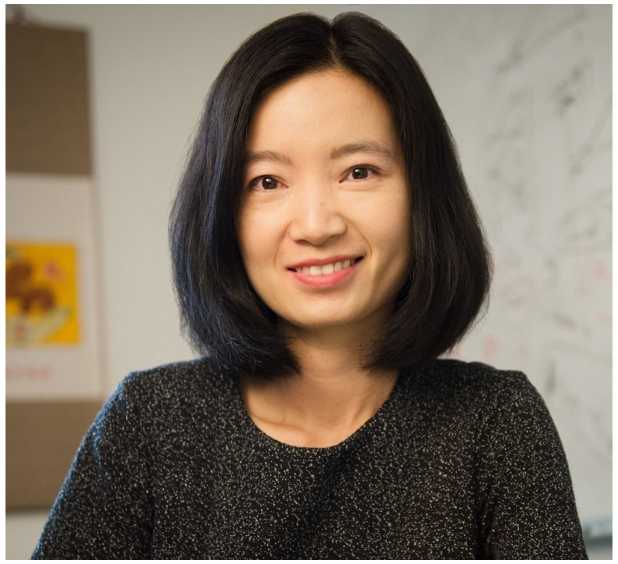
Theoretical and Empirical Insights into the Origins of Degree Bias in Graph Neural Networks

Arjun Subramonian, Jian Kang, Yizhou Sun

NeurIPS 2024



Contributions

- We provide thorough, empirically-validated **theoretical analysis** of why GNNs perform better for high-degree nodes on node classification tasks

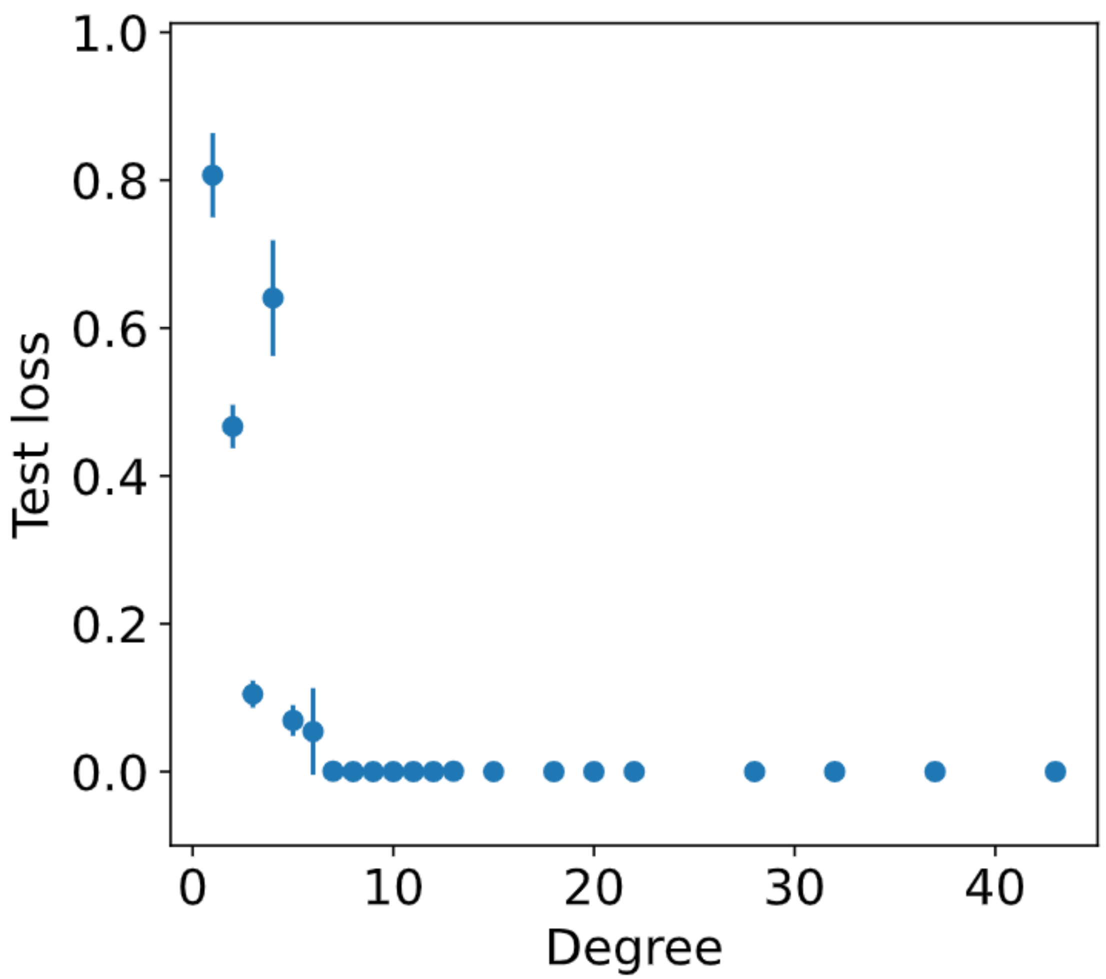
- We prove degree bias arises from variety of factors associated with node's degree, e.g., homophily, neighbor diversity

