

NeurIPS 2025 · The 39th Conference on Neural Information Processing Systems · San Diego Convention Center

# *Over-squashing* in Spatiotemporal Graph Neural Networks

---

Ivan **Marisca**, Jacob Bamberger, Cesare Alippi, Michael M. Bronstein

University of Oxford

IDSIA, USI (Lugano, Switzerland)



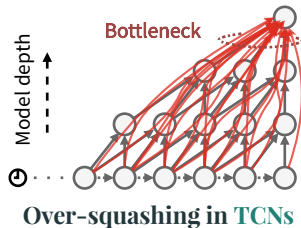
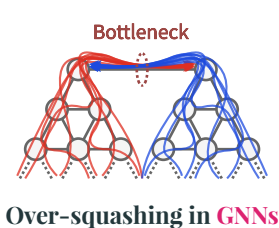
idsia



# Over-squashing

GNNs suffer from *over-squashing*: information is **compressed and lost** through **bottlenecks** [1].

STGNNs introduce a **temporal** dimension, **compounding** the issue [2].



How does *spatiotemporal over-squashing* affect representation learning in STGNNs?

[1] Alon et al., “On the bottleneck of graph neural networks and its practical implications”, ICLR 2021.

[2] Bengio et al., “Learning long-term dependencies with gradient descent is difficult”, IEEE TNN 1994.

# Contributions

---

Formal characterization of **spatiotemporal over-squashing**.

⚠️ The temporal dimension is an **additional axis** for information propagation...

😞 ...which **amplifies the compression** effects observed in static GNNs.

Proof that convolutional STGNNs are **more sensitive to information far apart in time**.

😞 Counterintuitive behavior and **opposite** to graph over-squashing.

😊 We outline architectural **modifications that mitigate** this imbalance when required.

Proof that spatiotemporal over-squashing affects T&S and TTS models **to the same degree**.

😊 Theoretical support for **scalable** TTS designs.

# Sensitivity analysis

**Spatiotemporal** over-squashing can be assessed via the **spectral norm** of the **STGNN's Jacobian** :

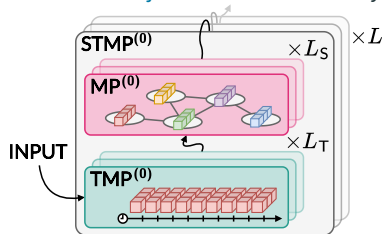
$$\left\| \nabla_i^u \mathbf{h}_t^{v(L)} \right\| = \left\| \frac{\partial \mathbf{h}_t^{v(L)}}{\partial \mathbf{h}_{t-i}^{u(0)}} \right\|.$$

💡 How much **features** of  **$u$**  influence the **output** of  **$v$** .

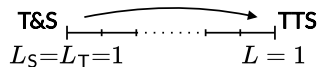
The Jacobian of the  $l$ -th **disjoint** layer **factorizes** as:

$$\frac{\partial \mathbf{h}_{t-j}^{v(l+1)}}{\partial \mathbf{h}_{t-i}^{u(l)}} = \underbrace{\frac{\partial \mathbf{h}_{t-j}^{v(l+1)}}{\partial \mathbf{z}_{t-j}^{u(l)}}}_{\text{space}} \underbrace{\frac{\partial \mathbf{z}_{t-j}^{u(l)}}{\partial \mathbf{h}_{t-i}^{u(l)}}}_{\text{time}}.$$

We consider  $L$  **disjoint STMP** stacked layers:



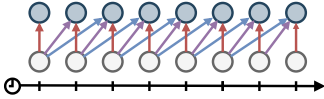
$LL_S$  and  $LL_T$  are **computational budgets**.



[3] Topping et al., “Understanding over-squashing and bottlenecks on graphs via curvature”, ICLR 2022.

