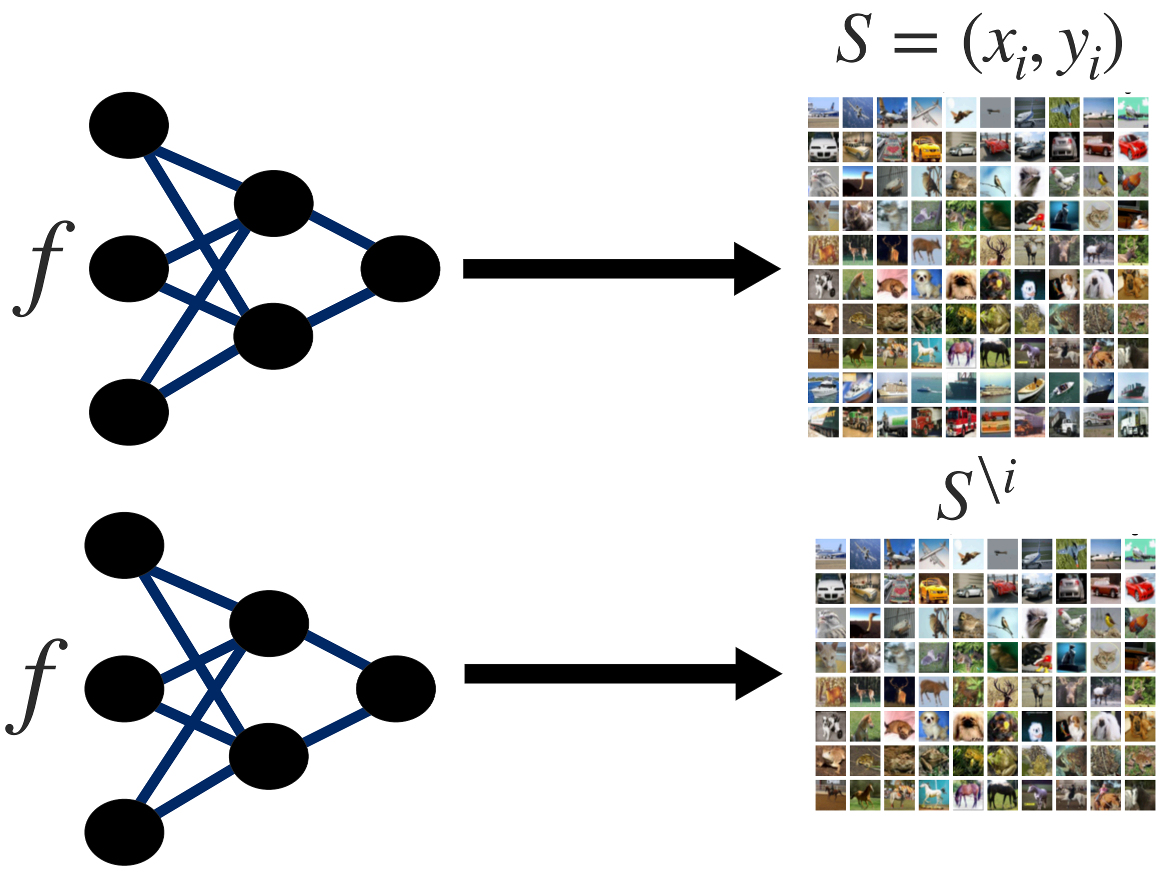
Memorization in Graph Neural Networks

Adarsh Jamadandi[1], Jing Xu[2], Adam Dziedzic[2] and Franziska Boenisch[2]

CNRS, IRISA, Rennes[1]
CISPA Helmholtz Centre for Information Security, Saarland, Germany[2]

