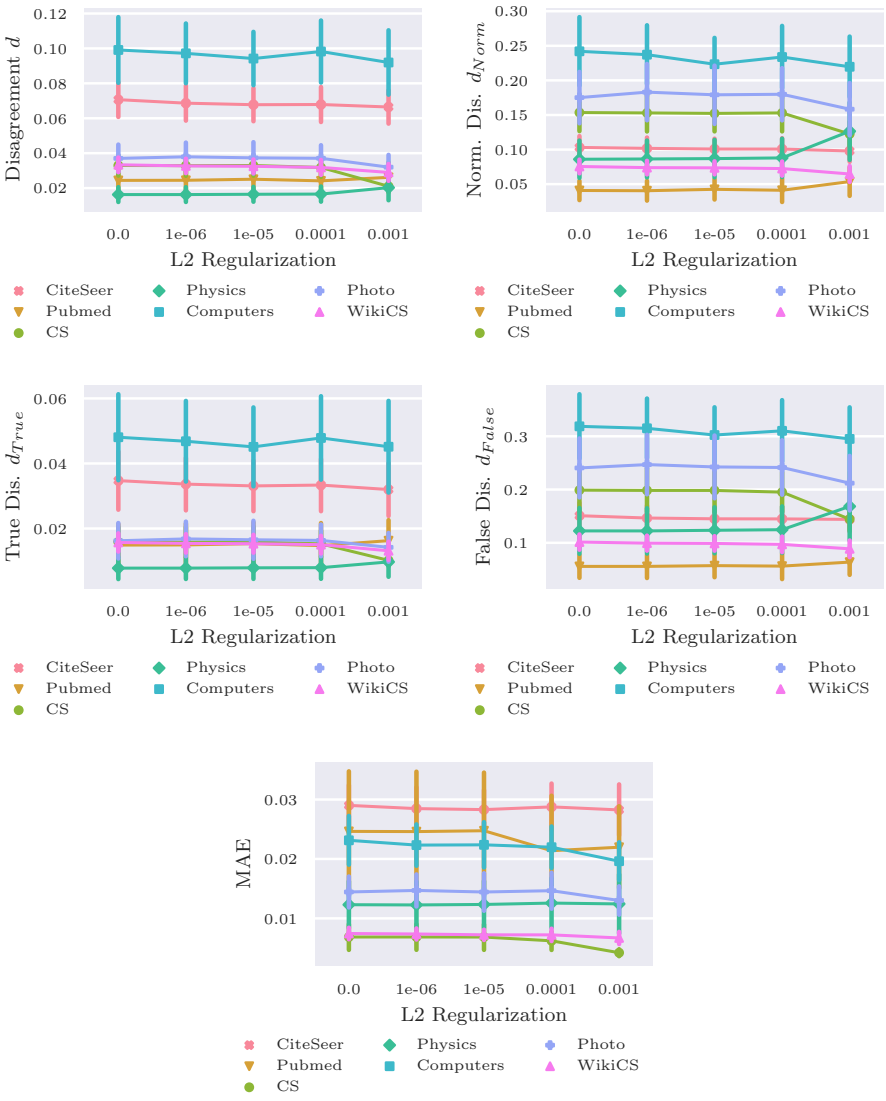

Fig. 34: Results for varying L2 regularization for GCN.

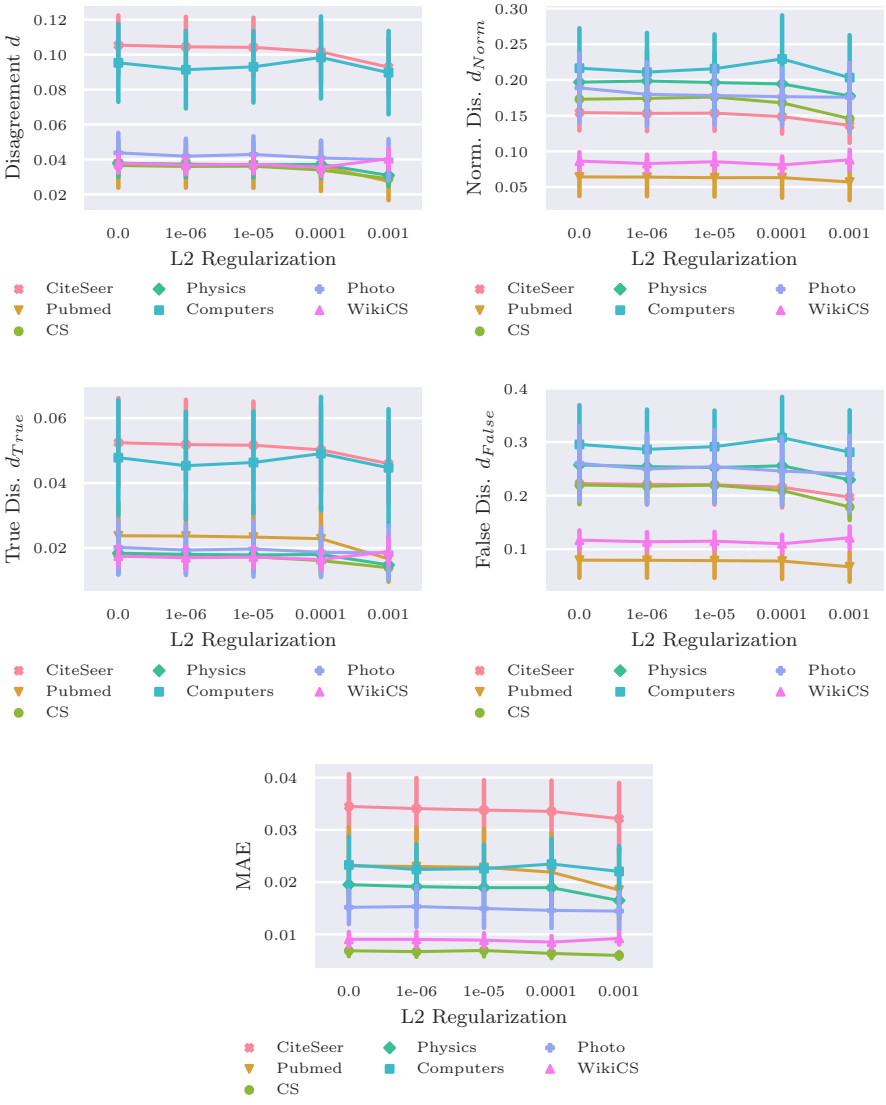

Fig. 35: Results for varying L2 regularization for GAT.