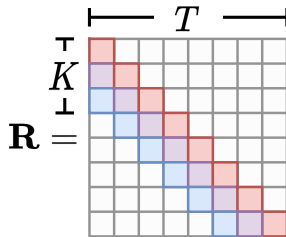
# Message-passing temporal convolutional networks

We focus on **MPTCNs**, where TMP is implemented using **causal convolutions** with  $K$ -sized filter  $\tau$ :

$$h_{t-i}^{(l+1)} = \sum_{k=0}^{K-1} \tau[k] \cdot h_{t-i-k}^{(l)}$$


This can be written as a matrix multiplication using **Toeplitz matrix**  $\mathbf{R}$ :

$$h_{t-T:t}^{(l+1)} = \mathbf{R}^\top h_{t-T:t}^{(l)}$$



$\mathbf{R}$  acts as a **temporal topology matrix**:

- $(\mathbf{R}^l)_{ij}$  is the **number of paths** from  $t-i$  to  $t-j$  after  $l$  layers.

# Temporal sensitivity upper bound

We aim at studying  $\left\| \nabla_i^u \mathbf{h}_t^{v(L)} \right\|$  in MPTCNs, starting from the **temporal component**:  $\left\| \nabla_i^v \mathbf{h}_t^{v(L_T)} \right\|$

## Theorem (4.1 — Over-squashing in TCNs)

Consider a TCN with  $L_T$  successive layers, all with kernel size  $K$ , and assume that  $\left\| \mathbf{W}_k^{(l)} \right\| \leq w$  for all  $k < K$  and  $l \leq L_T$ , and that  $|\sigma'| \leq c_\sigma$ . For each nonnegative  $i < T$ , we have:

$$\left\| \nabla_i^v \mathbf{h}_t^{v(L_T)} \right\| \leq \underbrace{(c_\sigma w)^{L_T}}_{\text{model}} \underbrace{(\mathbf{R}^{L_T})_{i0}}_{\text{temporal topology}}.$$

😊 **Model** and **temporal topology** have **distinct and multiplicative effects** on sensitivity.

🔗 **R** is a **lower-triangular, Toeplitz, band** matrix ( $\mathbf{R}_{ii} = 0$  for  $i \geq K$ ), so asymptotically we have...

# Amplification of long-range temporal dependencies

## *Proposition (4.2 — Sink effect in causal TCNs)*

Let  $\mathbf{R} \in \mathbb{R}^{T \times T}$  be a lower-triangular Toeplitz matrix with lower bandwidth  $K \geq 2$ . For any  $i > j$ :

$$\left| \frac{(\mathbf{R}^l)_{j0}}{(\mathbf{R}^l)_{i0}} \right| \rightarrow 0 \text{ as } l \rightarrow \infty.$$

The final token receives considerably more influence from tokens positioned **earlier** in the sequence.

Counterintuitively, TCNs **overemphasize earlier information** when depth increases.

**Sink effect:** recent inputs **vanish** with depth, with all focus towards the **first token**.

# Spatiotemporal sensitivity bound

Theoretical upper bound for common GNNs decomposed into **model vs topology** too (from [4]):

$$\left\| \nabla_0^u \mathbf{h}_t^{v(L)} \right\| \leq \underbrace{(c_\xi \theta_m)^L}_{\text{model}} \underbrace{(\mathbf{S}^L)_{uv}}_{\text{topology}}, \quad \leftarrow \mathbf{S} \text{ is a function of } \mathbf{A}$$

We combine both results to analyze the **spatiotemporal case**.

---

[4] Di Giovanni et al., “How does over-squashing affect the power of GNNs?”, TMLR 2024.



# Spatiotemporal sensitivity bound

## Theorem (5.1 — Over-squashing in MPTCNs)

Consider an MPTCN with  $L$  STMP layers, each consisting of  $L_T$  temporal (TMP) and  $L_S$  spatial (MP) layers. Then, for any  $i, j \in \mathcal{V}$  and  $p, q \in [0, T)$ , the following holds:

$$\left\| \nabla_i^u \mathbf{h}_t^{v(L)} \right\| \leq \underbrace{(c_\xi \theta_m)^{LL_S} (c_\sigma \mathbf{w})^{LL_T}}_{\text{model}} \underbrace{(\mathbf{S}^{LL_S})_{uv} (\mathbf{R}^{LL_T})_{i0}}_{\text{spatiotemporal topology}}.$$

Again, the bound decomposes into **model vs topology** and **time vs space**. Two observations:

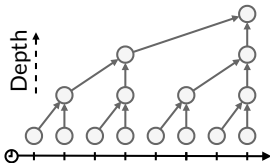
😊 TTS ( $L = 1$ ) and T&S ( $L > 1$ ) architectures with fixed computational budget **share the same bound**.