Measuring Memorization

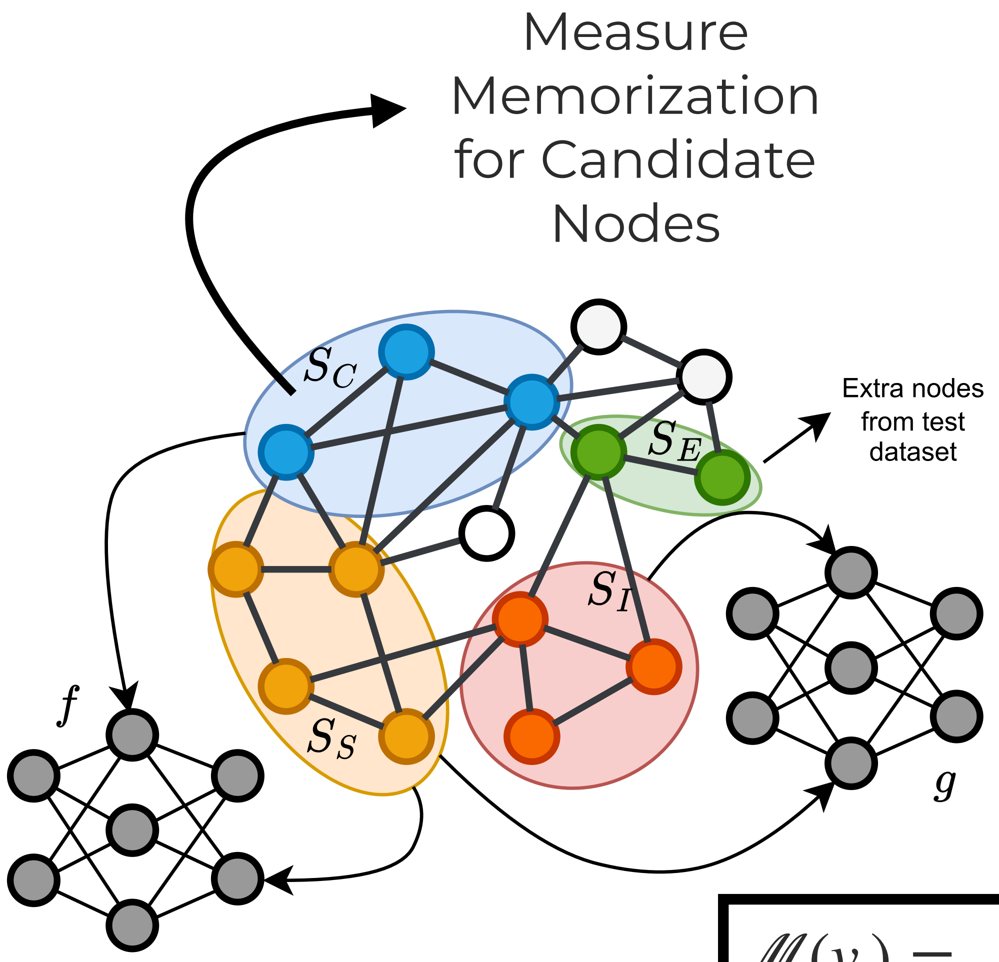
- Train model f on S and on $S^{\backslash i}$.
- Compare the behavior of the two models.
- A model needs to see the label of the sample to correctly predict the label $\rightarrow$ Memorized.



$$\boxed{\mathcal{M}(x_i) = \mathop{\mathbb{E}}_{f\sim\mathcal{T}(S)}[\Pr[f(x_i) = y_i]] - \mathop{\mathbb{E}}_{f\sim\mathcal{T}(S\backslash x_i)}[\Pr[f(x_i) = y_i]]}$$

Credits: [Does Learning Require Memorization? A Short Tale about a Long Tail]

Measuring Memorization in GNNs



- Models $f = S_s \cup S_c$ and $g = S_s \cup S_I$ are trained on various subsets of nodes.
- We isolate the effect of one node's label on model behavior.

