Linear GCNs Need Better Bias, Not More Expressive Power

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Celia Rubio-Madrigal Rebekka Burkholz

CISPA Helmholtz Center for Information Security

ICML International Conference On Machine Learning

Findings

We prove that linear GCNs are WL-expressive, but they do not perform on par with SOTA GNNs on their own.

However, equally performing solutions exist. We can find them with knowledge distillation (KD).

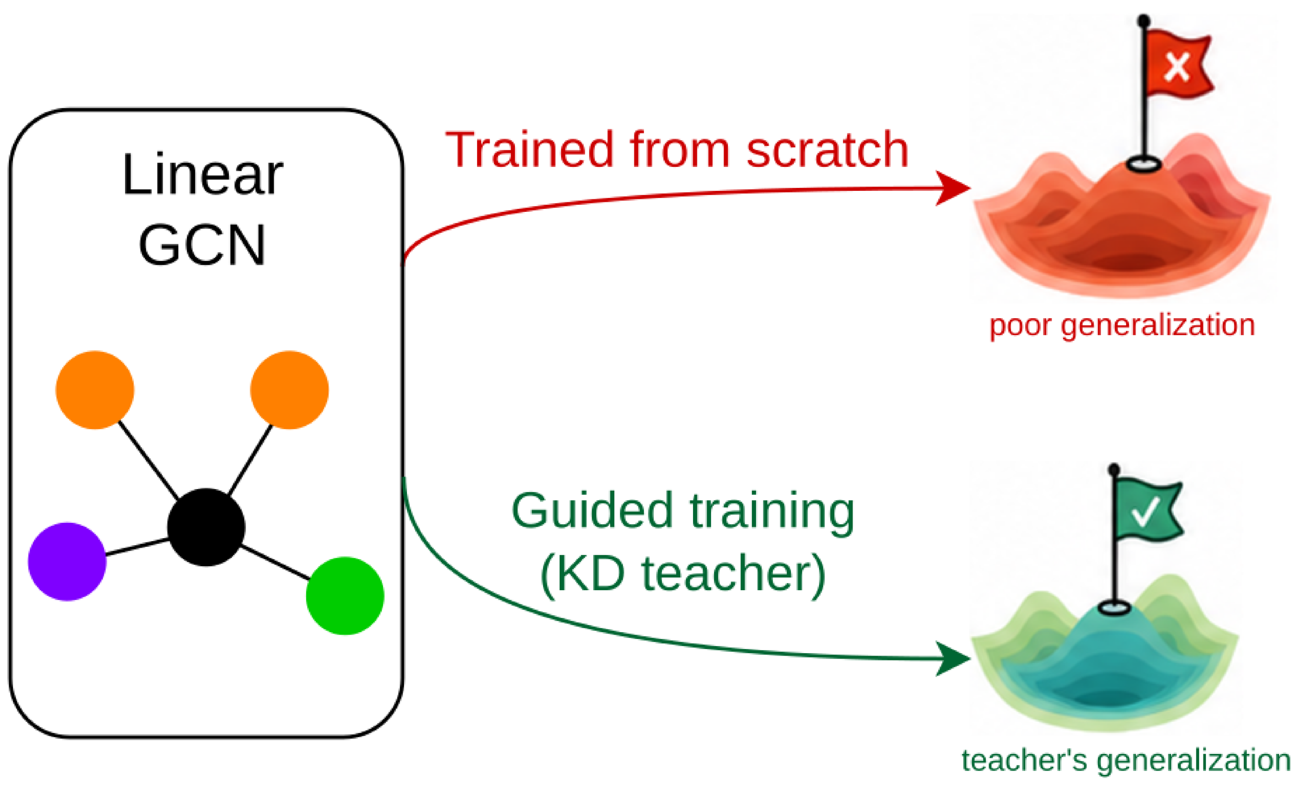
GNNs do not win because of expressiveness, but because they impose better inductive biases.

