



We Need Metrics for the Localisation and Factorisation of Learning Tasks on Graphs

Johannes Lutzeyer

Data Science and Mining Team, Laboratoire d'Informatique (LIX),
École Polytechnique, Institut Polytechnique de Paris

April 30, 2025

Overview of Today's Talk

- 1) Brief Introduction to the Discussed Methodology;
- 2) Contexts in Which Better Metrics Would Be Beneficial:
 - Oversmoothing,
 - Oversquashing,
 - Long Range Interactions,
 - Interaction Effects in Neighbourhoods and Beyond,
 - Global Representations.

Graph Neural Networks with Virtual Nodes

We consider graph-structured data to be the combination of

- a graph $G = (V, E)$;
- node-features $X = [x_1, \dots, x_n]^T$;
- and add a **virtual node** connected to all nodes in V (Pham, et al., 2017; Gilmer et al., 2017).

The model equations for MPNN+VN,

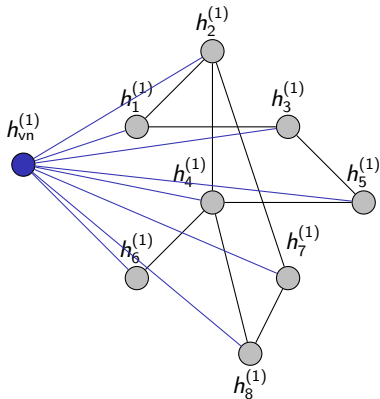
$$h_{\text{vn}}^{(\ell+1)} = U_{\text{vn}}^{(\ell)}(h_{\text{vn}}^{(\ell)}, M_{\text{vn}}^{(\ell)}(\{h_j^{(\ell)} : j \in V\})),$$

$$h_i^{(\ell+1)} = U_i^{(\ell)}(h_i^{(\ell)}, M_i^{(\ell)}(\{h_j^{(\ell)} : j \in \mathcal{N}(i)\}), h_{\text{vn}}^{(\ell)}).$$

E.g., one layer of the GCN + VN,

$$h_{\text{vn}}^{(\ell+1)} = \sigma \left(\Omega_{\text{vn}}^{(\ell)} h_{\text{vn}}^{(\ell)} + \frac{1}{n} \sum_{j=1}^n W_{\text{vn}}^{(\ell)} h_j^{(\ell)} \right), \quad h_i^{(\ell+1)} = \sigma \left(\Omega_i^{(\ell)} h_i^{(\ell)} + \sum_{j \in \mathcal{N}_i} \frac{1}{\sqrt{d_i d_j}} W_i^{(\ell)} h_j^{(\ell)} + h_{\text{vn}}^{(\ell)} \right),$$

where $\Omega_{\text{vn}}^{(\ell)}$ and $W_{\text{vn}}^{(\ell)}$ denote trainable weight matrices. where d_i denotes the node degree of node i and $\Omega_{\text{vn}}^{(\ell)}, W_{\text{vn}}^{(\ell)}, \Omega_i^{(\ell)}, W_i^{(\ell)}$ denote trainable weight matrices.



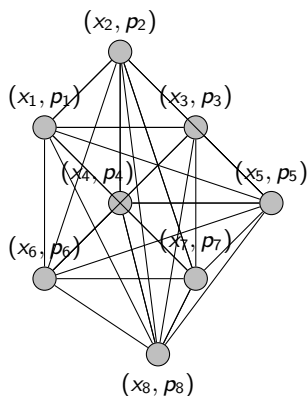
Graph Transformers (GTs)

We consider graph-structured data to be the combination of

- a graph $G = (V, E)$;
- node-features $X = [x_1, \dots, x_n]^T$;
- we construct positional encoding p_i for node $i \in V$ (random walk or spectral) and concatenate them with x_i .

We use the dense 'query, key, value self-attention scheme' to message pass the concatenation of node features and positional encodings (Kreuzer et al., 2021; Ying et al., 2021).

The GraphGPS architecture (Rampášek et al., 2022) runs both a GT and MPNN layer in parallel and then sums node representations.



Complexity Comparison

MPNN: $\mathcal{O}(|E|)$,

MPNN+VN: $\mathcal{O}(|E| + |V|)$,

GT: $\mathcal{O}(|V|^2)$.