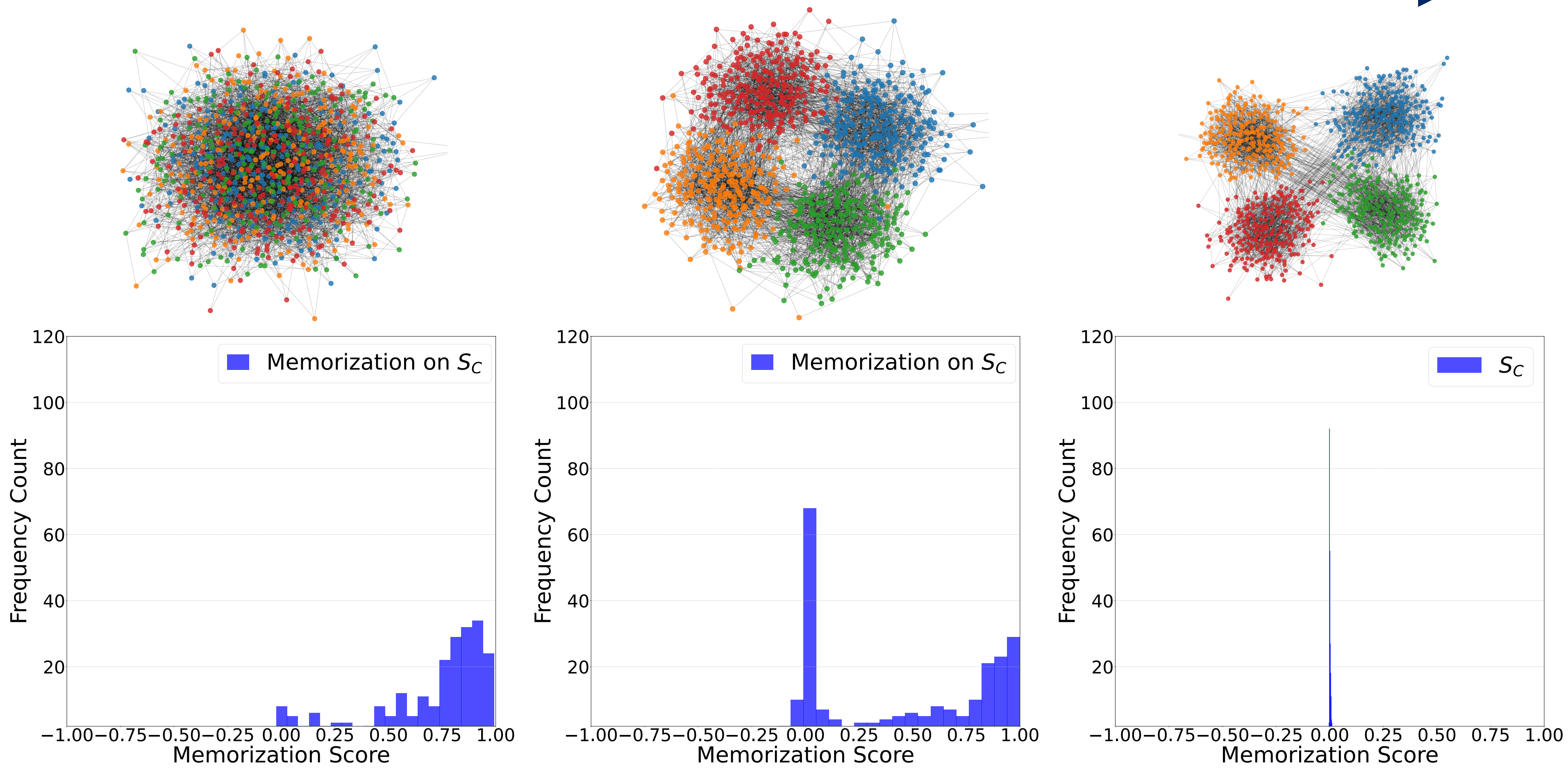
$$\boxed{\mathscr{M}(v_i) = \mathop{\mathbb{E}}_{f \sim \mathscr{T}(S)}[\Pr[f(v_i) = y_i]] - \mathop{\mathbb{E}}_{g \sim \mathscr{T}(S \backslash x_i)}[\Pr[g(v_i) = y_i]]}$$

Credits: [Does Learning Require Memorization? A Short Tale about a Long Tail]

Homophily and Memorization

Low Homophily

High Homophily



High Memorization

Low Memorization

Explaining Memorization in GNNs

- We uncover 3 internal mechanisms to explain the emergence of memorization in GNNs.
- Graph homophily - like nodes connected to like nodes.
- Implicit bias of GNNs to leverage the graph structure.
- Label-Feature Inconsistency.

Analyzing the Training Dynamics of GNNs

- We will define an alignment metric, cosine-similiarity-like applied to matrices given by $\mathscr{A}(\mathbf{K_1}, \mathbf{K_2}) = \dfrac{\langle \mathbf{K_1}, \mathbf{K_2} \rangle_F}{||\mathbf{K_1}||_F ||\mathbf{K_2}||_F}$
- Adjacency Matrix: $\mathbf{A}$
- Optimal Kernel Matrix: $\mathbf{\Theta}^* = \bar{\mathbf{Y}}\bar{\mathbf{Y}}^T$
- NTK Matrix: $\mathbf{\Theta}_t^l(x, \tilde{x}; \mathbf{A}) = \nabla_W f(x; \mathbf{A})^T \cdot \nabla_W f(\tilde{x}; \mathbf{A})$

Credits: How Graph Neural Networks Learn: Lessons from Training Dynamics

We Will Track

Kernel-Graph Alignment

- Alignment between NTK matrix $\mathbf{\Theta}_t$ and adjacency matrix $\mathbf{A}$ ($\mathscr{A}(\mathbf{\Theta}_t, \mathbf{A})$).

- Represents the implicit bias of GNNs to leverage the graph structure.

