Understanding and Tackling Over-Dilution in Graph Neural Networks

Junhyun Lee[1] Veronika Thost[2] Bumsoo Kim[3] Jaewoo Kang[1†] Tengfei Ma[4†]

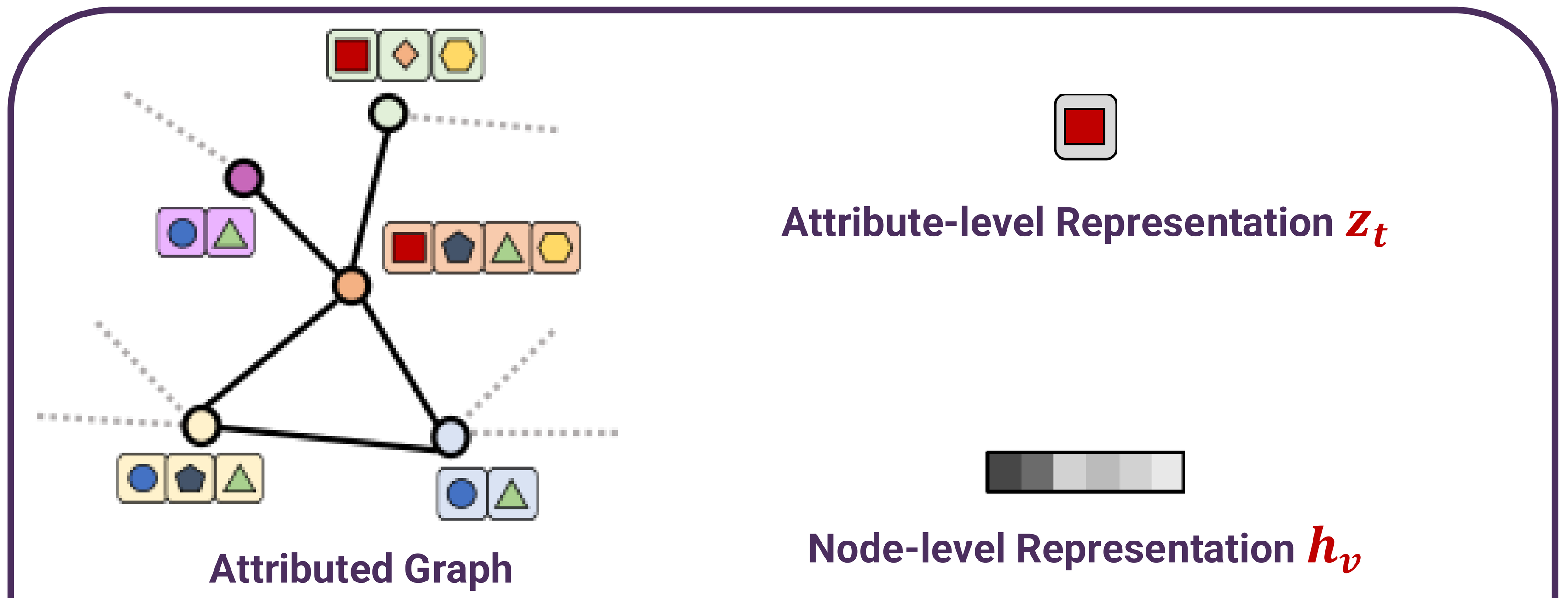
1 Korea University 2 IBM Research 3 Chung-Ang University 4 Stony Brook University

