Mechanistic interpretability workshop @ ICML 2026

Yong-Min Shin | LG AI Research

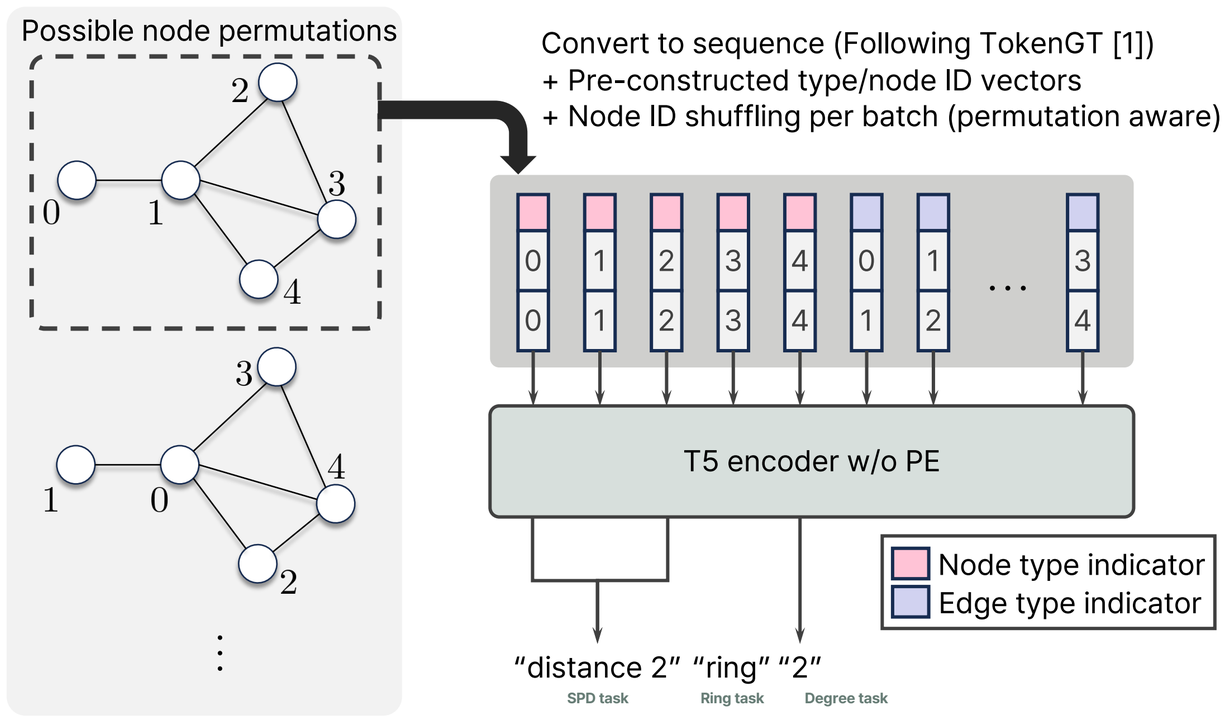
Discovering Mechanisms in Tokenized Graph Transformers

Yong-Min Shin, Dae-Woong Jeong, Sehui Han | LG AI Research / Materials Intelligence Lab

OVERVIEW

- **Extension of mechanistic interpretability to graphs.**
  To the best of our knowledge, this is the first mechanistic interpretability work applied to understand graph transformers.
- **[Observation 1] Early degree-like feature extraction.**
  We observe the model extracts degree-like information for all node token in early layers (layer 0~1).
- **[Observation 2] Task-specific composition.**
  The models build upon early degree-like features towards solving more complex tasks (ring membership, shortest-path)
- **[Observation 3] Universal node ID matching behavior.**
  Across tasks, the model first learns how to match ID vectors, behaving like a general incident-edge retrieval module at layer 0. This is crucial for permutation-aware computations.