Kernel-Target Alignment

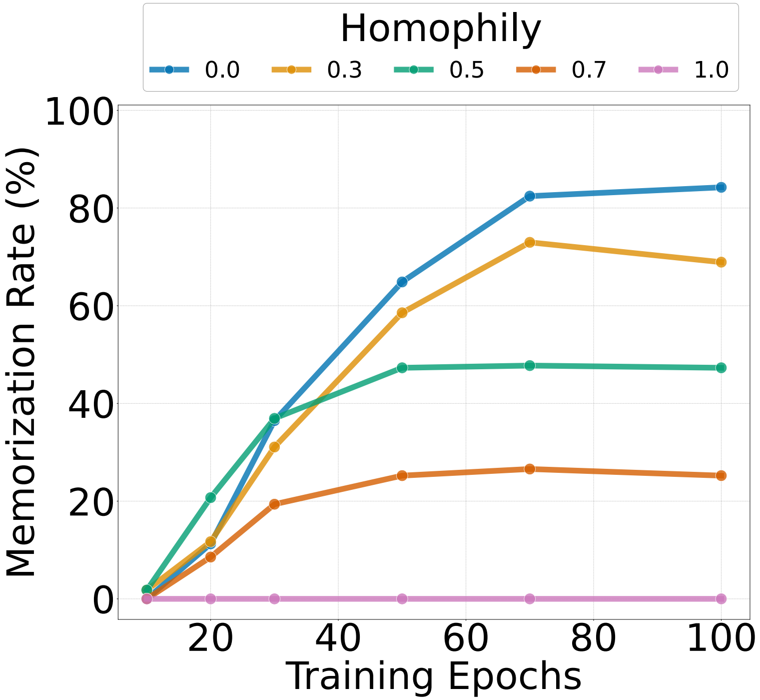
- Alignment between the NTK matrix $\mathbf{\Theta}_t$ and optimal kernel matrix $\mathbf{\Theta}^*$.

- This metric measures how well a classifier generalizes, a higher alignment implies good generalization.

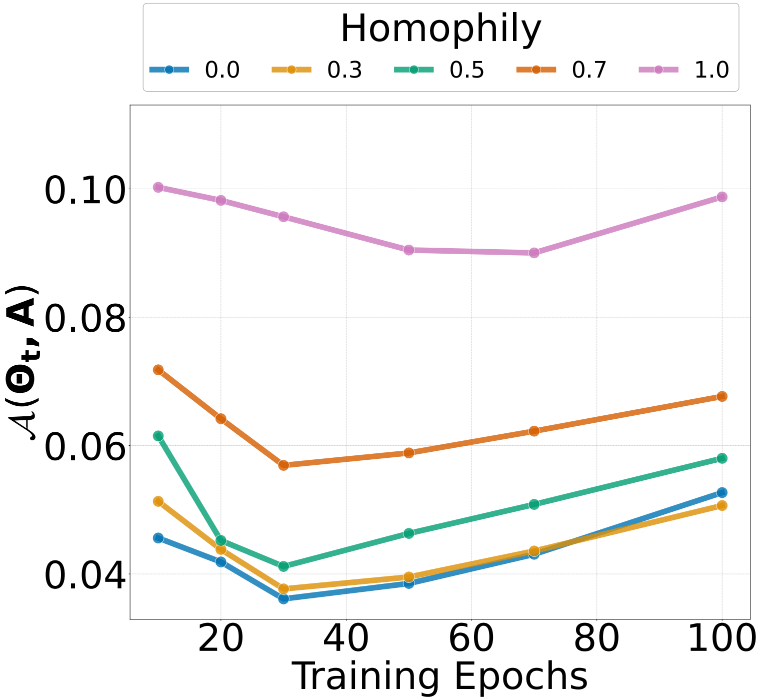
Credits: How Graph Neural Networks Learn: Lessons from Training Dynamics