⚠️ Improving **only one** dimension is **insufficient** if the other is **bottlenecked**.

# Rewiring the temporal graph

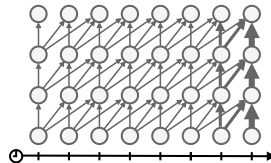
Two approaches to mitigate temporal over-squashing by *rewiring* the temporal graph.

## Dilated convolutions



- 😊 Exponentially expanding the receptive field **reduces the paths** from earlier tokens.
- 😞 Sink behavior after dilation resets.

## Row-normalized convolutions



- 😊 Slowly converges to **uniform attention** distribution over the sequence.
- 😞 Has only effect on **last token**.

# Assessment on real-world benchmarks

Empirical evaluation on traffic and weather data.

- 😊 T&S and TTS approaches perform **comparably** on average.
- 😊 Row-normalized convolutions **improve accuracy**.
- 😊 Findings remain valid even with more **sophisticated architectures**. (see Graph WaveNet [5])

**Table 1:** Forecasting error (MAE  $\pm$  std) with fixed computational budget.

Models		$L$	METR-LA	PEMS-BAY	EngRAD
MPTCN	<b>R</b>	6	$3.19_{\pm 0.02}$	$1.66_{\pm 0.00}$	$44.43_{\pm 0.41}$
		3	$3.19_{\pm 0.01}$	$1.65_{\pm 0.01}$	<b><math>43.83_{\pm 0.03}</math></b>
		1	<b><math>3.14_{\pm 0.02}</math></b>	<b><math>1.63_{\pm 0.01}</math></b>	$44.47_{\pm 0.42}$
	<b>R<sub>N</sub></b>	6	$3.17_{\pm 0.02}$	$1.65_{\pm 0.01}$	$41.82_{\pm 0.38}$
		3	$3.17_{\pm 0.01}$	$1.65_{\pm 0.00}$	$41.78_{\pm 0.09}$
		1	$3.16_{\pm 0.01}$	$1.65_{\pm 0.01}$	<b><math>40.38_{\pm 0.08}</math></b>
Graph WaveNet (orig.)			$3.02_{\pm 0.02}$	<b><math>1.55_{\pm 0.01}</math></b>	<b><math>40.50_{\pm 0.27}</math></b>
Graph WaveNet (TTS)			<b><math>3.00_{\pm 0.01}</math></b>	$1.57_{\pm 0.00}$	$40.64_{\pm 0.29}$



# THE END

Questions? Write to [ivan.marisca@usi.ch](mailto:ivan.marisca@usi.ch)

Live poster presentation:  
San Diego Convention Center  
Friday 5 December 2025  
4:30 p.m. – 7:30 p.m. PST

# References

---

- [1] Alon and Yahav. “**On the bottleneck of graph neural networks and its practical implications**”. In: *International Conference on Learning Representations*. 2021.
- [2] Bengio, Simard, and Frasconi. “**Learning long-term dependencies with gradient descent is difficult**”. In: *IEEE Transactions on Neural Networks* 5.2 (1994).
- [3] Topping, Di Giovanni, Chamberlain, Dong, and Bronstein. “**Understanding over-squashing and bottlenecks on graphs via curvature**”. In: *International Conference on Learning Representations*. 2022.
- [4] Di Giovanni, Rusch, Bronstein, Deac, Lackenby, Mishra, and Veličković. “**How does over-squashing affect the power of GNNs?**” In: *Transactions on Machine Learning Research* (2024). ISSN: 2835-8856.
- [5] Wu, Pan, Long, Jiang, and Zhang. “**Graph WaveNet for Deep Spatial-Temporal Graph Modeling**”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019 [[🔗 URL](#)].