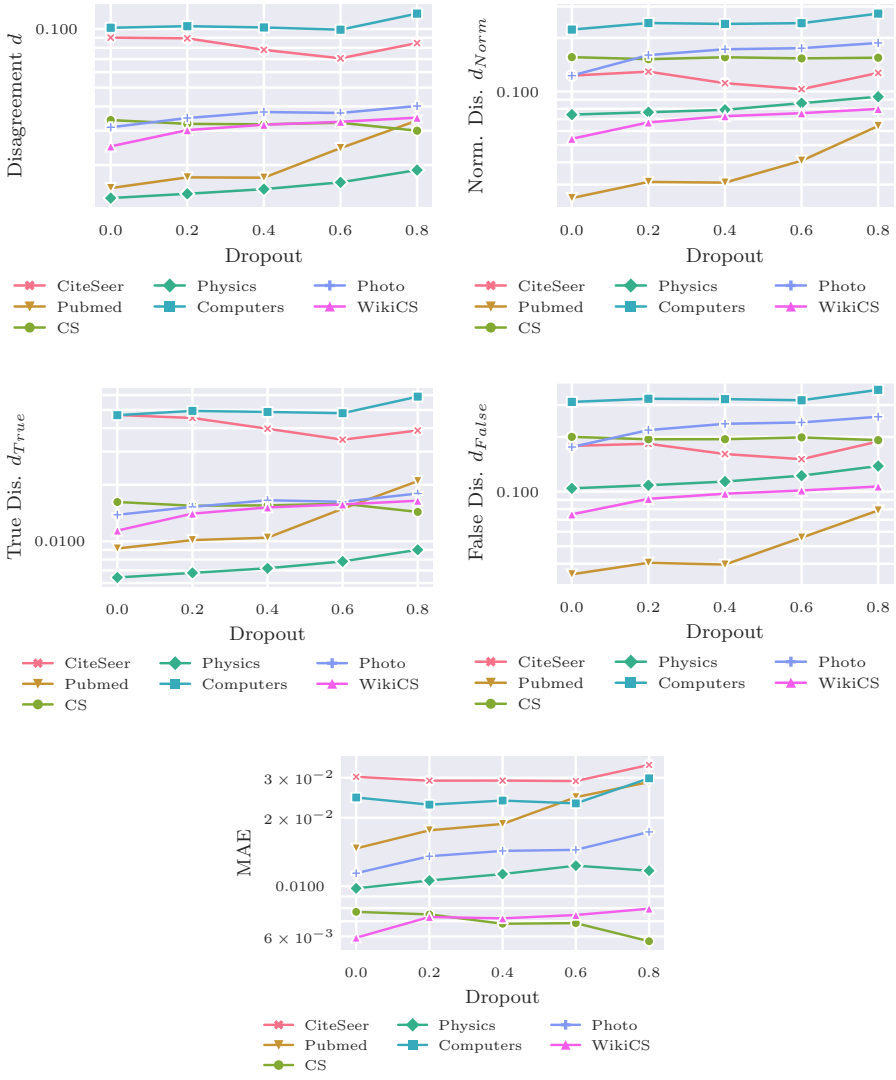

Fig. 36: Results for varying dropout for GCN.

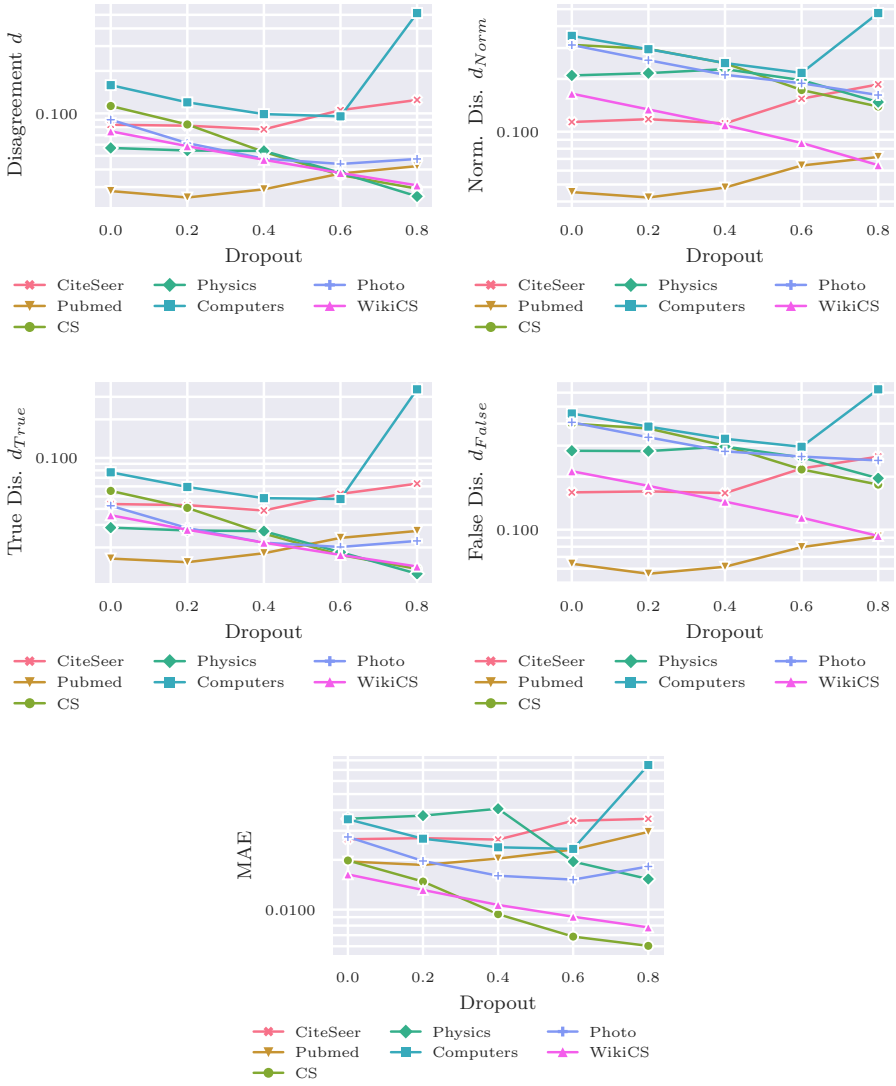

Fig. 37: Results for varying dropout for GAT.

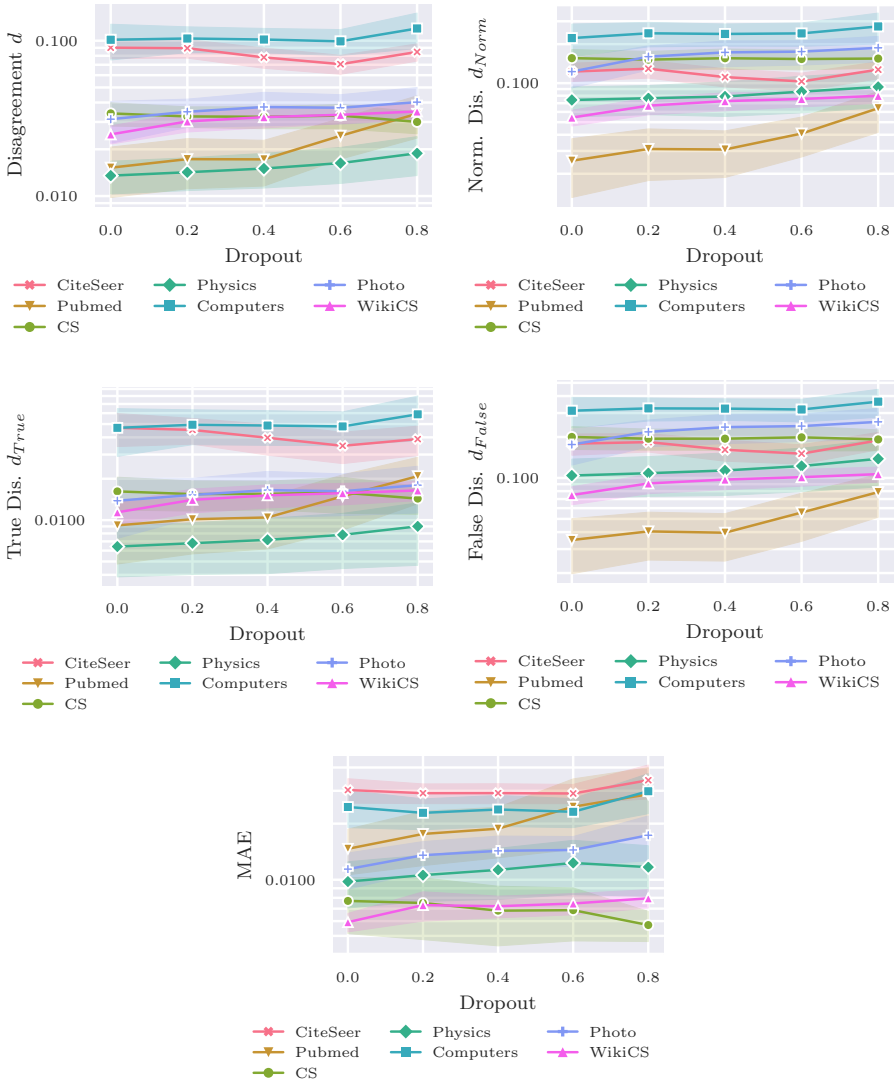

Fig. 38: Results for varying dropout for GCN.