- We prove GNNs reduce loss on low-degree nodes more slowly

Motivation

- GNNs exhibit better performance for high-degree nodes on **node classification** tasks

- Privileges high-degree actors during social and content recommendation



Motivation

- Researchers have proposed various hypotheses for why GNN degree bias occurs
- We find via a survey of 38 degree bias papers that these hypotheses are often not rigorously validated, and can even be contradictory

Hypothesis	**Papers**
(H1) Neighborhoods of low-degree nodes contain insufficient or overly noisy information for effective representations.	[115], [190], [193], [53], [219], [112], [113], [118], [116], [84], [110], [72], [109], [195], [222], [174], [46], [31], [76], [220], [197]
(H2) High-degree nodes have a larger influence on GNN training because they have a greater number of links with other nodes, thereby dominating message passing.	[163], [190], [219], [87], [208], [109], [209]
(H3) High-degree nodes exert more influence on the representations of and predictions for nodes as the number of GNN layers increases.	[219], [29], [106], [46], [210]
(H4) In semi-supervised learning, if training nodes are picked randomly, test predictions for high-degree nodes are more likely to be influenced by these training nodes because they have a greater number of links with other nodes.	[163], [208], [71]
(H5) Representations of high-degree nodes cluster more strongly around their corresponding class centers, or are more likely to be linearly separable.	[122], [178], [105]

Test-Time Degree Bias

Theorem 1. Consider a test node i with label $Y_i = c$. Furthermore, consider a label $c' \neq c$. Let $\mathbb{P}\left(\ell(\mathscr{M} \mid i, c) > \ell(\mathscr{M} \mid i, c')\right)$ be the probability of any model $\boxed{\mathscr{M}}$ misclassifying i. Then:

GNN, MLP, logistic regression, etc.

$$\mathbb{P}\left(\ell(\mathscr{M} \mid i, c) > \ell(\mathscr{M} \mid i, c')\right) \leq \frac{1}{1 + \boxed{R_{i,c'}}},$$