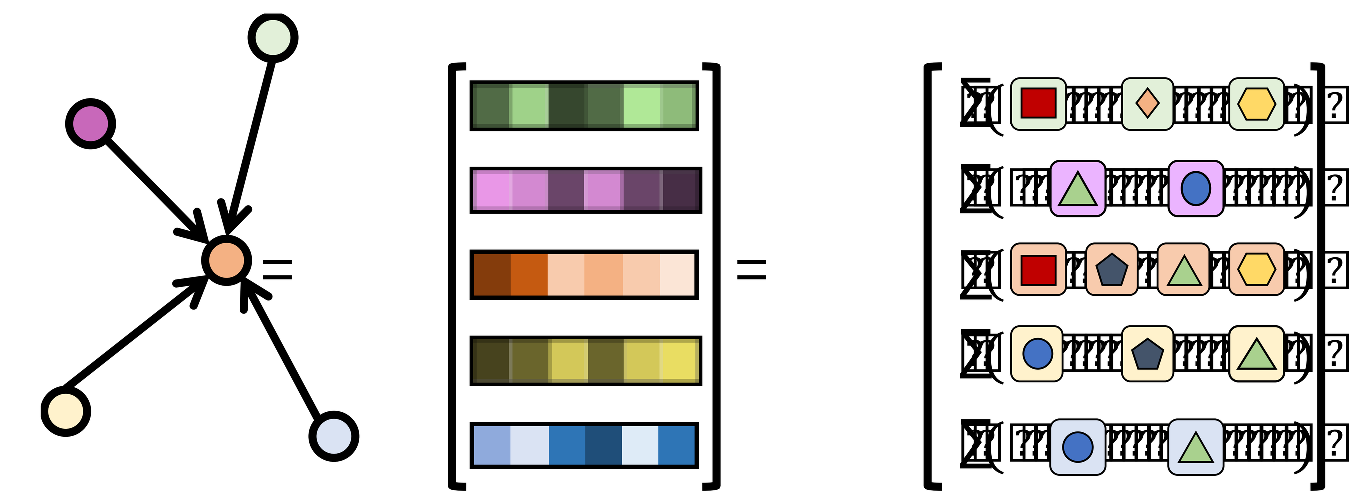
† Corresponding Authors

KDD2025 August 3–7, 2025

Abstract

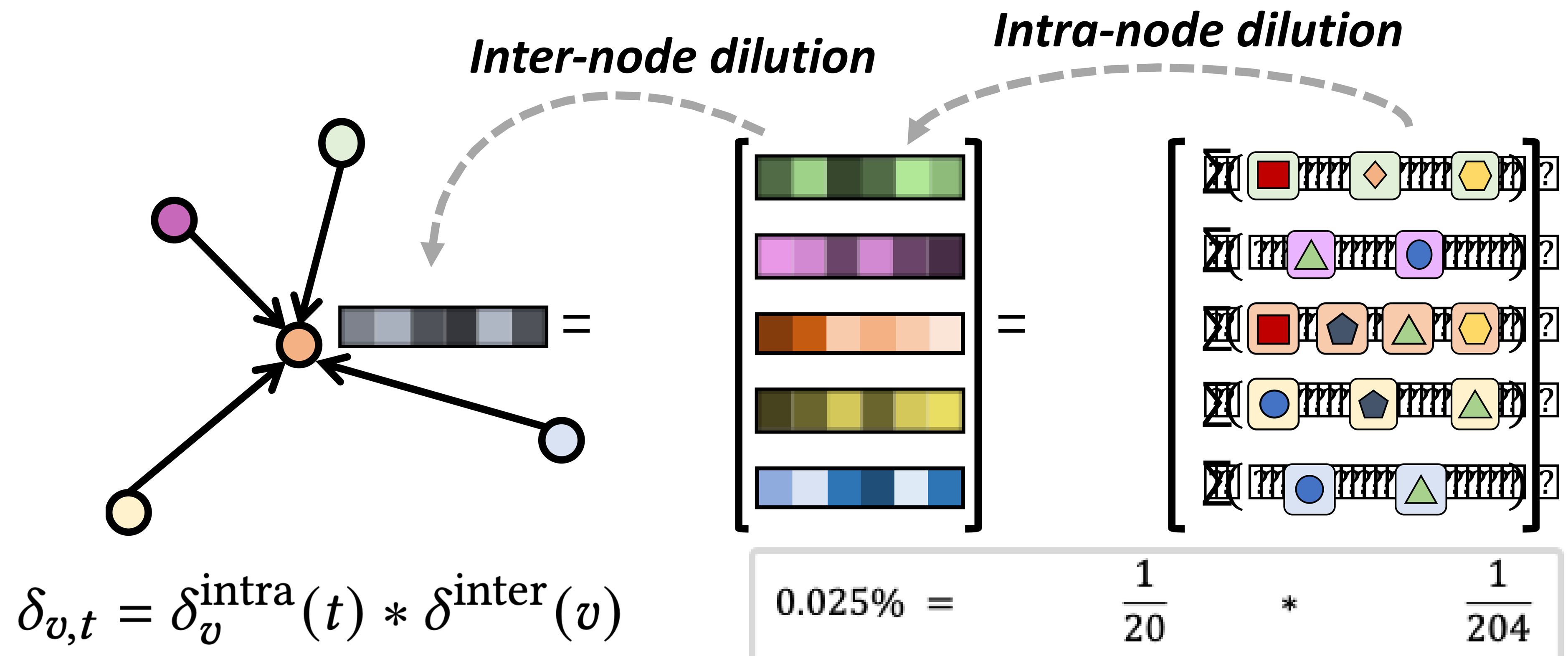
Message Passing Neural Networks (MPNNs) hold a key position in machine learning on graphs, but they struggle with unintended behaviors, such as over-smoothing and over-squashing, due to irregular data structures. The observation and formulation of these limitations have become foundational in constructing more informative graph representations. In this paper, we delve into the limitations of MPNNs, focusing on aspects that have previously been overlooked. Our observations reveal that even within a single layer, the information specific to an individual node can become significantly diluted. To delve into this phenomenon in depth, we present the concept of **Over-dilution** and formulate it with two dilution factors: **intra-node dilution** for attribute-level and **inter-node dilution** for node-level representations. We also introduce a transformer-based solution that alleviates over-dilution and complements existing node embedding methods like MPNNs. Our findings provide new insights and contribute to the development of informative representations.