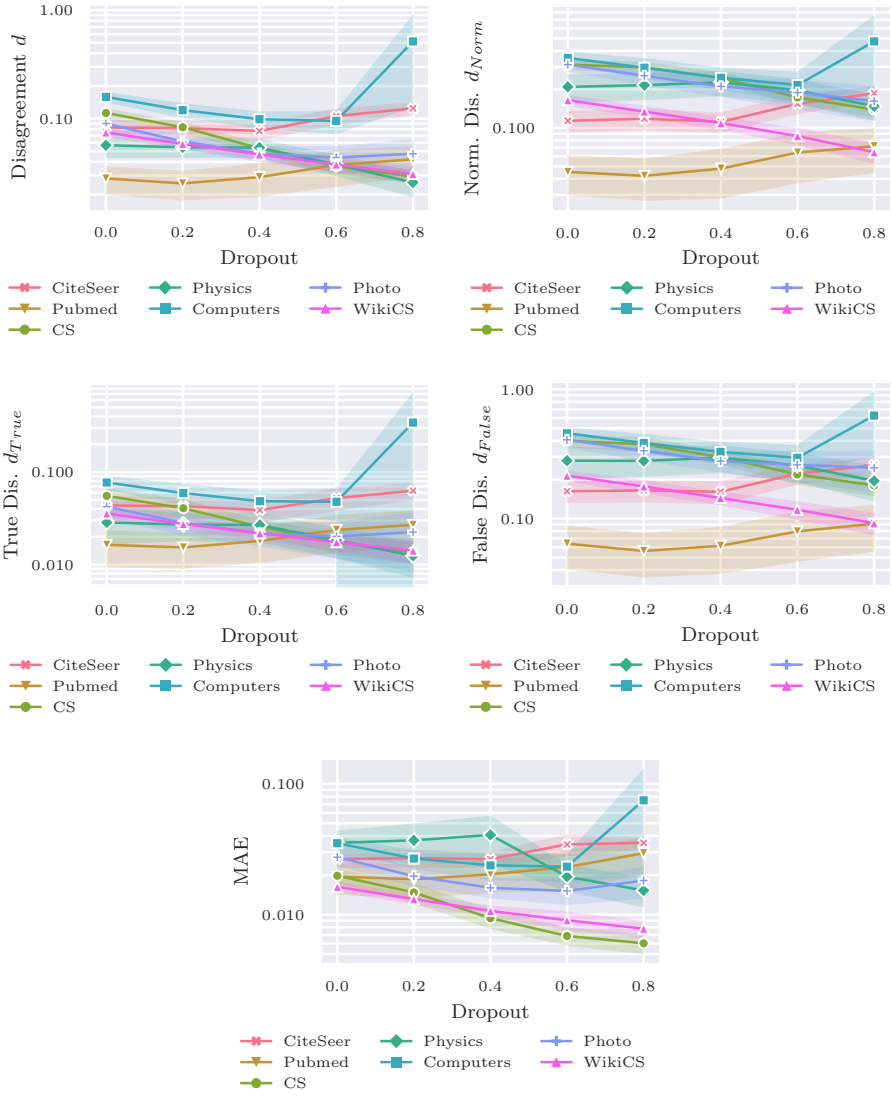

Fig. 39: Results for varying dropout for GAT.

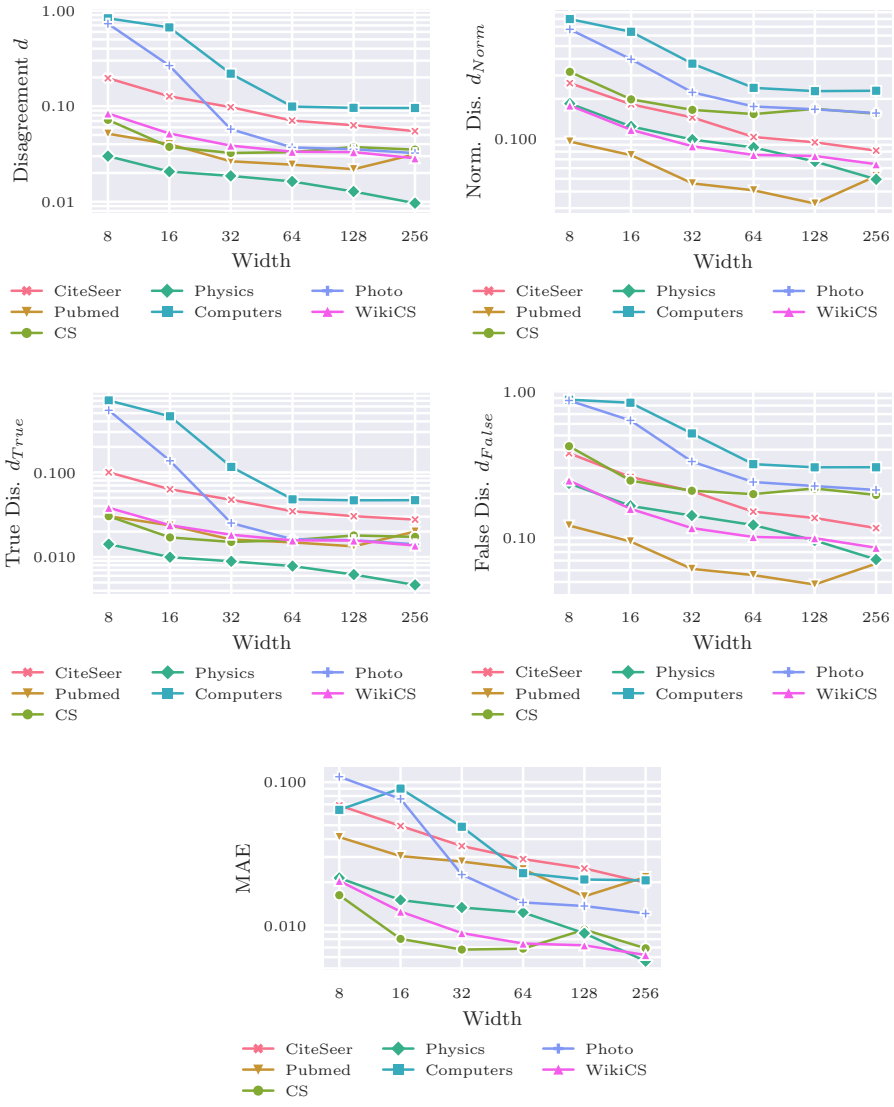

Fig. 40: Results for different model width for GCN.

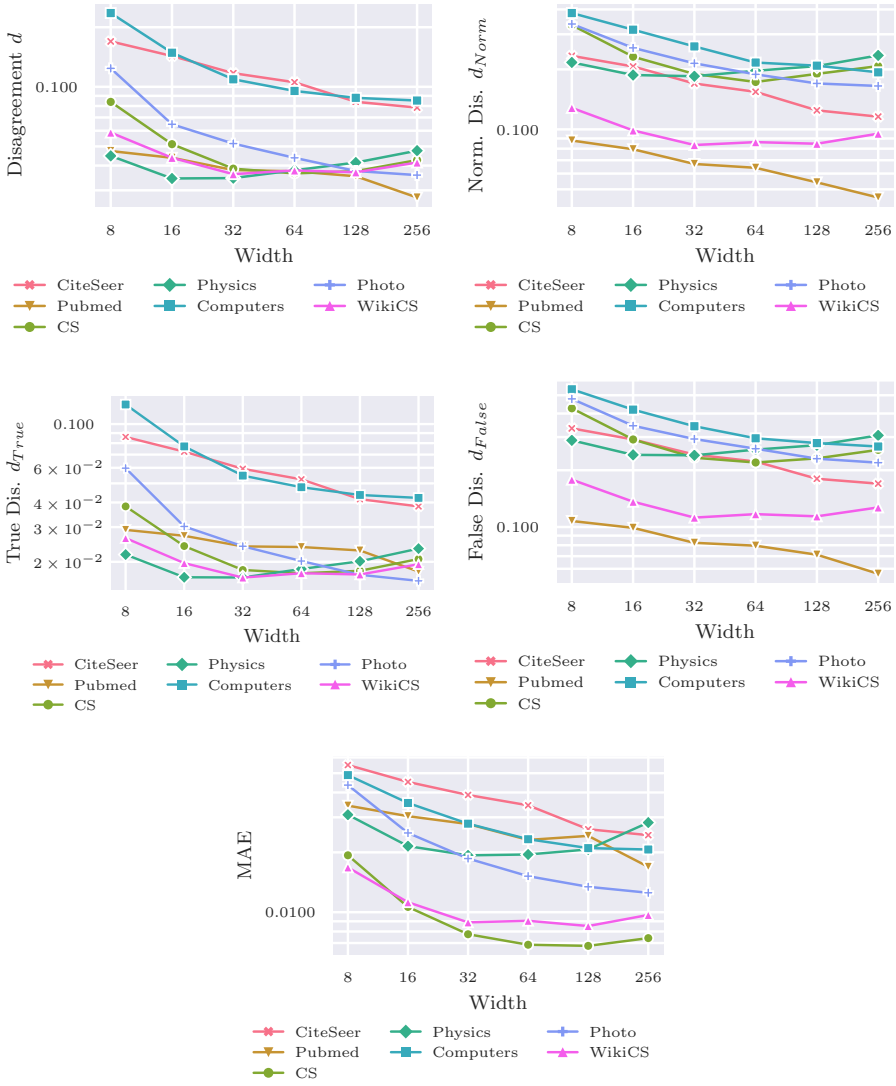

Fig. 41: Results for different model width for GAT.