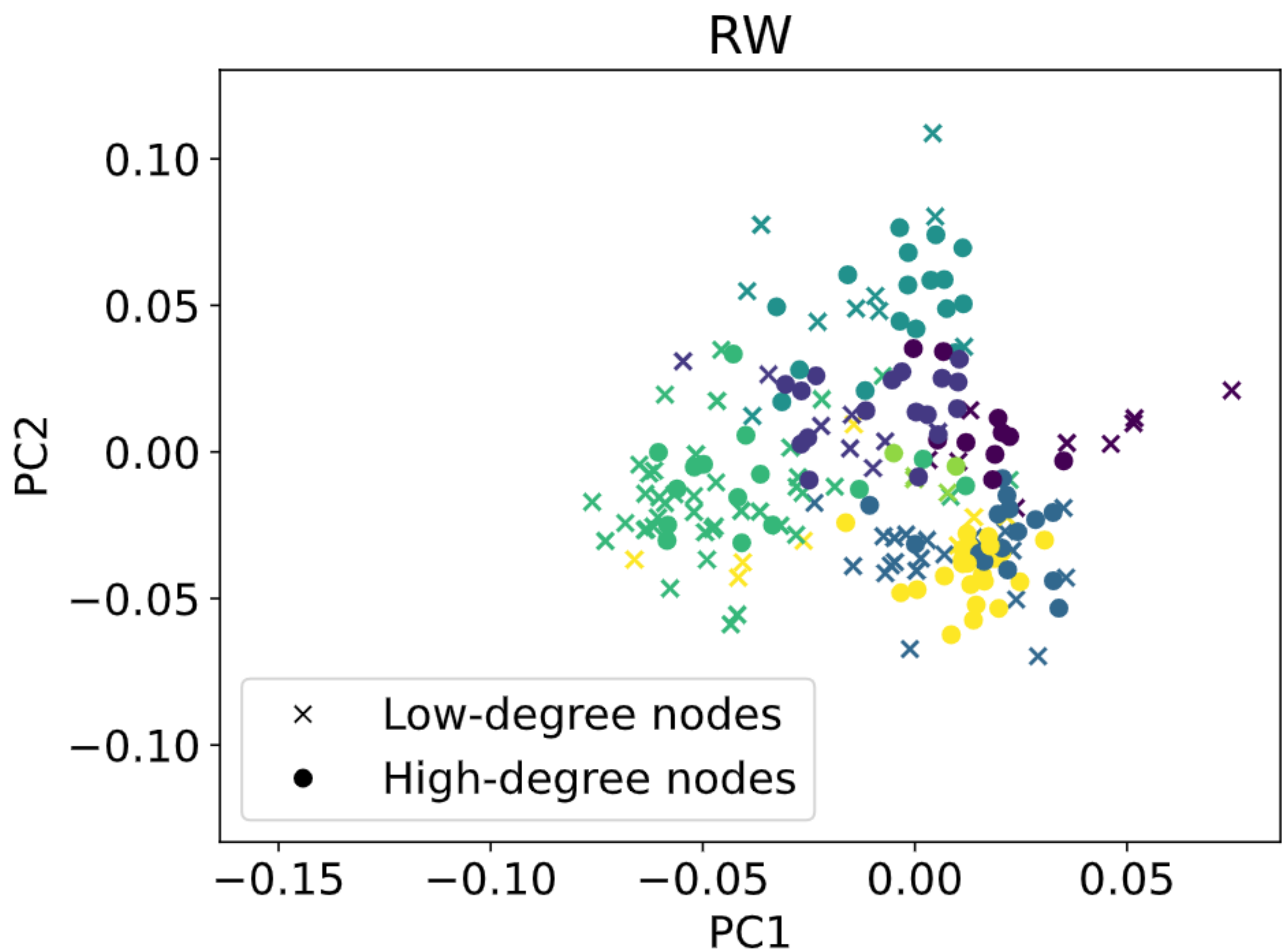
normalized measure of dispersion often used in economics to quantify inequality

where the **squared inverse coefficient of variation**

$$R_{i,c'} = \frac{\left(\mathbb{E}\left[Z_{i,c'}^{(L)} - Z_{i,c}^{(L)}\right]\right)^2}{\mathrm{Var}\left[Z_{i,c'}^{(L)} - \boxed{Z_{i,c}^{(L)}}\right]}.$$

c logit for node i

Visualization: Test-Time Degree Bias



RW: Test-Time Degree Bias

Theorem 2. $\forall l \in [L], \forall j \in \mathcal{V}$, $\textbf{Var}_{x \sim \mathcal{D}_{Y_j}}\left[x^T w^{(l)}_{c'-c}\right] \leq M$.

Then:

Prediction homogeneity

$$R_{i,c'} \geq \frac{\left(\sum_{l=0}^{L} \boxed{\beta^{(l)}_{i,c'}}\right)^2}{M(L+1)\sum_{l=0}^{L} \boxed{\alpha^{(l)}_i}}.$$

Collision probability

RW:

l-hop Prediction Homogeneity

$$\beta^{(l)}_{i,c'} = \mathbb{E}_{j\sim \boxed{\mathcal{N}^{(l)}(i)}}\left[\mathbb{E}_{x\sim\mathcal{D}_{Y_j}}\left[x^T \boxed{w^{(l)}_{c'-c}}\right]\right]$$

Distribution over terminal nodes of length-l random walks starting from i

Boundary that separates classes c and c':

$$w^{(l)}_{c'-c} = W^{(l)}_{.,c'} - W^{(l)}_{.,c}$$

High level: measures expected prediction score for nodes j, weighted by probability of being reached by length-l random walk starting from i

RW:
l-hop Collision Probability

$$\alpha_i^{(l)} = \sum_{j \in \mathcal{V}} \left[\left(P_{\mathrm{rw}}^l \right)_{ij} \right]^2$$

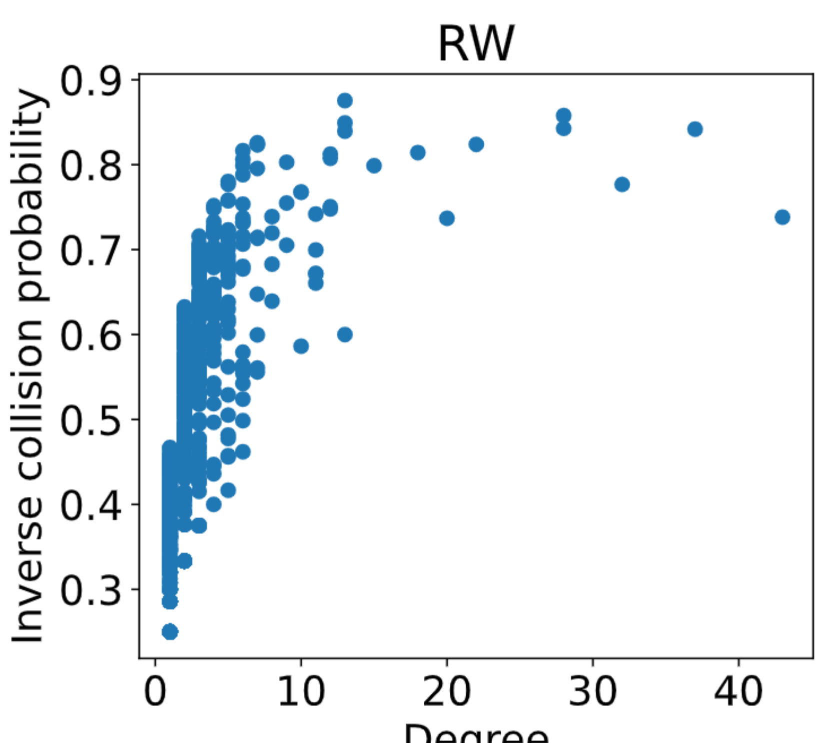
- **High level:** quantifies probability of two length-l random walks starting from i colliding at same end node j
- When collision probability is lower, random walks are more *diverse*

RW: Test-Time Degree Bias

- To make $R_{i,c'}$ larger (i.e., minimize probability of misclassification), sufficient (although not necessary) that $\frac{1}{\sum_{l=0}^{L} \alpha_i^{(l)}}$ is larger

$$\alpha_i^{(l)} = \sum_{j \in \mathcal{V}} \left[\left(P_{\mathrm{rw}}^{l} \right)_{ij} \right]^2$$