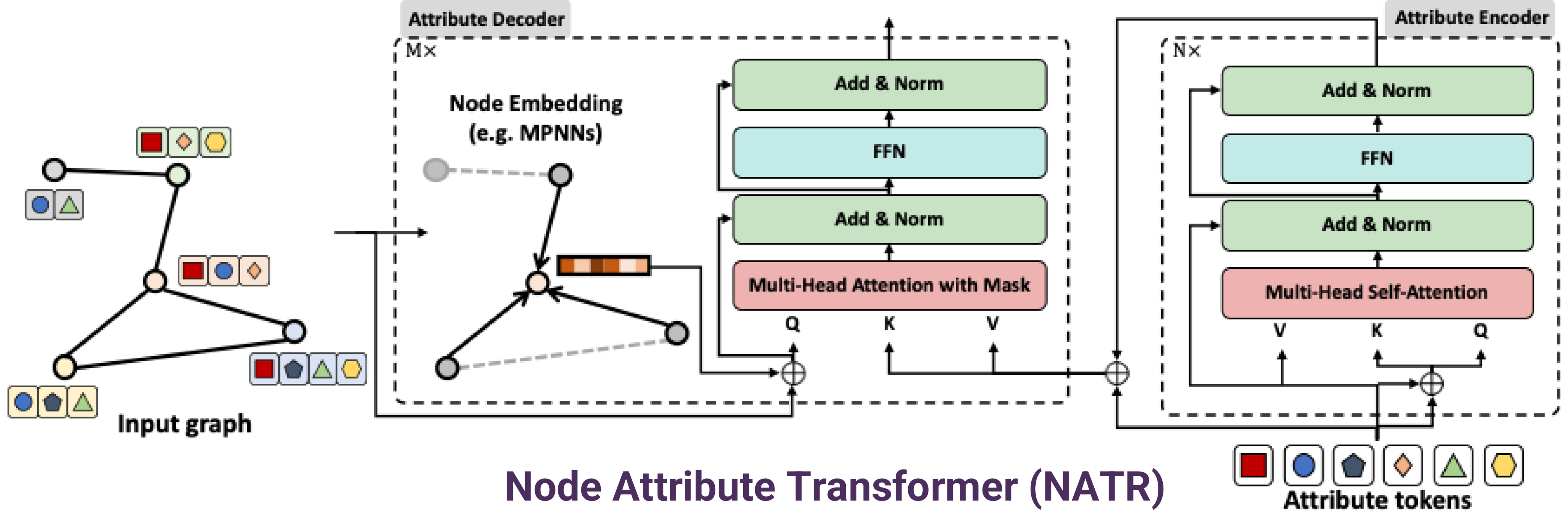


Message Passing Neural Networks

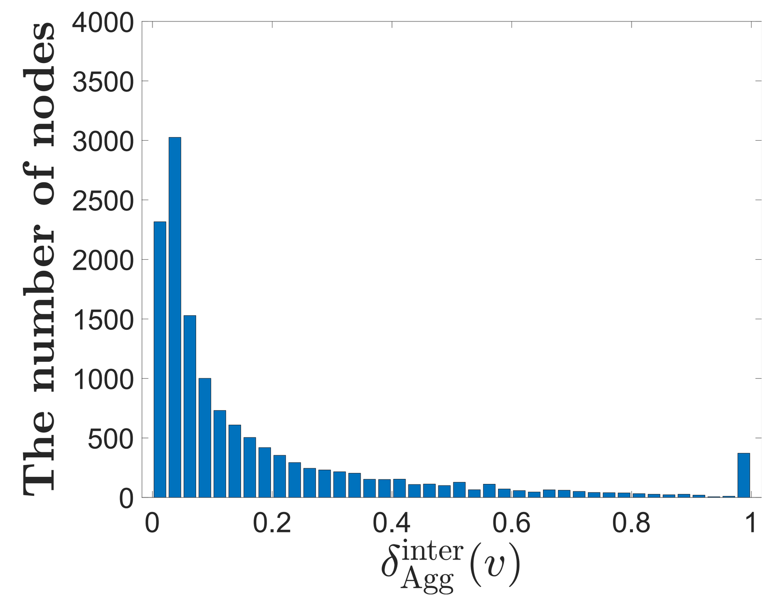
	Median Degree	$\lvert\mathcal{T}\rvert$	Avg. $\lvert\mathcal{T}_v\rvert$	Median $\lvert\mathcal{T}_v\rvert$
