Logical Characterizations of Graph Transformers

Veeti Ahvonen

Tampere University

Mathematics Research Centre

9.1.2026

This talk is based on joint work "Expressive Power of Graph Transformers via Logic" with Maurice Funk, Damian Heiman, Antti Kuusisto and Carsten Lutz, to appear in AAAI 2026.

Background

Transformers form the basis of modern large language models (LLMs) such as ChatGPT, Copilot, etc. (Vaswani et al., NeurIPS 2017).

Little is known about their precise expressive power on graphs.

Standard transformers have been studied via temporal logics by Yang et al. (NeurIPS 2024), Chiang et al. (ICML 2023), Li and Cotterel (NeurIPS 2025), Jerad et al. (ACL 2025).

Background

We study Graph Transformers (GTs) with reals and floating-point numbers.

Restricted to first-order logic, we characterize real-based GTs.

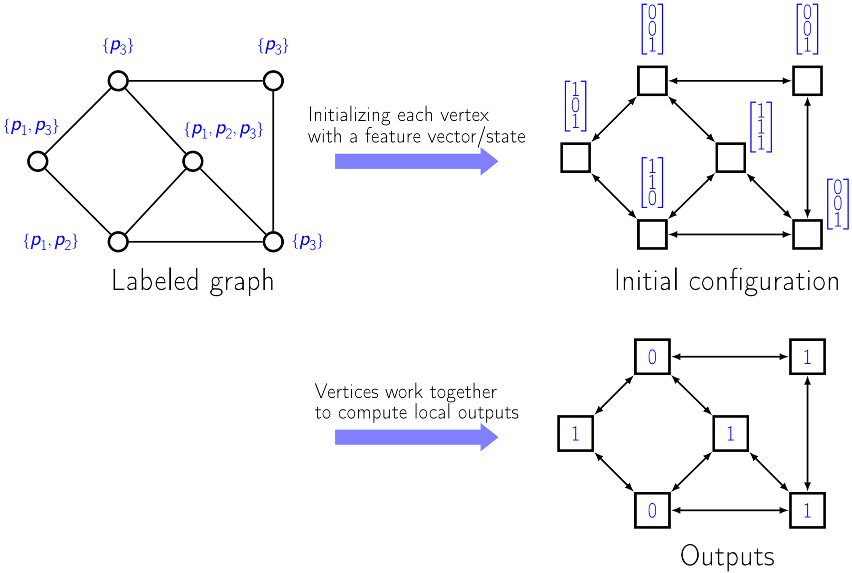
We give an unrestricted logical characterization of float-based GTs.

Why is this interesting?

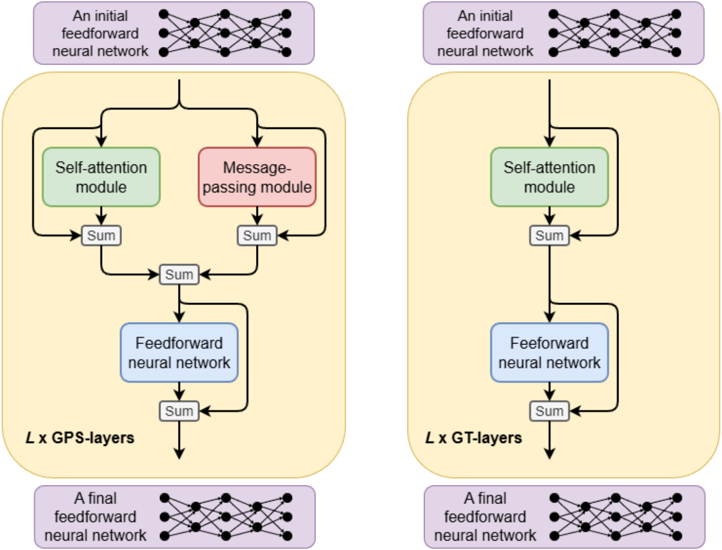
$\Longrightarrow$ Helps to understand theoretical limits of graph transformers.

$\Longrightarrow$ Helps practitioners select appropriate GT architectures.

Graphs


{p_3}
{p_3}
{p_1, p_3}
{p_1, p_2, p_3}
{p_1, p_2}
{p_3}
Labeled graph
Initializing each vertex
with a feature vector/state
$\begin{bmatrix}0\\0\\1\end{bmatrix}$ $\begin{bmatrix}0\\0\\1\end{bmatrix}$ $\begin{bmatrix}1\\0\\1\end{bmatrix}$ $\begin{bmatrix}1\\1\\1\end{bmatrix}$ $\begin{bmatrix}1\\1\\0\end{bmatrix}$ $\begin{bmatrix}0\\0\\1\end{bmatrix}$
Initial configuration
Vertices work together
to compute local outputs
0 1 1 1 0 1
Outputs

GPS-networks and Graph Transformers



Message passing

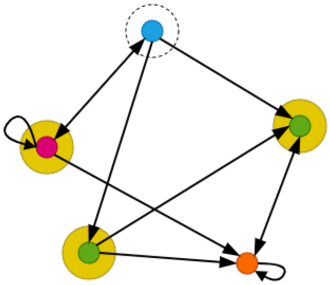


Figure: A graph where feature vectors are identified with colors.

Each vertex updates its feature vector using its previous features and the aggregated features of its out-neighbors

A demonstration of how the blue vertex's feature vector is updated:

$\bullet = \mathrm{COM}(\bullet, \mathrm{AGG}(\{\{\bullet, \bullet, \bullet\}\}))$

Often, AGG is sum, mean or max, and COM is realized by an FNN.