  - Indicates more diverse and possibly informative L-hop neighborhood

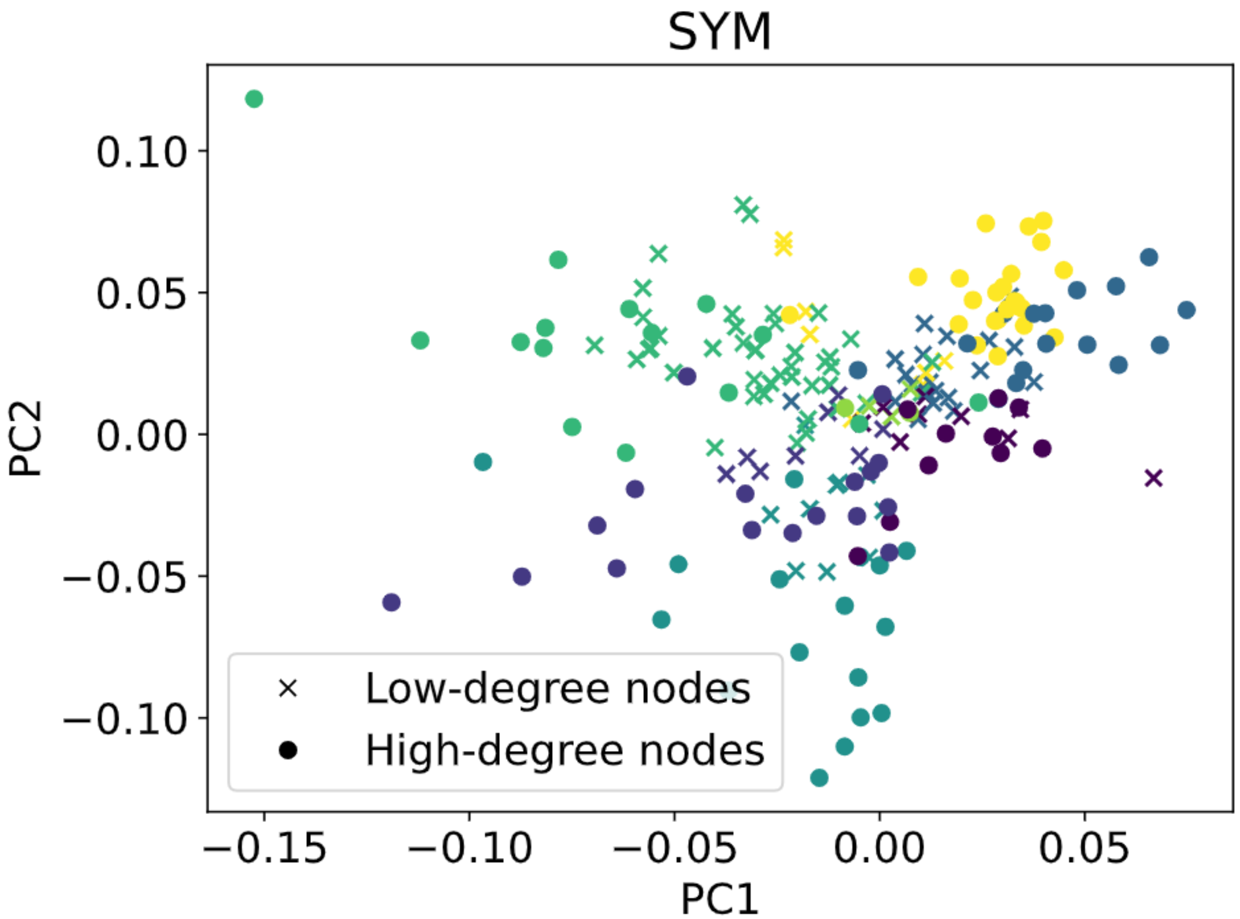


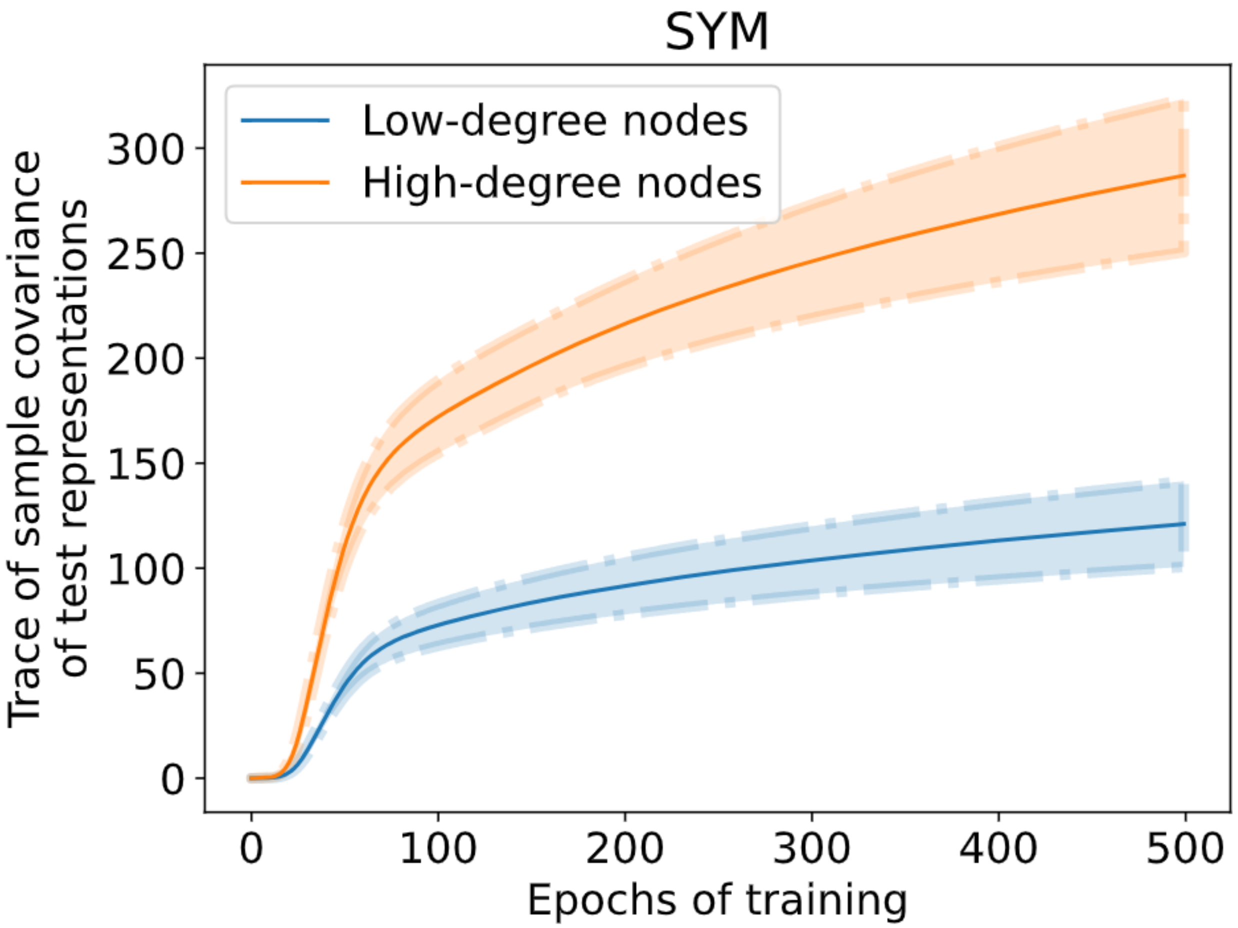
RW: Test-Time Degree Bias

$$\beta_{i,c'}^{(l)} = \mathbb{E}_{j \sim \mathcal{N}^{(l)}(i)} \left[\mathbb{E}_{x \sim \mathcal{D}_{Y_j}} \left[x^T w_{c'-c}^{(l)} \right] \right]$$