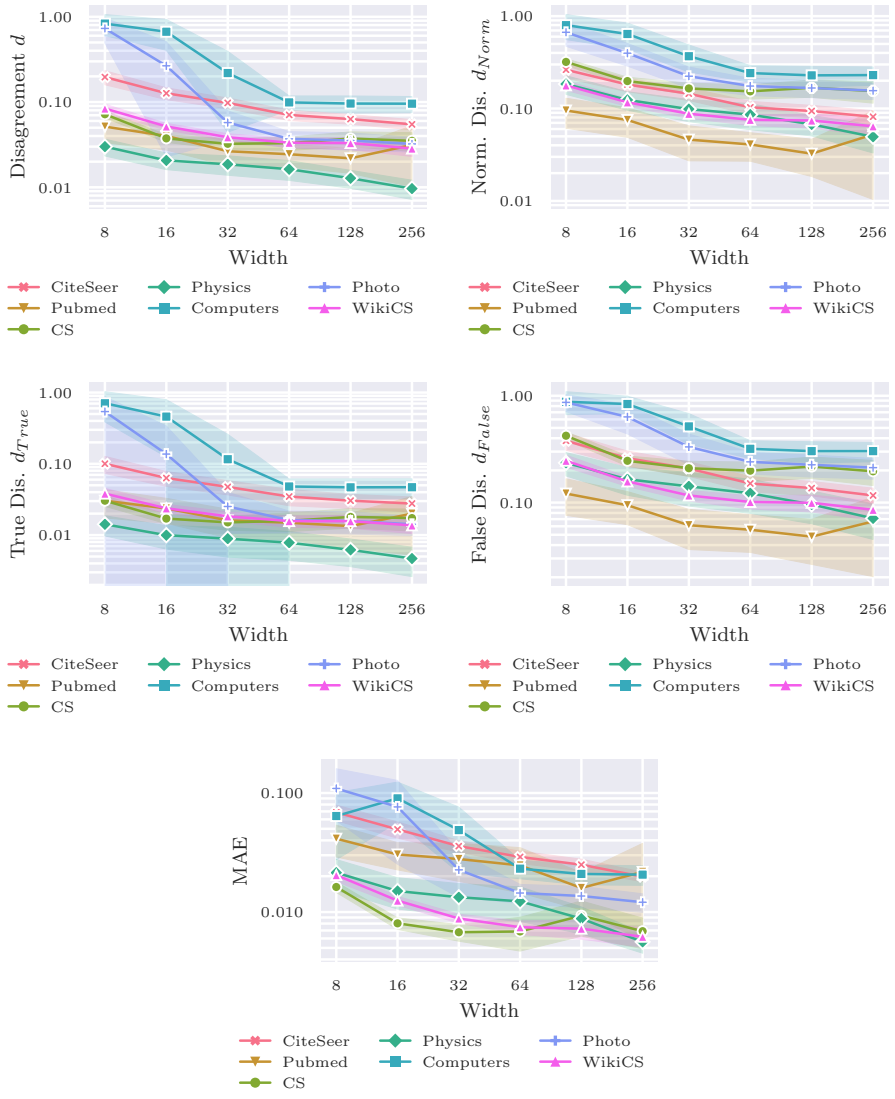

Fig. 42: Results for different model width for GCN.

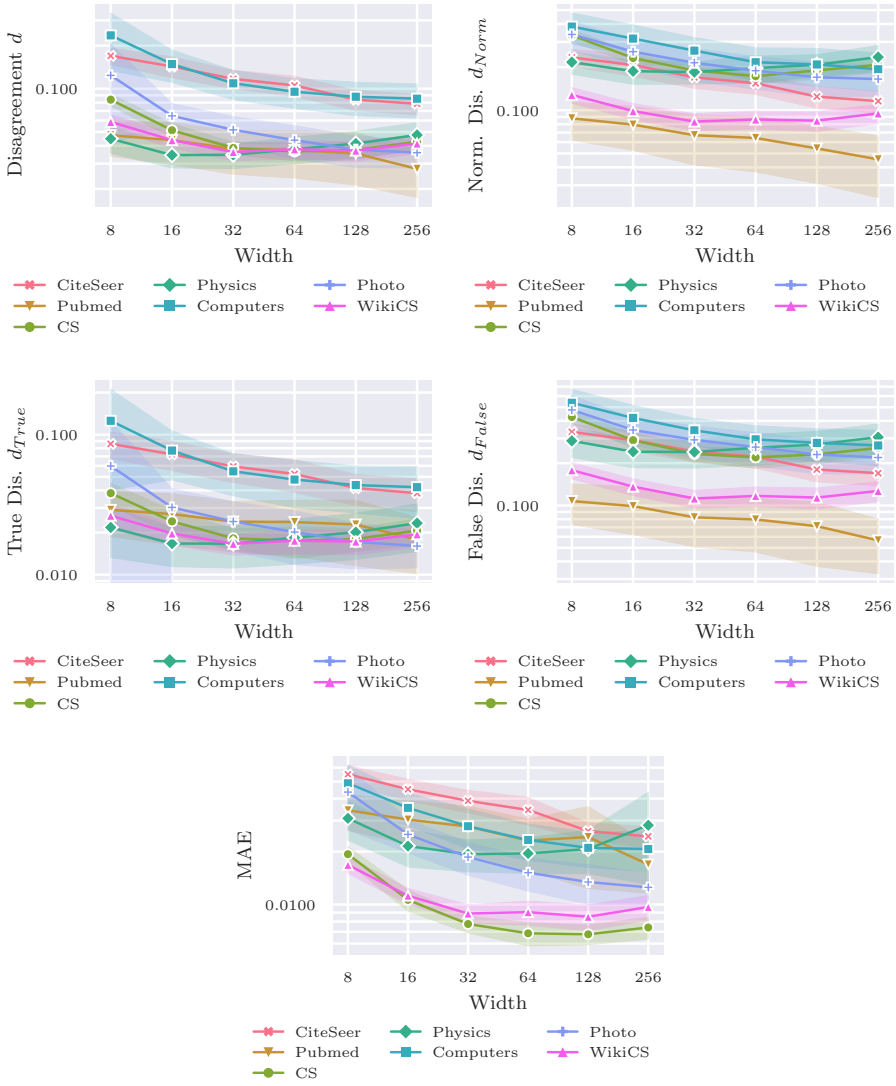

Fig. 43: Results for different model width for GAT.

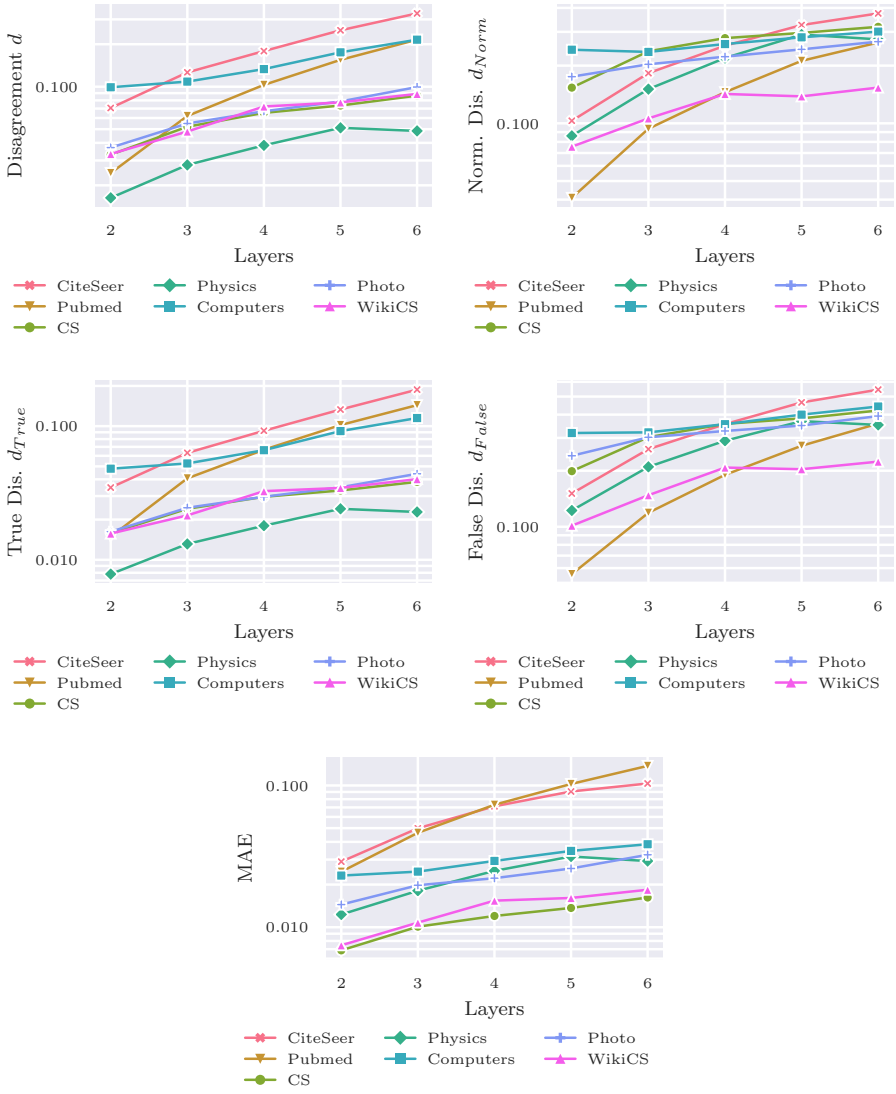

Fig. 44: Results for varying model depth for GCN.

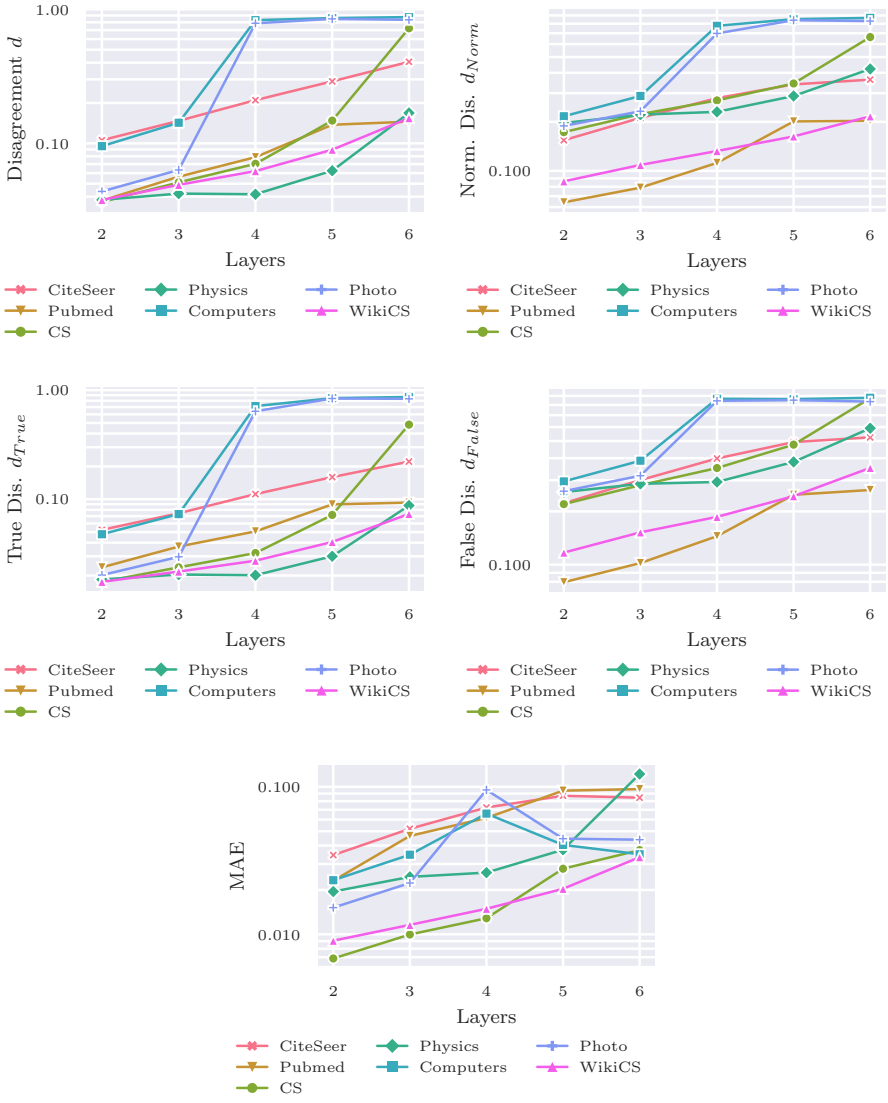

Fig. 45: Results for varying model depth for GAT.

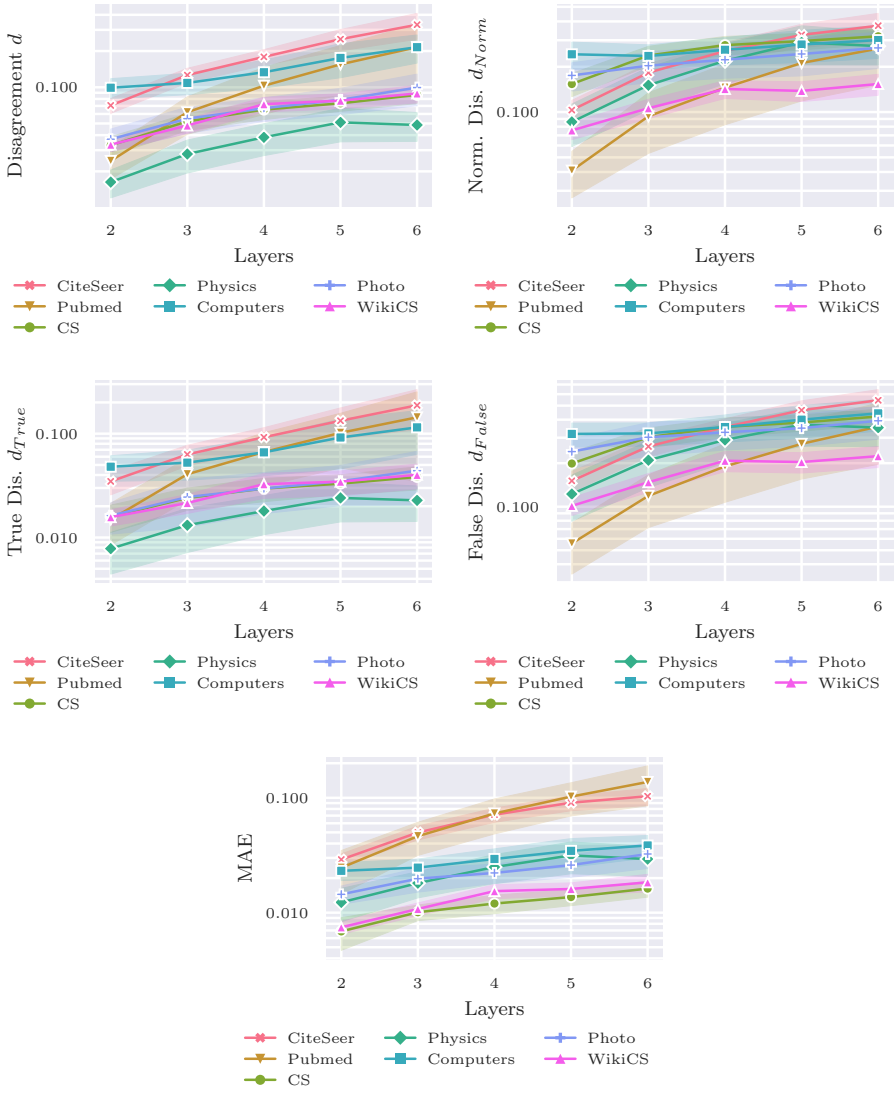

Fig. 46: Results for varying model depth for GCN.