Computers	19	767	267.2	204
Photo	18	745	258.8	193
Cora ML	3	2879	50.5	49
OGB-DDI$_{\text{SUBSET}}$	500	1024	58.2	56
OGB-DDI$_{\text{FULL}}$	446	1024+1	49.1	51



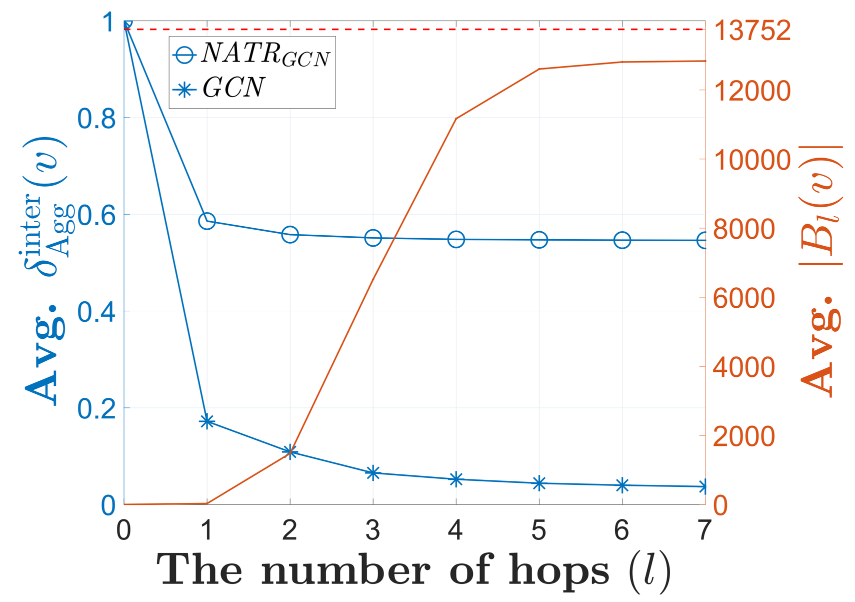
The dilution factor of attribute t at node v