Diagnostic: KD as model selection

Linear student with labels y_{train} and teacher logits z:

$$\mathcal{L} = \mathcal{L}_{\text{label}}(\hat{y}_{\text{train}}, y_{\text{train}}) + \lambda\, \mathcal{L}_{\text{teacher}}(\hat{y}, z),$$

If the student matches the teacher, KD has selected a solution already expressible by the linear model class.



Theory: Linear GCNs are WL-expressive

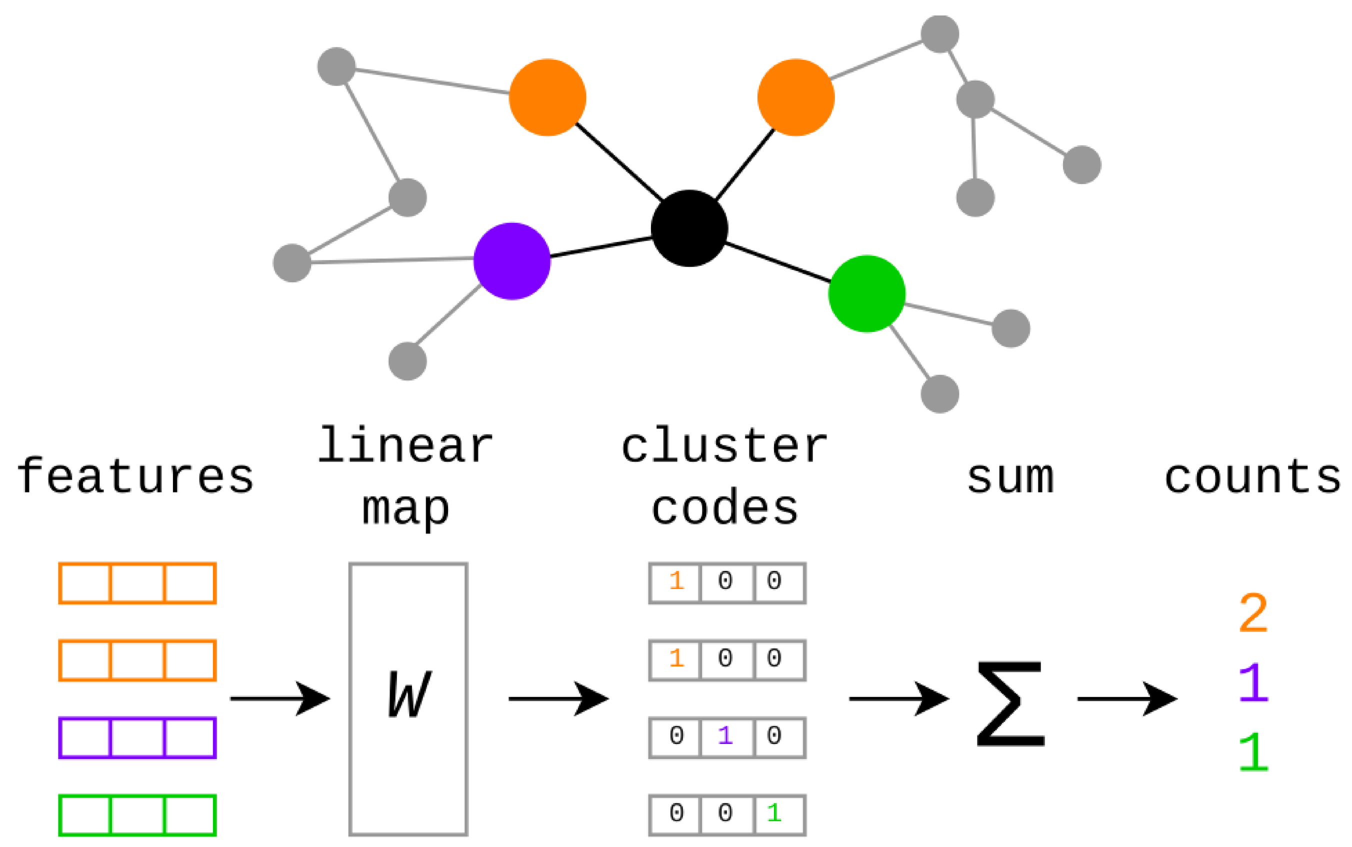
For separable node representations, a linear map can send clustered features to near one-hot codes.