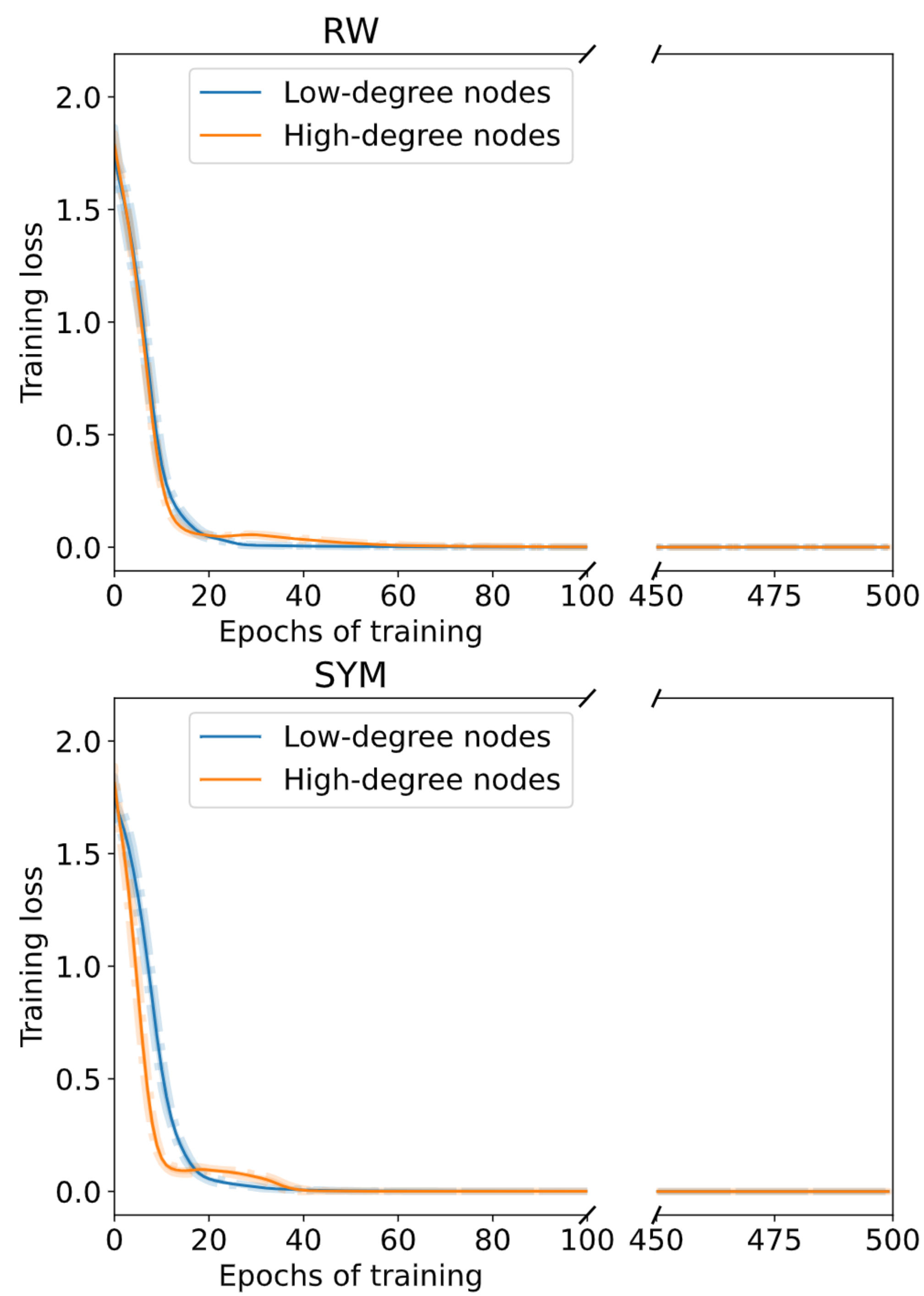
- To make $R_{i,c'}$ larger, it is sufficient that for all $l \in [L]$, $\beta_{i,c'}^{(l)}$ is more negative:
    - e.g., when more nodes in l-hop neighborhood of i are in class c and were part of training set S

Visualization: Test-Time Degree Bias





Visualization: Training-Time Degree Bias



SYM: Why do we care?

- As GNNs are applied to increasingly large networks, only few epochs of training may be possible due to limited compute
    - Which nodes receive superior utility from limited training?
- GNNs may serve as efficient lookup mechanism for nodes in deployed systems
    - If partially-trained, can perform poorly for low-degree nodes

SYM:
Training-Time Degree Bias

Theorem 2. The change in loss for i after an arbitrary training step t obeys:

$$\left| \ell[t+1](\overline{\mathsf{SYM}} \,|\, i, c) - \ell[t](\overline{\mathsf{SYM}} \,|\, i, c) \right| \leq C[t] \boxed{\sqrt{D_{ii}}} \sum_{l=0}^{L} \left\| \boxed{\widetilde{\chi}_i^{(l)}[t]} \right\|_2$$