Self-attention

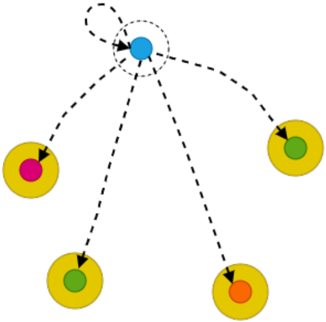


Figure: The blue vertex observes *all* vertices in the graph.

Vertices update their feature vector by aggregating the feature vectors of all vertices in the graph.

A common method of aggregation applied to graphs is self-attention:

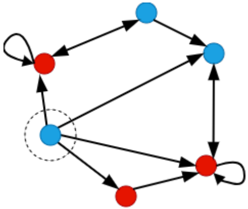
$$\mathrm{softmax}\left(\frac{(XW_Q)(XW_K)^\top}{\sqrt{d}}\right)XW_V,$$

where W_Q, W_K and W_V are real-valued $d \times d$-matrices and X is the feature matrix of the graph.

Logics

Graded modal logic with counting global modality (or $\mathrm{GML} + \mathrm{GC}$): Propositional logic + diamonds $\Diamond_{\geq k}$ and $\langle G \rangle_{\geq k}$.

- A vertex satisfies $\Diamond_{\geq k}\varphi$ iff at least k out-neighbours satisfy φ.
- A vertex satisfies $\langle G \rangle_{\geq k}\varphi$ iff at least k vertices in the graph satisfy φ.



$\mathrm{GML} + \mathrm{G}$: Propositional logic + diamonds $\Diamond_{\geq k}$ and $\langle G \rangle_{\geq 1}$.

$\mathrm{PL} + \mathrm{GC}$: Propositional logic + $\langle G \rangle_{\geq k}$.

$\mathrm{PL} + \mathrm{G}$: Propositional logic + $\langle G \rangle_{\geq 1}$.

Real-based characterizations

Theorem 1

Relative to FO, the following pairs have the same expressive power:

$$\mathrm{GML} + \mathrm{G} \equiv \textit{GPS-networks}$$

$$\mathrm{PL} + \mathrm{G} \equiv \textit{Graph Transformers}$$

Proof.

"$\Rightarrow$" By induction on the structure of a formula.

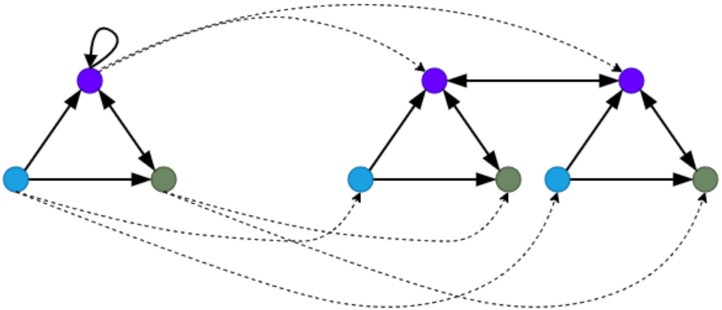
"$\Leftarrow$" We introduce **global-ratio graded bisimilarity**, $\sim_{G\%}$, which considers the ratios of graded bisimilarity types within a graph.

We show that GPS-networks are invariant under $\sim_{G\%}$.

Then we prove a corresponding **van Benthem/Rosen theorem**: Every FO-formula invariant under $\sim_{G\%}$ is equivalent to a formula of $\mathrm{GML} + \mathrm{G}$.

Analogous results are obtained for GTs and $\mathrm{PL} + \mathrm{G}$. $\square$

Global-ratio graded bisimilarity



An illustration of two global-ratio graded bisimilar graphs, where bisimilar vertices are connected with dotted lines.

Float-based characterizations

Float-based sums are bounded, i.e., there exists some k such that it makes no difference whether a float appears k or ℓ times in the sum for all $\ell > k$.

This is because:

- Float sum is not associative due to rounding errors, so the order of the sum has to be fixed.
- Summing in a random vertex order violates isomorphism invariance.
- Instead, summing in increasing order of floats is reasonable for numerical accuracy, but leads to boundedness due to rounding.

Thus, floating-point aggregation functions are also bounded.

Float-based characterizations

Theorem 2

The following pairs have the same expressive power:

$$\mathrm{GML} + \mathrm{GC} \equiv \textit{GPS-networks with floats}$$

$$\mathrm{PL} + \mathrm{GC} \equiv \textit{Graph Transformers with floats}$$

Proof.

"$\Rightarrow$" By structural induction, self-attention can simulate diamonds $\langle G \rangle_{\geq k}$ due to the underflow effect in float arithmetic (values near 0 round to 0).

"$\Leftarrow$" PL can handle all local steps (e.g., FNNs), since it is Boolean complete.

Message passing modules can be simulated by using diamonds $\Diamond_{\geq k}$ since aggregation functions are bounded.

For self-attention, we carefully simulate each matrix operation step-by-step; diamonds $\langle G \rangle_{\geq k}$ suffice due to the sum operations being bounded. □

Future work

- Study other attention mechanisms (our results already generalize to average-hard attention).
- Characterize common positional encodings (e.g. graph Laplacian).
- Generalize all of our results for graph classifications (our float results already generalize for graph classification tasks).

Thank you!