$$\mathbf{X}^{(l)} = \mathbf{TC} \Rightarrow \mathbf{M}^{(l)} = [\mathbf{C}^{+}, \mathbf{0}]$$

Sum aggregation acts as counting, preserving neighborhood multisets for WL-level message passing.

Error: up to perturbation, conditioning, and degree.

$$\left\| \sum_v a_{uv} \mathbf{X}_v^{(l)} \mathbf{M}^{(l)} - \sum_v a_{uv} \mathbf{T}_v^{(l)} \right\|_2 \leq \sum_v |a_{uv}| \frac{\epsilon'}{\sigma_{\min}(\mathbf{C})}.$$



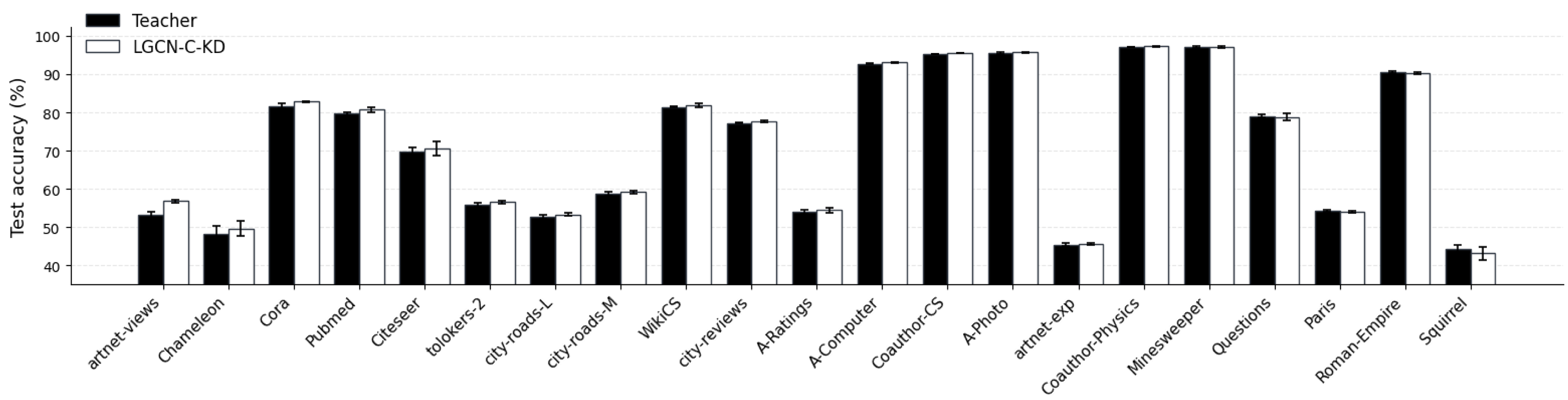
Architectures

- Teachers are well-tuned GCN, GAT, or GraphSAGE.
- Linear GCN: no ReLUs before the prediction head.
- -C: Concat all layer-wise embeddings to the head.

Empirical observations

- LGCN-KD matches teachers. Training from scratch (LGCN) or without depth (MLP-KD) underperform.
- Concat (LGCN-C-KD) helps trainability. Long-range tasks (e.g. Paris) require more width and longer training.

LGCN-C-KD matches SOTA teachers from benchmarks: Luo et al. (2024), GraphLand (2025), City-Networks (2026).



Linear GCNs (with nonlinear prediction heads) can express the right solutions; we find them with knowledge distillation.

Progress in graph learning should target implicit bias, trainability, and model selection, not only expressive capacity.