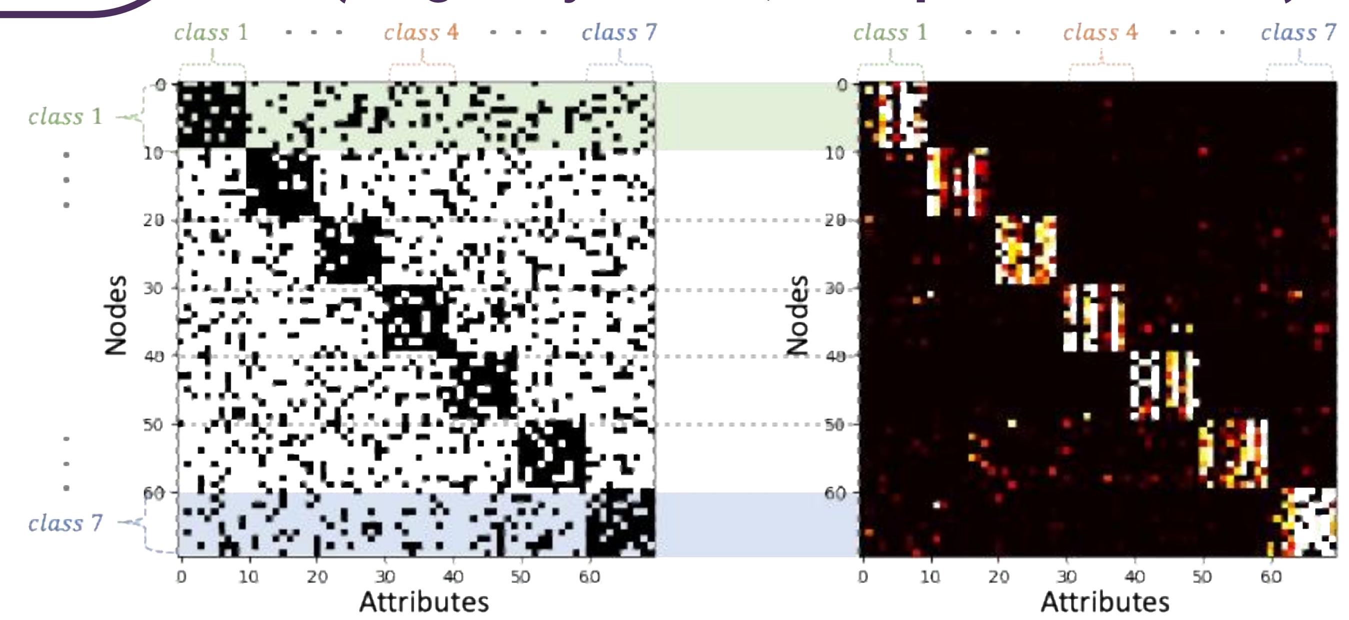
Node Attribute Transformer (NATR)



The histogram of inter-dilution factors (single layer *GCN*, Computers Dataset)



The Avg. of inter-dilution factors
The Avg. size of the receptive field



Intra-node dilution factors (synthetic)

	2 Layers	3 Layers	4 Layers	5 Layers
GCN	**31.01**	30.84	28.97	26.99
GCN$_{JK}$	**29.47**	27.85	28.00	27.49
NATR$_{GCN}$	39.81	41.54	40.96	**42.38**
GAT	**24.73**	21.07	11.52	4.15
GAT$_{JK}$	**27.22**	24.54	23.90	23.98
NATR$_{GAT}$	39.51	39.58	**40.63**	40.21
SGC	**30.37**	25.78	24.30	23.87
NATR$_{SGC}$	**36.99**	36.47	35.31	34.01

Link Prediction Task (Hits@20)

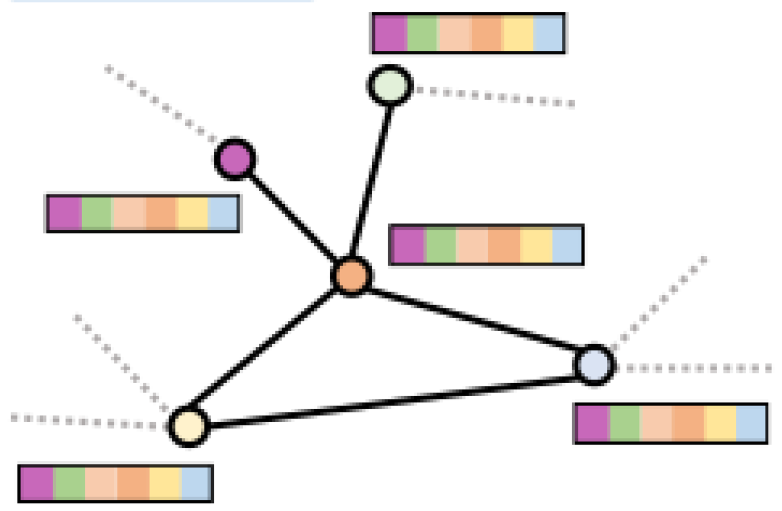
Definition 3.1. (Intra-node dilution factor). *For a graph $\mathcal{G} = (\mathcal{T}, \mathcal{V}, \mathcal{E})$, let z_t be the representation of attribute $t \in \mathcal{T}$ and $h_v^{(0)}$ denote the initial feature representation of node $v \in \mathcal{V}$, which is calculated from the representations of attribute subset $\mathcal{T}_v$ that node v possesses. The influence score $I_v(t)$ attribute t on node v is the sum of the absolute values of the elements in the Jacobian matrix $\left[\frac{\partial h_v^{(0)}}{\partial z_t}\right]$. We define the intra-node dilution factor as the influence distribution by normalizing the influence scores: $\delta_v^{\text{intra}}(t) = I_v(t)/\Sigma_{s\in\mathcal{T}_v} I_v(s)$. In detail, with the all-ones vector e:*

$$\delta_v^{\text{intra}}(t) = e^T \left[\frac{\partial h_v^{(0)}}{\partial z_t}\right] e \Big/ \sum_{s\in\mathcal{T}_v} e^T \left[\frac{\partial h_v^{(0)}}{\partial z_s}\right] e \tag{3}$$