Position: Graph Learning Will Lose Relevance Due To Poor Benchmarks (Bechler-Speicher, et al., 2025)

Issues:

- “Missing transformative **real-world applications** and supporting benchmarks”
- “Graphs are not necessarily **constructed** in a meaningful way”
- “Bad benchmarking culture”

“This position paper calls for a paradigm shift toward more meaningful benchmarks, rigorous evaluation protocols, and stronger collaboration with domain experts to drive impactful and reliable advances in graph learning research, unlocking the potential of graph learning.”

⇒ These are good points. In addition to better datasets, **we need better metrics for the associated learning tasks.**

Metrics for Oversmoothing

Oversmoothing: Refers to the phenomenon of node representations becoming indistinguishable in deep GNNs (Alon and Yahav, 2020; Keriven, 2022; Southern et al., 2025).

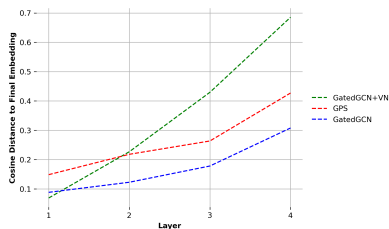
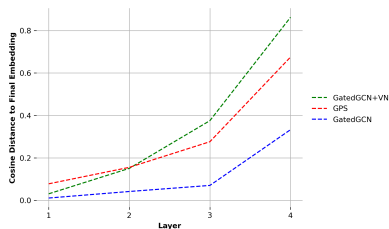
However,

- we can still fit arbitrarily deep update steps. Oversmoothing only arises as a result of arbitrarily many message passing steps, i.e., **arbitrarily large receptive fields** over the graph. Do we really want to solve problems of this kind?
- I suspect the performance decrease on real-world data to arise not from oversmoothing, but as a consequence of **uninformative receptive** fields.
- for graph-level tasks oversmoothing may be **beneficial**.

We need better metrics for the label-relevance of different size receptive fields.

Oversmoothing for Graph-Level Tasks (Southern et al., 2025)

In fact, we observe MPNN+VN to smooth node features more drastically than the baselines on the Peptides-func and Peptides-struct datasets:



Oversmoothing: Refers to the phenomenon of node representations becoming indistinguishable in deep GNNs. We further specify,

- for node-level tasks indistinguishable node representations are clearly **harmful**.
- However, for graph-level tasks oversmoothing may be **beneficial** if the common node representation aligns with the label distribution.

Metrics for Oversquashing

Oversquashing can also be defined as the impossibility of losslessly compressing a receptive field that grows with the depth of the network in fixed size node representations.

Relatedly, **oversquashing** sometimes refers to the difficulty of exchanging information across “bottlenecks” in the graph.

However,

- How many real-world learning tasks require the **joint observation** of entire receptive fields?
- How often do we want to exchange information **between structural communities**?

It would be beneficial to be able to quantify

- 1) **factorisation**: whether this relevant information truly needs to be jointly observed or the label distribution factorises and subsets of the data can be processed independently.
- 2) **localisation**: where in the k -hop neighbourhood of a node the information relevant to our learning task is located.

Metrics for Long Range Interactions

An often only loosely defined concept is that of **long range interactions** in a graph, where features at a large shortest path distance from a given central node v need to be considered to perform a certain learning task on v .

Sometimes related to: **Oversquashing** prevents the underlying model from exchanging information between nodes at large **commute time** τ , where $\tau(i, j)$ measures the expected number of steps for a random walk to commute between i and j .

The hypothesised presence of such interactions is often used to motivate the use of graph rewiring techniques, virtual nodes or graph transformers.

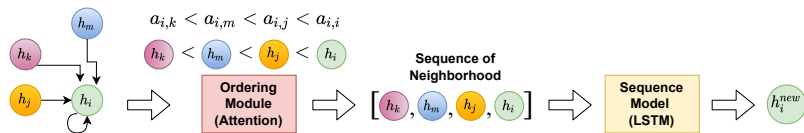
However,

- Do we always want the ability to exchange information between distant nodes?
- In what contexts do long range interactions arise?

There is recent progress on this issue (Zhou et al., 2025) and also Jacob Bamberger has upcoming work on this topic.

Metrics for Interaction Effects in Neighbourhoods and Beyond

We proposed the GOAT architecture (Chatzianastasis et al., 2023) to capture interaction effects in neighbourhoods.



- 1) A self-attention mechanism is used to obtain a ranking between the nodes of the neighbourhood.
- 2) Then, the ordered neighbourhood is given as input into a sequence model (LSTM) to produce the updated representation of node v_i .

However,

- When are such effects present and how could we quantify them?
- How else may we want to capture such interaction effects?

Metrics for Global Representations

Observation: First layer attention maps of the self-attention matrix in the GraphGPS framework for different datasets (Southern et al., 2025).

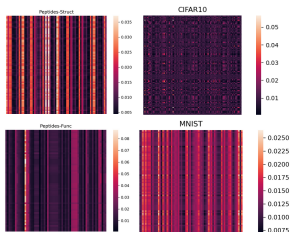


Table: Comparing GraphGPS (a GT) to an MPNN+VN. Arrows indicate if the performance improves with higher (\uparrow) or lower (\downarrow) scores. We also report the mean standard deviation within each column in the first attention layer.

Method	Pept-Func (\uparrow)	Pept-Struct (\downarrow)	MNIST (\uparrow)	CIFAR10 (\uparrow)
GraphGPS	0.6534 \pm .0091	0.2509 \pm .0014	98.051 \pm .126	72.298 \pm .356
GatedGCN+PE+VN	0.6712 \pm .0066	0.2481 \pm .0015	98.122 \pm .102	70.280 \pm .380
std attention layer	0.0011	0.0007	0.0006	0.0038

Here it would be nice to be able to **quantify** when we can benefit from more than one global representation to perform our learning task.

Conclusions

We conclude,

- We should find more meaningful benchmarks for graph learning and when doing so aim to quantify properties of associated learning tasks.
- I think we have an insufficient understanding of what makes graph learning tasks complex. A principled, quantified categorisation of graph learning tasks would be of value.

I recently helped Sohir Maskey with a submission, in which we generalise the Tree Movers Distance to obtain generalisation bounds on expressive GNNs. Here a bespoke metric helps us get a better understanding of the problem.

Thank you for your attention!

 @jlutzeyer.bsky.social

References

- U. Alon & E. Yahav, "On the Bottleneck of Graph Neural Networks and its Practical Implications," In: *International Conference on Learning Representations (ICLR)*, 2020.
- M. Bechler-Speicher, B. Finkelshtein, F. Frasca, L. Müller, J. Tönshoff, A. Siraudin, V. Zaverkin, M. M. Bronstein, M. Niepert, B. Perozzi, M. Galkin & C. Morris, "Position: Graph Learning Will Lose Relevance Due To Poor Benchmarks," *arXiv:2502.14546*, 2025.
- M. Chatzianastasis, J. F. Lutzeyer, G. Dasoulas & M. Vazirgiannis, "Graph Ordering Attention Networks," *Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2023.
- J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals & G. E. Dahl, "Neural message passing for Quantum chemistry," *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.
- N. Keriven, "Not too little, not too much: a theoretical analysis of graph (over) smoothing," *Advances in Neural Information Processing Systems (NeurIPS)*, pp.2268-2281, 2022.
- Thomas N. Kipf & M. Welling, "Semi-supervised classification with graph convolutional networks," *International Conference on Learning Representations (ICLR)*, 2017.
- D. Kreuzer, D. Beaini, W. Hamilton, V. Létourneau & P. Tossou, "Rethinking graph transformers with spectral attention," *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

- G. Mateos, S. Segarra, A. G. Marques & A. Ribeiro, "Connecting the dots," *IEEE Signal Processing Magazine*, vol. 36, pp. 16–43, 2019.
- T. Pham, T. Tran, H. Dam & Svetha Venkatesh, "Graph classification via deep learning with virtual nodes," *arXiv:1708.04357*, 2017.
- L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf & D. Beaini, "Recipe for a general, powerful, scalable graph transformer," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- J. Southern, F. Di Giovanni, M. Bronstein & J. F. Lutzeyer, "Understanding Virtual Nodes: Oversmoothing, Oversquashing, and Node Heterogeneity," *International Conference on Learning Representations (ICLR)*, 025.
- C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen & T.-Y. Liu, "Do transformers really perform badly for graph representation?," *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- D. Zhou, E. Kharlamov & E. V. Kostylev, "GLoRa: A Benchmark to Evaluate the Ability to Learn Long-Range Dependencies in Graphs," *International Conference of Learning Representations (ICLR)*, 2025.