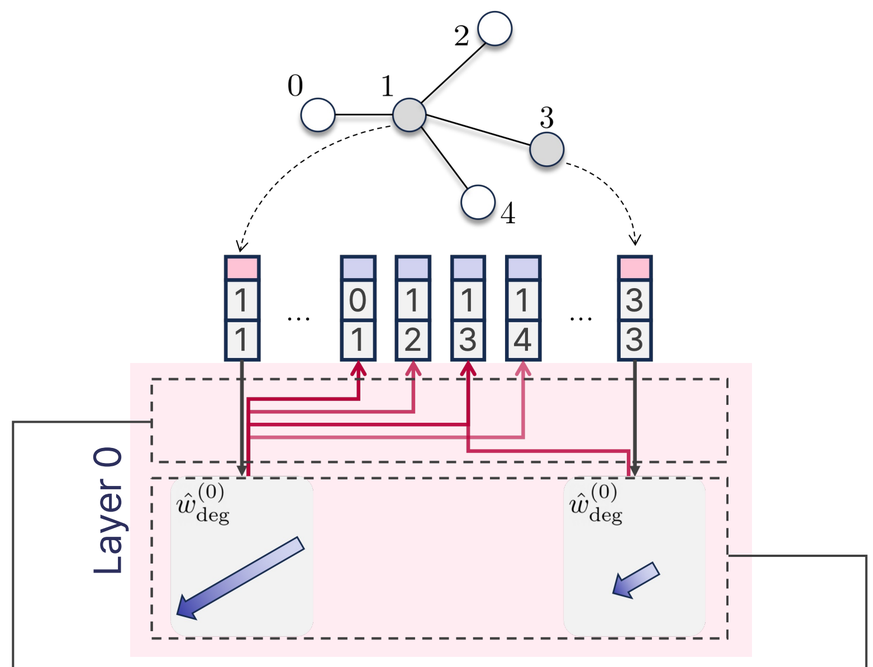
PROBLEM SETUP



[1] Kim et al., "Pure Transformers are Powerful Graph Learners", NeurIPS 2022

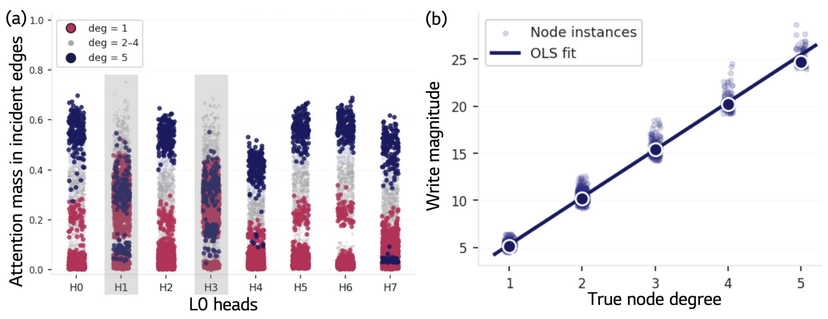
DEGREE COUNTING

- **Degree direction $\hat{w}_{\deg}^{(l)}$ at layer 0 (residual)**



Allocated attention mass to edge tokens roughly matches degree.

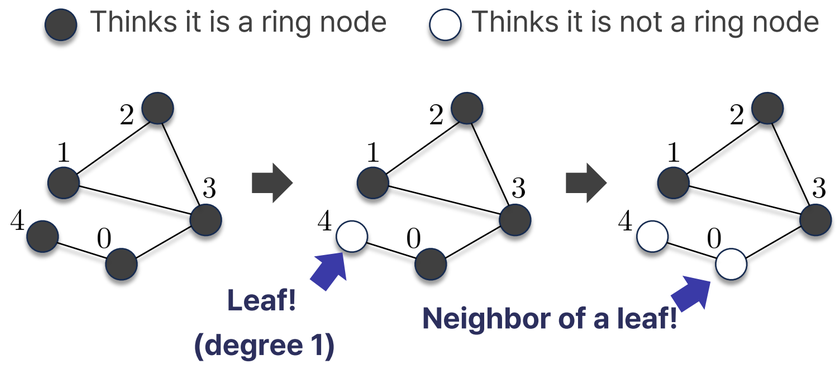
Write magnitude is **proportional to degree**.



- **Interventions along the degree direction.**
  We observe the model's performance changes for all three tasks when we add vector $\alpha\hat{w}_{\deg}^{(l)}$ at the residual stream.

α	Mean degree shift	Ring/Non-ring acc.	SPD acc.
-5	-0.9125	0.5513/1.0000	0.9612
-3	-0.5437	0.9708/1.0000	0.9898
0	$+0.0000$	1.0000/0.9986	0.9891
$+3$	-0.1958	1.0000/0.8622	0.9116
$+5$	$+0.2271$	1.0000/0.0000	0.7539

RING MEMBERSHIP



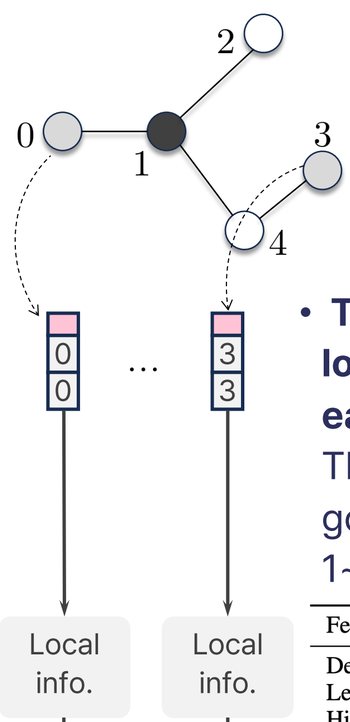
- **The model assumes all nodes are part of a ring by default.**
  We observe several evidence to support this. First, the ring membership task is nearly solved at layer 1, a much shallower position than expected. Second, zero-ablation on layer 1 mainly drops non-ring accuracy of the model.

Ring size	Logit margin	Accuracy	Δ accuracy
3	$+12.2$	100.0%	-0.0%
4	$+11.8$	99.0%	-1.0%
5	$+11.4$	99.1%	-0.9%
6	$+9.8$	94.4%	-5.6%
7	$+11.4$	95.5%	-4.5%
8	$+9.3$	100.0%	-0.0%
15	$+1.8$	100.0%	-0.0%
Non-ring	–	63.6%	$\mathbf{-36.4\%}$

- **The model uses leaf-neighbor evidence.**
  By training a linear probe, we observed that to override the ring assumption, the model looks if any of the neighbors are leaf nodes (degree 1).
- **Explicit algorithm test.**
  This shows that the approach is valid (95% acc.)

```
def predict_ring(v, adj, degree, threshold):
    """Explicit leaf-fraction algorithm (best single-threshold variant)."""
    if degree[v] < 2:
        return 0  # non-ring: leaves cannot be on a cycle
    frac_leaf = sum(1 for u in adj[v] if degree[u] == 1) / degree[v]
    return 1 if frac_leaf <= threshold else 0  # ring if few/no leaf neighbors
```

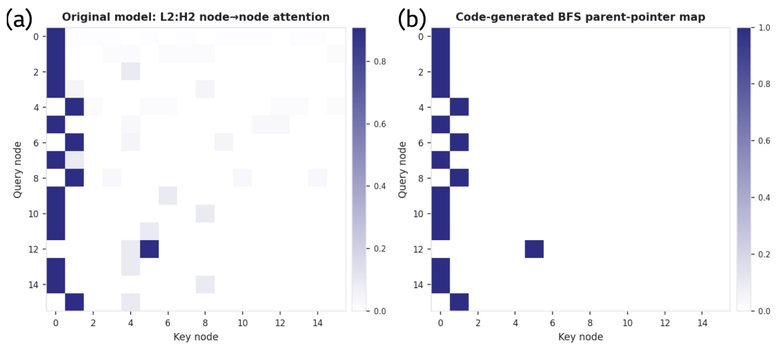
SHORTEST-PATH DISTANCE



- **The model constructs a local structural features in early layers (L0/L1).**
  This can effectively provide good evidence for distance 1~2 nodes.

Feature	Embed	L1-post
Degree	0.405	0.942
Leaf neigh. count	0.737	0.824
High-degree neigh. count	0.566	0.842

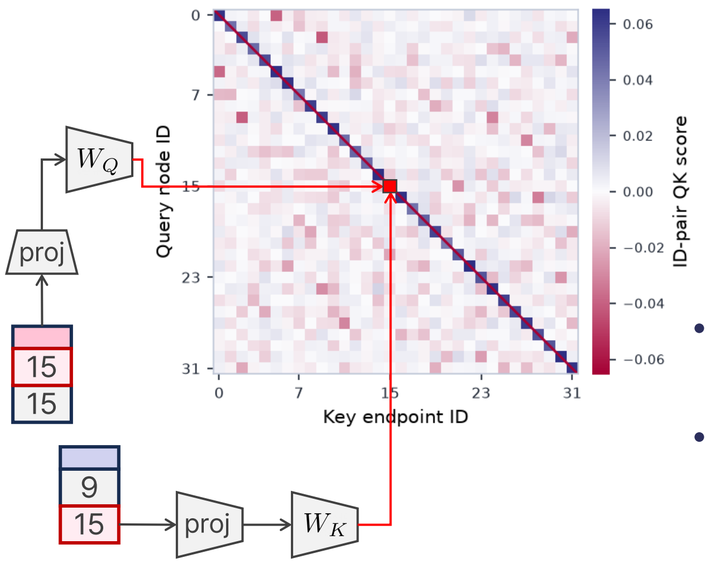
- **L2:H2 (BFS-parent copy head) selectively copies information** (including degree) **from the parent node.**



- **L3 + task head are responsible for refining more distant node pairs.**

SPD	Clean	L3-attn ablated	Δ acc.
1 (adjacent)	1.000	0.987	-0.013
2	0.999	0.869	-0.130
3	0.996	0.880	-0.116
4	0.990	0.605	$\mathbf{-0.385}$
5+	0.985	0.299	$\mathbf{-0.686}$

NODE ID MATCHING BEHAVIOR



Task	Sel(M) > 3?	Strongest runtime heads
DC	8/8	H2 2.31σ, H4 2.17σ, H5 1.95σ
RM	8/8	H4 2.27σ, H2 1.92σ, H7 1.59σ
SPD	7/8	H5 2.34σ, H0 2.15σ, H7 1.84σ

- **Heads that identify same node IDs in other tokens universally exist at layer 0 across tasks.**
- **This is even possible when # of IDs > QK dim., since a soft 'nearly' orthogonal matching is sufficient.**

LIMITATIONS & DISCUSSIONS

- **Limited experimental settings.**
  We have chosen a vanilla transformer with TokenGT-style input, with small synthetic graphs. Future directions for more general models, graphs, and tokenization are important. Also, several aspects (e.g., node ID matching emergence) needs to be investigated.
- **One of the first mechanistic interpretation of how a graph transformer model understands and process graph input.**