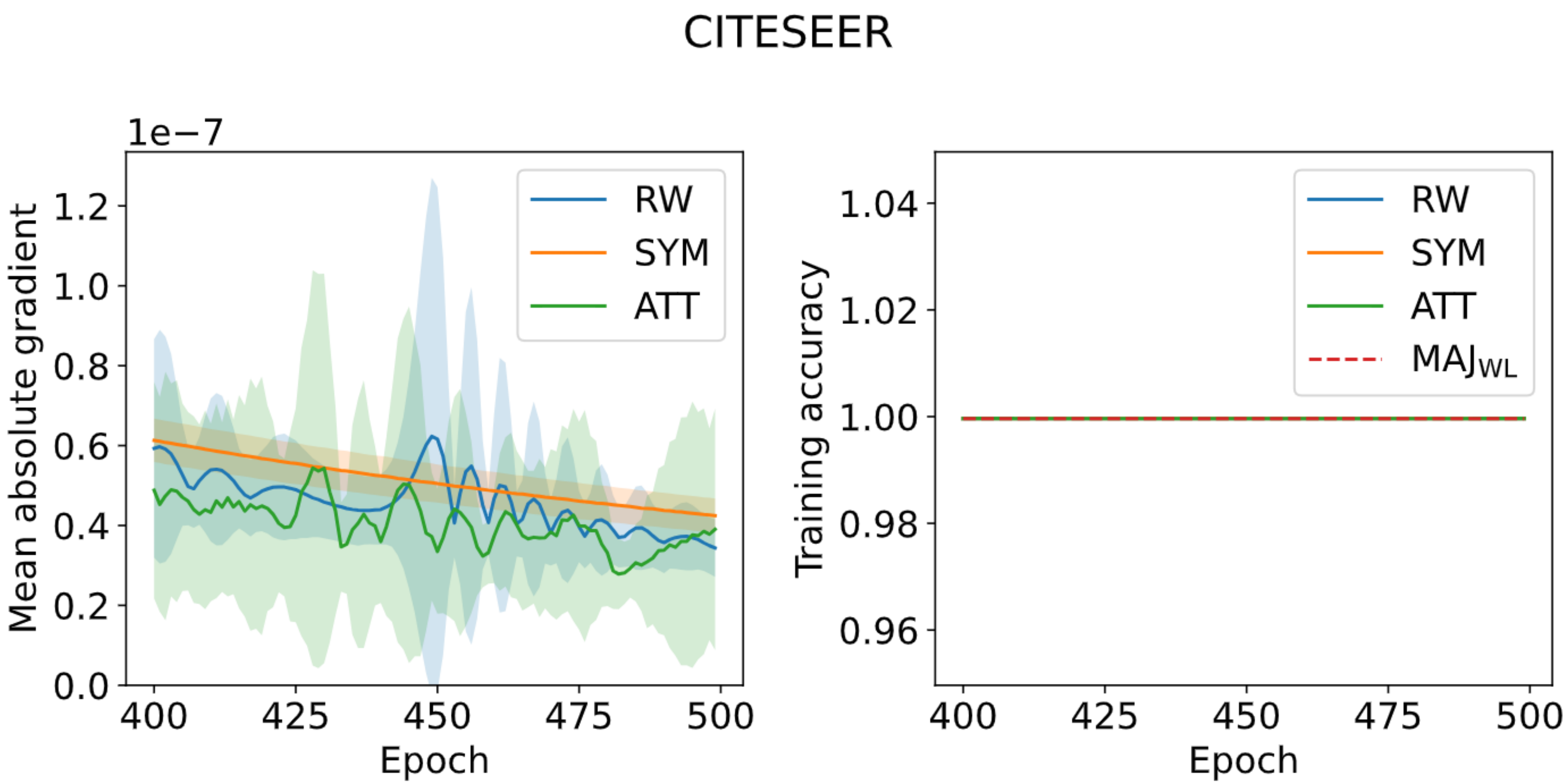
Expected similarity between neighbors of node i
and nodes in training batch $B[t]$

$$\forall m \in B[t], \left(\widetilde{\chi}_i^{(l)}[t]\right)_m = \sqrt{D_{mm}} \mathbb{E}_{j \sim \mathcal{N}^{(l)}(i), k \sim \mathcal{N}^{(l)}(m)} \left[\frac{1}{\sqrt{D_{jj} D_{kk}}} X_j X_k^T \right]$$

SYM:
Training-Time Degree Bias

- Change in loss for i after arbitrary training step has smaller magnitude if i is low-degree
- Loss for i changes more slowly when features of nodes in its L-hop neighborhood are not similar to the features in L-hop neighborhoods of nodes in training batch

SYM:
Training-Time Degree Bias



Principled Roadmap: Theoretically-Informed Criteria

- Maximizing inverse collision probability of low-degree nodes
- Increasing L-hop prediction homogeneity of low-degree nodes
- Minimizing distributional differences in representations of low and high-degree nodes
- Reducing training discrepancies with regards to rate at which GNNs learn for low vs. high-degree nodes

Conclusion

- **Contributions:**
  - Unify and distill hypotheses for origins of GNN degree bias
  - Prove degree bias arises from homophily, diversity, etc. of neighbors
  - Prove during training, some GNNs may adjust loss on low-degree nodes more slowly

Thank you!