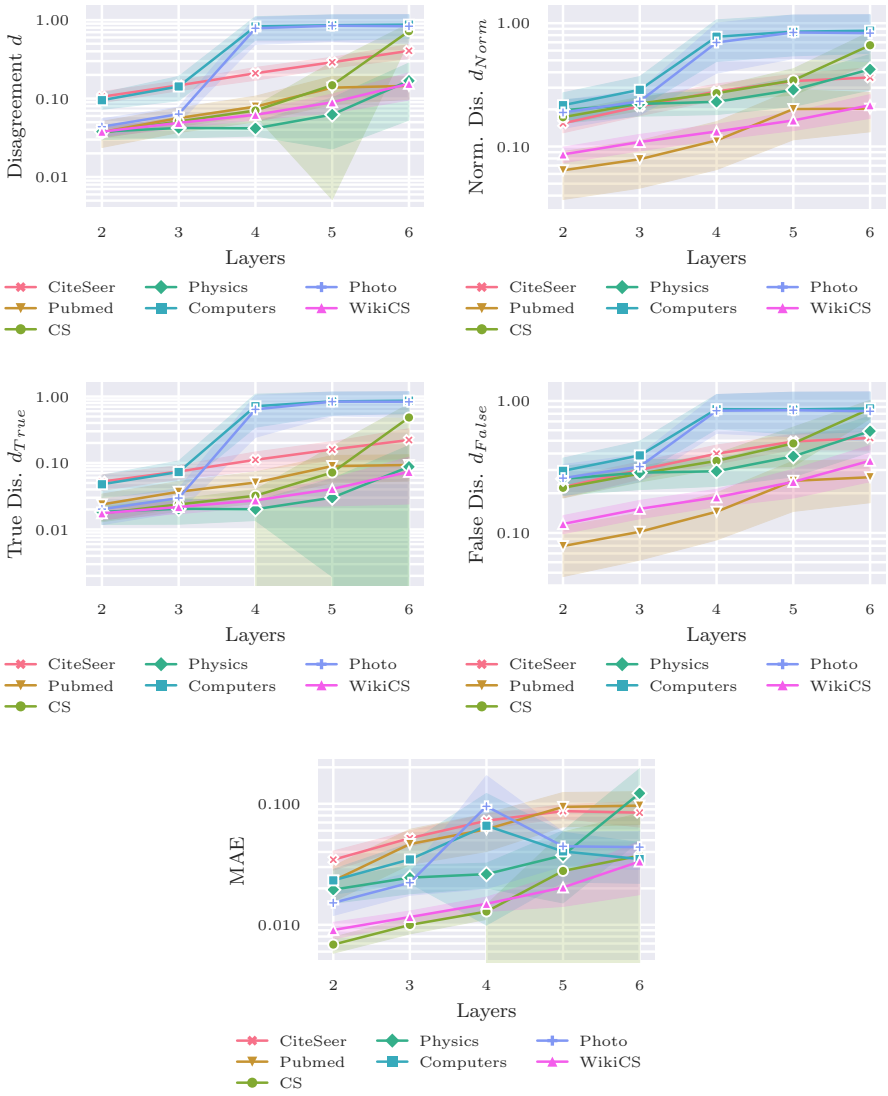

Fig. 47: Results for varying model depth for GAT.

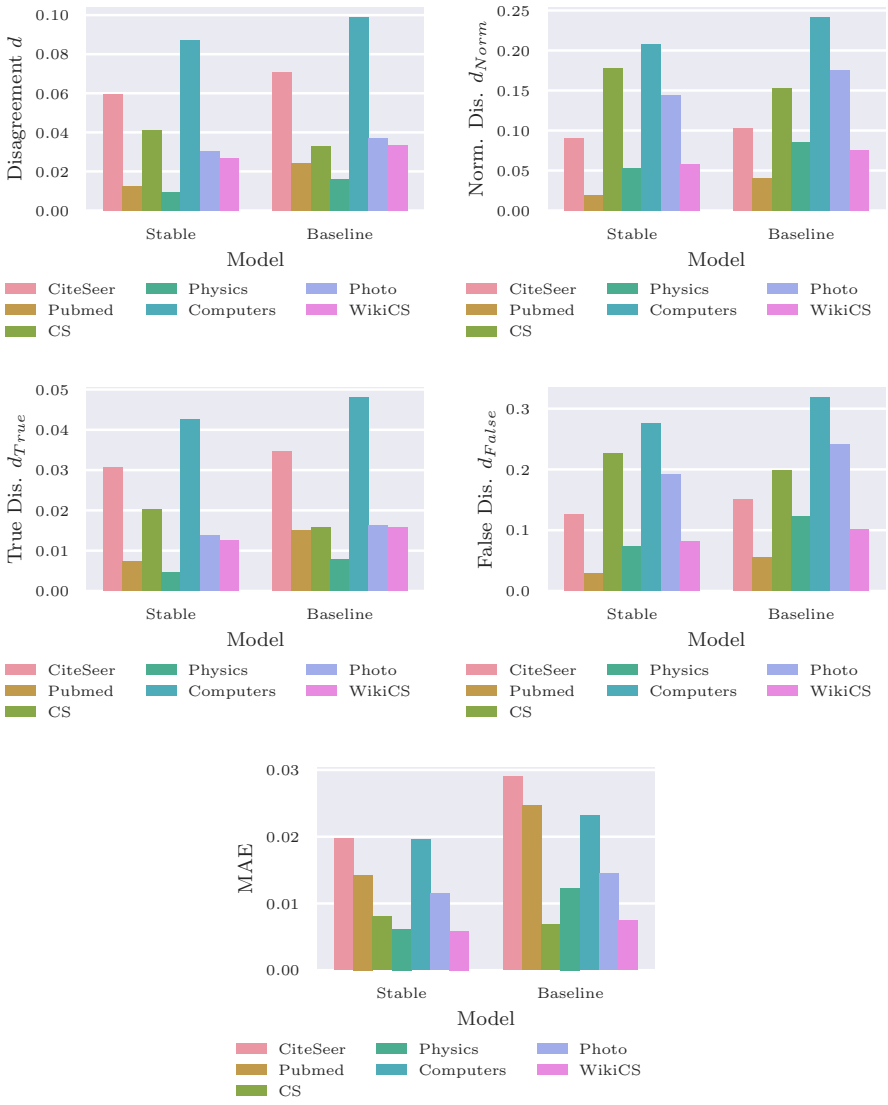

Fig. 48: Results of combining optimal hyperparameters for GCN.