Definition 3.2. (Inter-node dilution factor). *Let $h_v^{(0)}$ be the initial feature and $h_v^{(l)}$ be the learned representation of node $v \in \mathcal{V}$ at the l-th layer. We define the inter-node dilution factor as the normalized influence distribution of node-level representations: $\delta^{\text{inter}}(v) = I_v(v)/\Sigma_{u\in\mathcal{V}} I_v(u)$, or*

$$\delta^{\text{inter}}(v) = e^T \left[\frac{\partial h_v^{(l)}}{\partial h_v^{(0)}}\right] e \Big/ \sum_{u\in\mathcal{V}} e^T \left[\frac{\partial h_v^{(l)}}{\partial h_u^{(0)}}\right] e \tag{4}$$

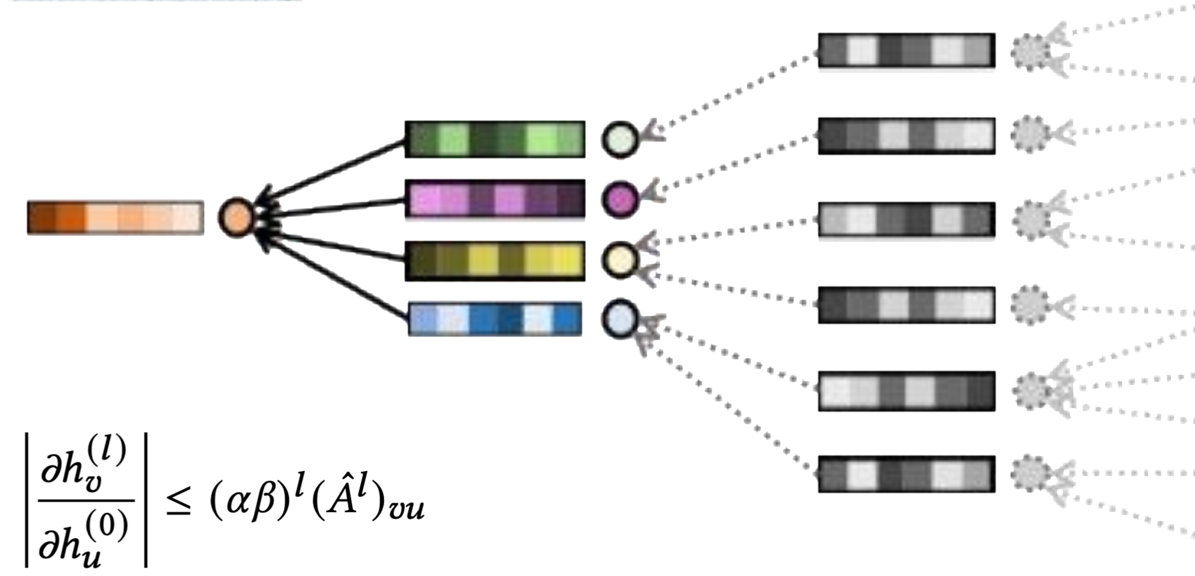
Over-smoothing Inability to distinguish node features



$$MAD(h^{(l)}) = \frac{1}{\mathcal{V}} \sum_{v\in\mathcal{V}} \sum_{u\in N_v} 1 - \frac{h_v^{(l)\top} h_u^{(l)}}{\|h_v^{(l)}\| \|h_u^{(l)}\|}$$

$$\mathcal{E}(h^{(l)}) = \frac{1}{\mathcal{V}} \sum_{v\in\mathcal{V}} \sum_{u\in N_v} \|h_v^{(l)} - h_u^{(l)}\|_2^2$$

Over-squashing Inefficacy in propagating long-distance node features



$$\left|\frac{\partial h_v^{(l)}}{\partial h_u^{(0)}}\right| \le (\alpha\beta)^l (\hat{A}^l)_{vu}$$