Alignment Matrices for Synthetic Datasets

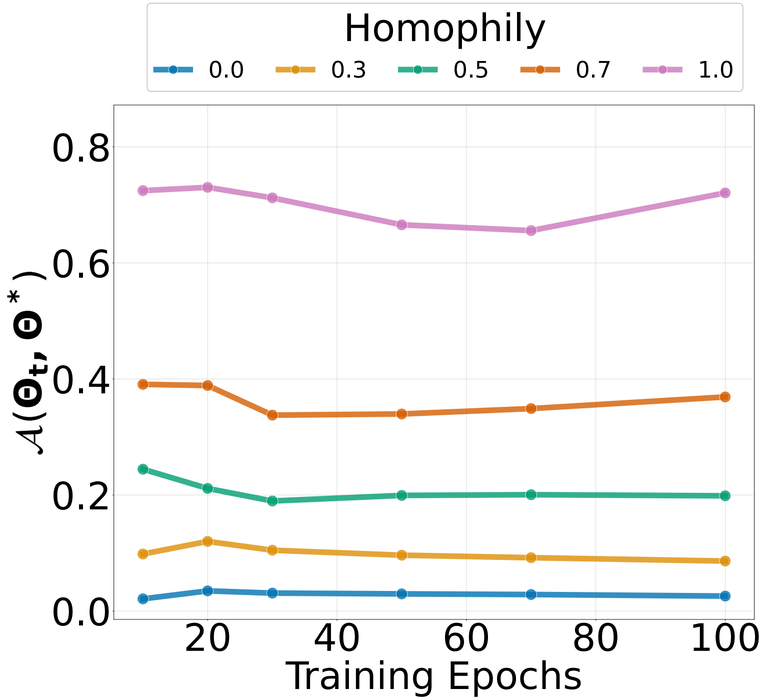
- Low homophily → Memorization rate increases.
- Low homophily (graph structure is less informative) → Still $\mathscr{A}(\mathbf{\Theta_t}, \mathbf{A})$ improves.
- Low homophily → $\mathscr{A}(\mathbf{\Theta_t}, \mathbf{\Theta}^*)$ poor, suggests memorization.



Memorization Rate



Kernel-Graph Alignment

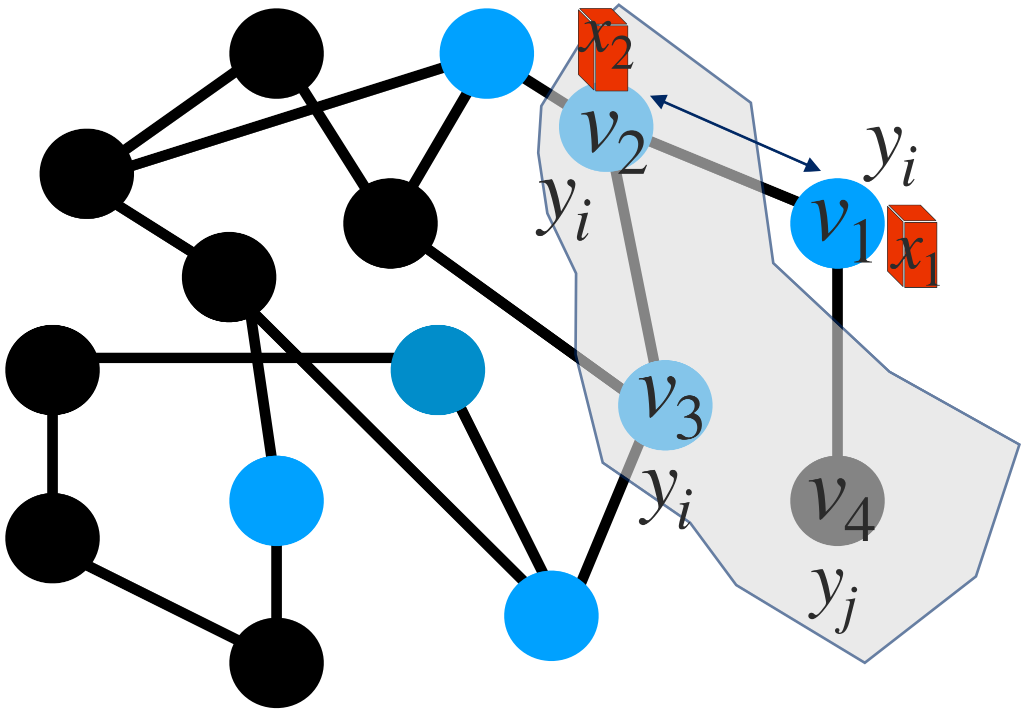


Kernel-Target Alignment

Node Atypicality

- Novel Label Disagreement Score (LDS) →local structural anomaly in the feature space of the nodes.

- Nodes with high LDS more likely to get memorized.



$$\mathrm{LDS}_k(v_i) = \frac{1}{k} \sum_{v_j \in N_k(v_i)} \mathbb{I}[y_j \neq y_i]$$

Summary

- GNNs also memorize node labels.
- Homophily$\uparrow$ Memorization Rate$\downarrow$.
- In low-homophily settings, the graph is unhelpful for the task. But GNNs have an implicit bias to use the graph structure.
- How to achieve 0 train loss? Memorize!
- Nodes with high label disagreement score usually get memorized.

Our Paper: