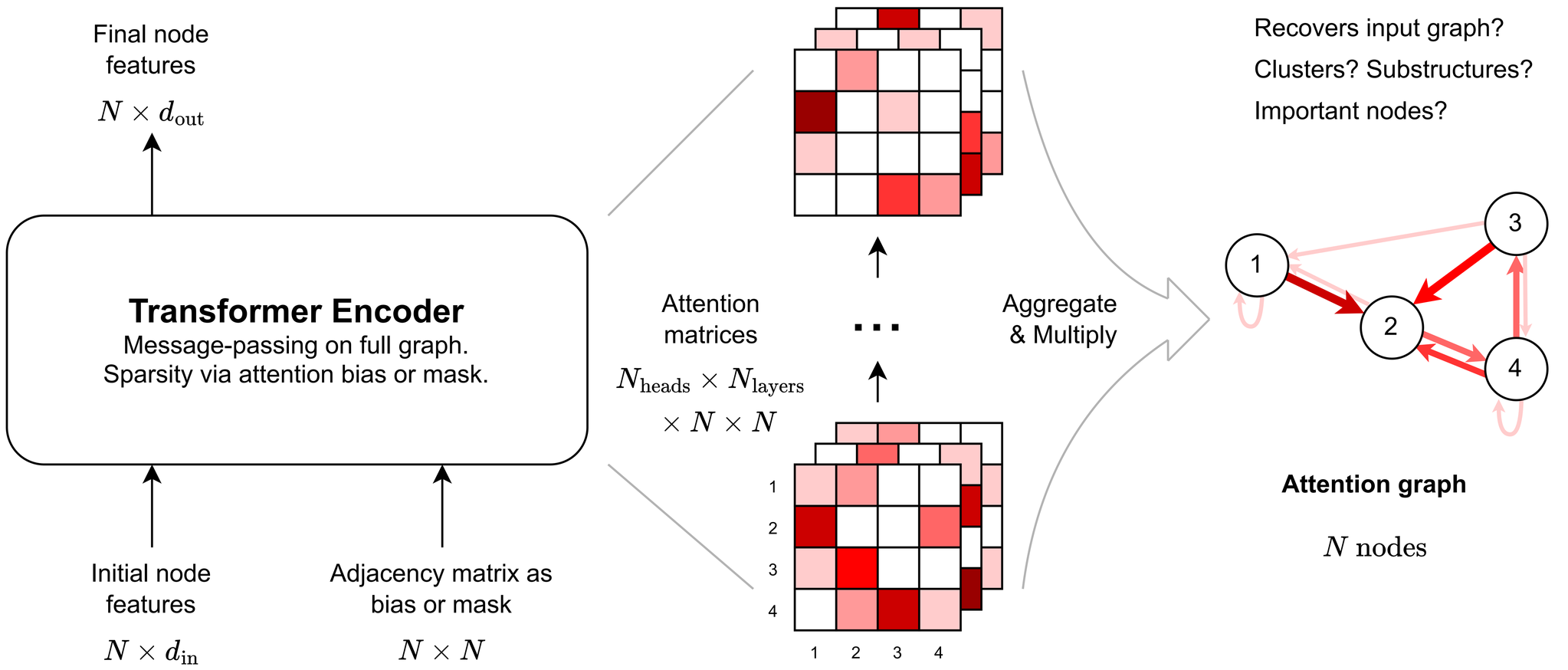
Towards Mechanistic Interpretability of Graph Transformers via Attention Graphs

Batu El*, Deepro Choudhury*, Pietro Liò, Chaitanya K. Joshi

Attention Graphs: network science-based mechanistic interpretability

We study the **graph structure of attention** from a **network science** perspective to mechanistically understand how GNNs and Graph Transformers learn!



1 **GNNs** and **Transformers** are mathematically equivalent with optional sparsity.

2 **Attention matrices** tell us how information flows among input tokens.

3 **Attention graphs:** directed graphs of information flow, aggregated across layers & heads.

Design space of Graph Transformers

Unifies the study of a large class of practically used GNN and Graph Transformer variants.

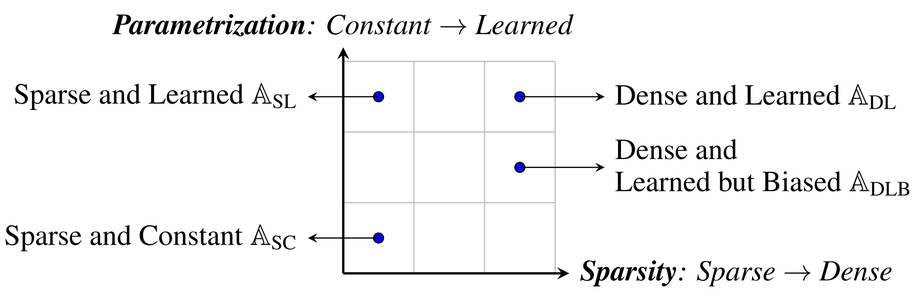


Figure 2: **Design space of Graph Transformers** based on two key dimensions: (1) sparsity of attention (sparse vs. dense) and (2) parametrization of attention (constant vs. learned).

Examples across scientific discovery:
AlphaFold, GraphCast, ProteinMPNN, ML Force Fields

From attention matrices to graphs

Each forward pass through a Graph Transformer creates N_{heads} x N_{layers} x N_{tokens} N_{tokens} matrices.

Across heads: simple averaging

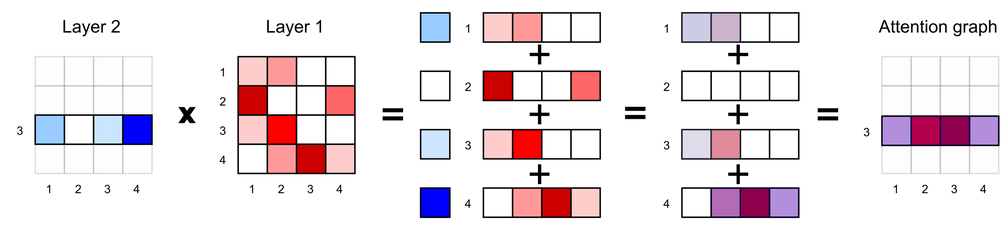


Figure 3: **Aggregating attention across layers by matrix multiplication.** Attention matrices from successive layers are combined to capture indirect information flow. For node i, row i in the attention matrix $\mathbb{A}_{L_2}$ represents how much it attends to each intermediate node j. Each row j in $\mathbb{A}_{L_1}$ captures how those intermediate nodes attend to other nodes k. Matrix multiplication $\mathbb{A}_{L_2}\mathbb{A}_{L_1}$ combines these patterns, revealing how node i indirectly attends to node k through intermediate nodes j.

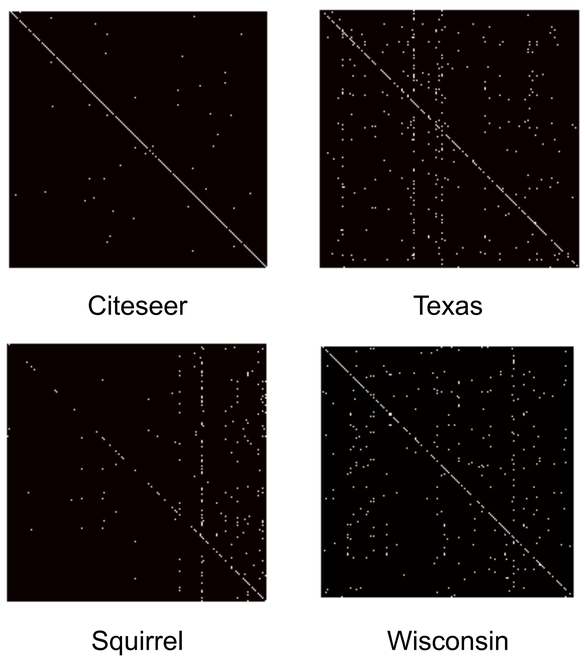
Different graph inductive biases in Transformers → distinct algorithms

We plot adjacency matrices from Attention Graphs.

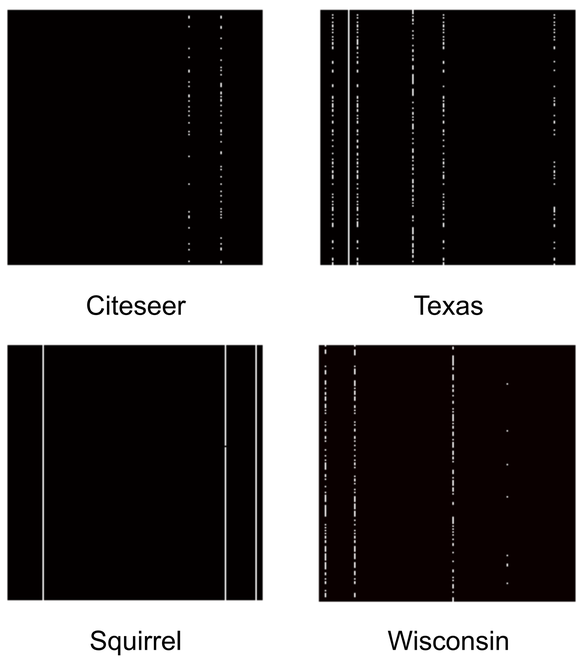
White = edge

Black = no edge

Dense, Learnt, Biased Graph Transformers — DLB
learn strong self-attention patterns (diagonals)



Dense & Learnt Graph Transformers — DL
learn to focus attention on reference nodes (verticals)



Lot's more in the paper:

batuel@stanford.edu, deepro.choudhary@stats.ox.ac.uk

github.com/batuel/understanding-inductive-biases-of-gnns