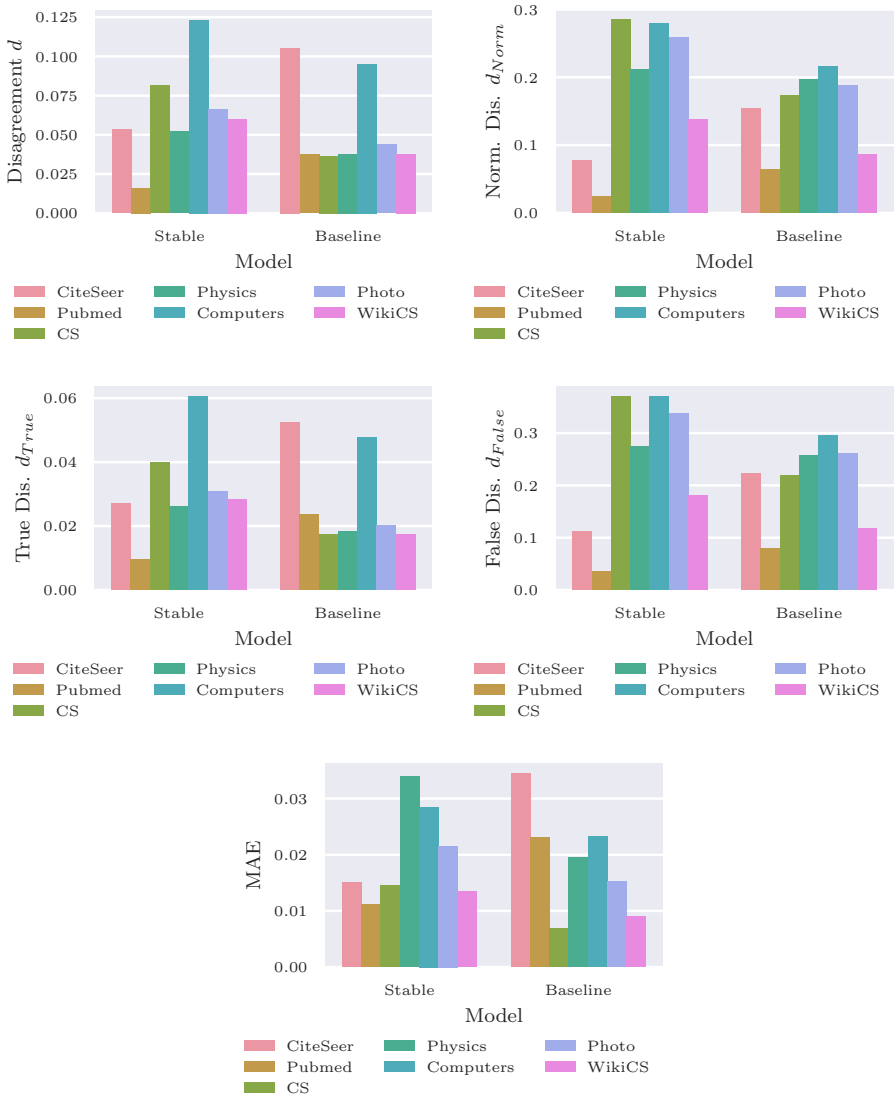

Fig. 49: Results of combining optimal hyperparameters for GAT.

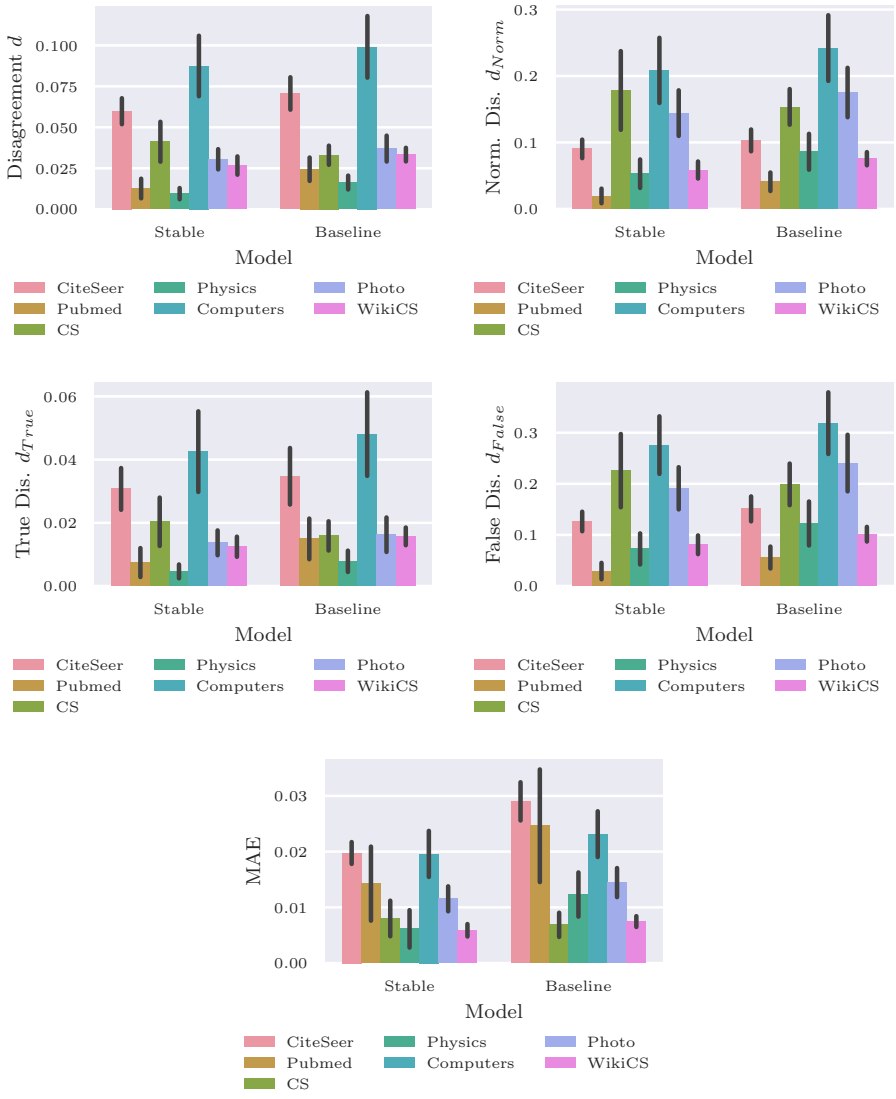

Fig. 50: Results of combining optimal hyperparameters for GCN.

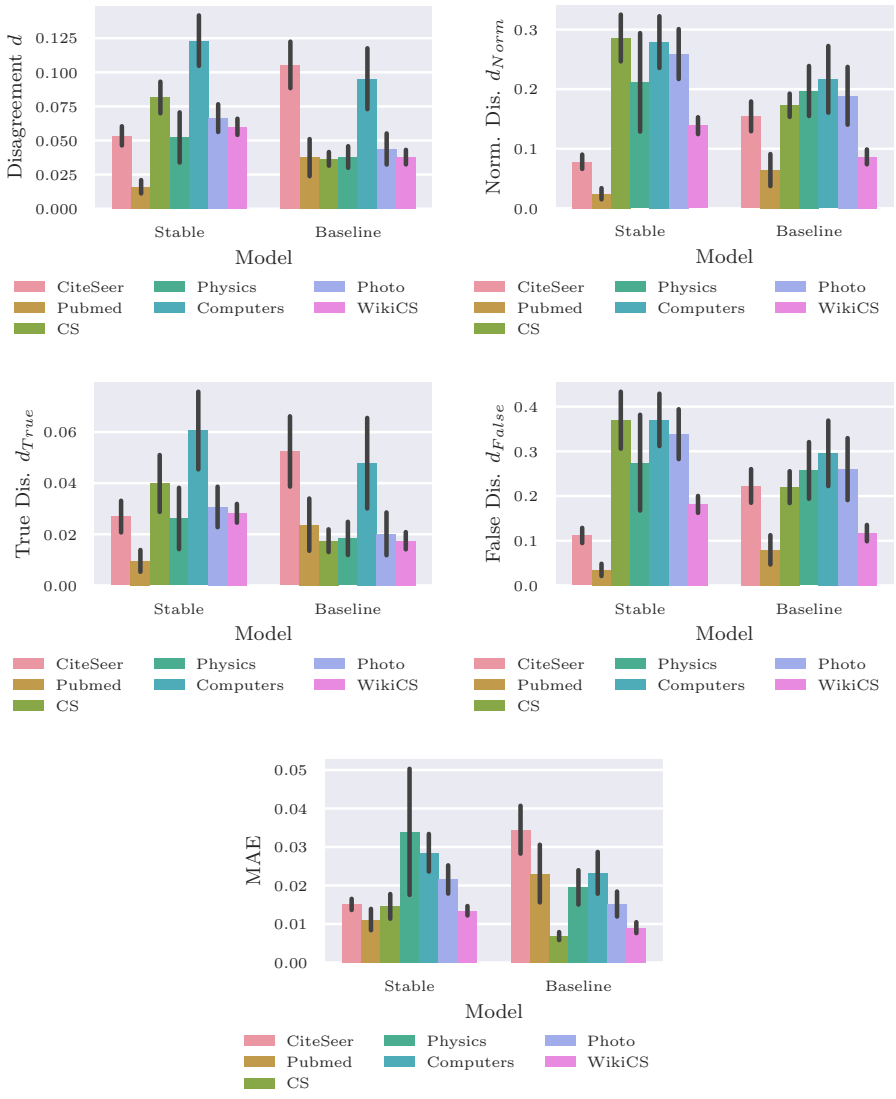

Fig. 51: Results of combining optimal hyperparameters for GAT.

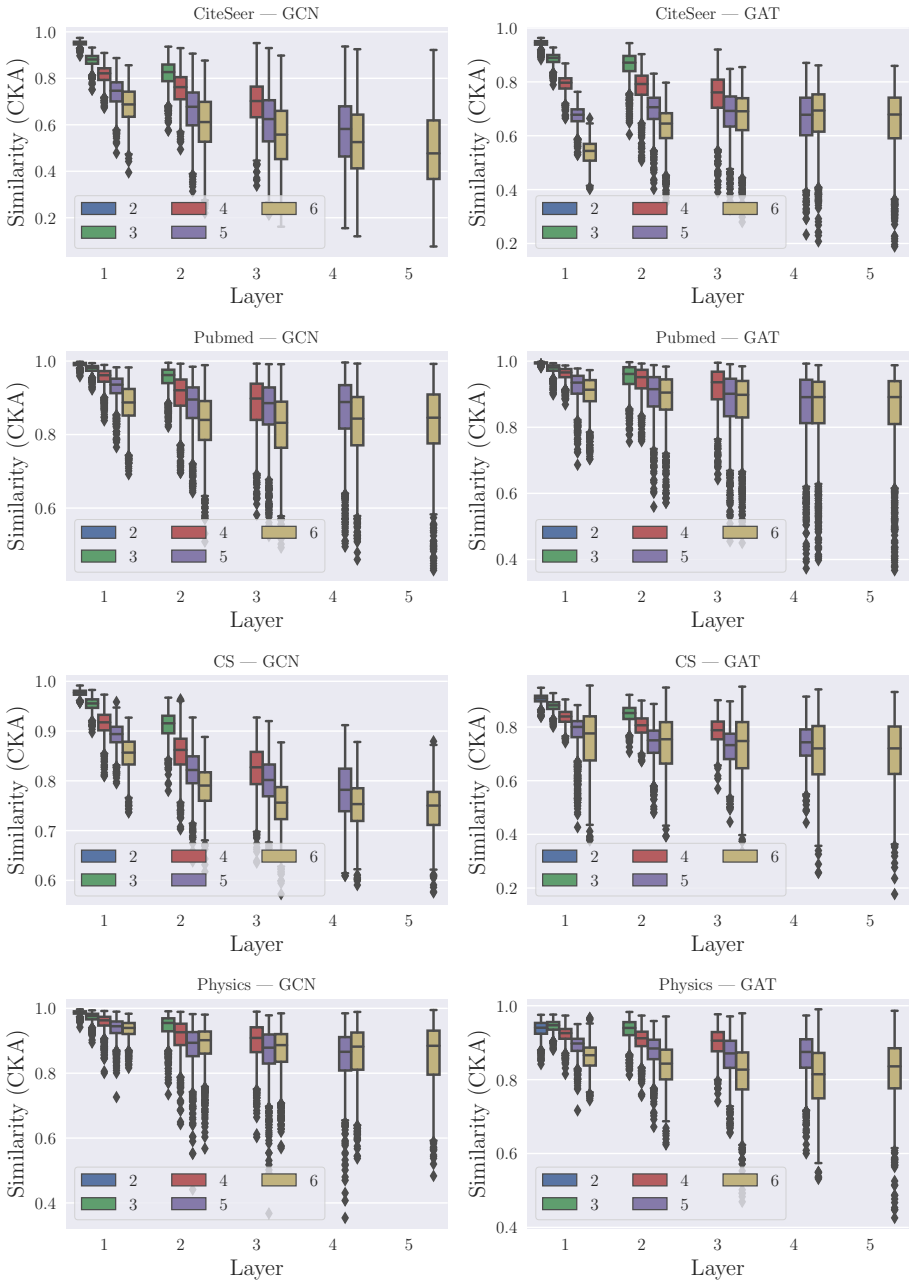

Fig. 52: Similarity of layers to corresponding layers of GCN (left) and GAT (right) on CiteSeer, Pubmed, CS, and Physics.

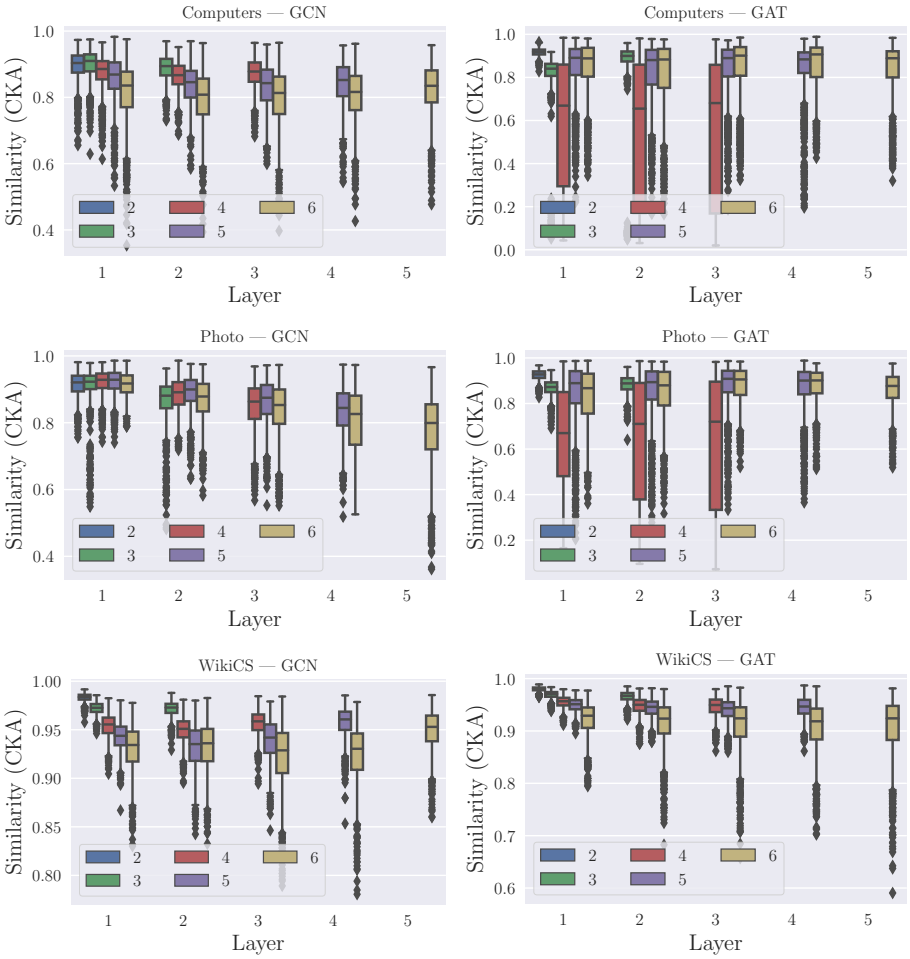

Fig. 53: Similarity of layers to corresponding layers of GCN (left) and GAT (right) on Computers, Photo, and WikiCS

References

1. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of Neural Network Representations Revisited. In: Proceedings of the 36th International Conference on Machine Learning. pp. 3519–3529. PMLR (May 2019), https://proceedings.mlr.press/v97/kornblith19a.html, iSSN: 2640-3498
2. Langville, A.N., Meyer, C.D.: A Survey of Eigenvector Methods for Web Information Retrieval. SIAM Review **47**(1), 135–161 (Jan 2005). https://doi.org/10.1137/S0036144503424786, http://epubs.siam.org/doi/10.1137/S0036144503424786
3. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. (1999)
4. Seidman, S.B.: Network structure and minimum degree. Social Networks **5**(3), 269–287 (Sep 1983). https://doi.org/10.1016/0378-8733(83)90028-X, https://www.sciencedirect.com/science/article/pii/037887338390028X
5. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–442 (Jun 1998). https://doi.org/10.1038/30918, https://www.nature.com/articles/30918, number: 6684 Publisher: Nature Publishing Group