A Low-Rank Perspective on Oversmoothing in Graph Neural Networks

Kaicheng Zhang[1], **Piero Deidda**[2, 3], **Desmond Higham**[1] **Francesco Tudisco**[1, 3]

[1] School of Mathematics and Maxwell Institute, University of Edinburgh, Edinburgh, UK
[2] Scuola Normale Superiore, Pisa, Italy
[3] Gran Sasso Science Institute, L'Aquila, Italy
k.zhang-60@sms.ed.ac.uk, piero.deidda@sns.it, d.j.higham@ed.ac.uk, f.tudisco@ed.ac.uk

Abstract

Oversmoothing is a fundamental challenge in graph neural networks (GNNs): as the number of layers increases, the node embeddings become progressively similar, leading to smoothened representations. This phenomenon often results in a sharp performance drop after only a few layers, significantly limiting the depth of GNNs. Traditionally, oversmoothing has been quantified using norm-based energy metrics, such as the Dirichlet energy, which measures the norm of differences between neighbouring node features. These metrics decay only when the embeddings converge to either a rank-one or an all-zero representation as depth increases. However, we argue that these metrics offer an overly simplistic view and fail to reliably capture oversmoothing in realistic scenarios, such as when network weights are unbounded, graph adjacency matrices are not stochastic, or activation functions are highly non-homogeneous (e.g. $\tanh$). In such cases, the embeddings may not collapse to a rank-one representation, and norm-based energy metrics fail to detect a drop in representational quality. Instead, we propose measuring the effective rank of the representations, which provides a more nuanced understanding of oversmoothing. Our findings reveal that a significant drop in effective rank corresponds closely with performance degradation, even in cases where energy metrics remain unchanged. Extensive evaluations across diverse graph architectures demonstrate that rank-based metrics consistently capture oversmoothing, unlike energy-based approaches, which often fail.

Introduction

Graph neural networks (GNNs) have emerged as a powerful framework for learning representations from graph-structured data, with applications spanning knowledge retrieval and reasoning (Peng et al. 2023; Tian et al. 2022), personalised recommendation systems (Damianou et al. 2024; Peng, Sugiyama, and Mine 2022), social network analysis (Fan et al. 2019), and 3D mesh classification (Shi and Rajkumar 2020). Central to most GNN architectures is the message-passing paradigm, where node features are iteratively aggregated from their neighbours and transformed using learned functions, such as multi-layer perceptrons or graph-attention mechanisms.

A significant challenge limiting the effectiveness of GNNs is the problem of oversmoothing. Although definitions of oversmoothing vary across the literature, it is broadly understood as the phenomenon where learned node features become increasingly similar as network depth increases (Li, Han, and Wu 2018). This often leads to a noticeable drop in performance, effectively constraining the maximum depth of message-passing GNNs.

Researchers have proposed various node similarity measures to quantify and understand oversmoothing, often relying on norm-based metrics such as the Dirichlet energy and its variants (Oono and Suzuki 2019; Cai and Wang 2020; Wu et al. 2023; Di Giovanni et al. 2023). These metrics are typically tailored to capture oversmoothing in a narrow range of architectures, particularly those with homogeneous activations like ReLU. As a result, different metrics can yield contradictory conclusions even for the same model. Notably, these metrics only decay when feature embeddings converge to a rank-one or all-zero representation as depth increases. Therefore, while they provide sufficient (but not necessary) conditions for oversmoothing, these metrics offer an overly simplistic view and may fail to capture the phenomenon reliably (Rusch, Bronstein, and Mishra 2023). Despite these limitations, they remain the standard tools for quantifying oversmoothing, even in recent studies (Rusch et al. 2022; Chen et al. 2022; Epping et al. 2024).

Motivated by these observations, we advocate for the use of the effective rank (Roy and Vetterli 2007) as a more robust and generalized metric for quantifying oversmoothing. The effective rank provides an information-theoretic measure of the dimensionality of the feature representation space.

Our results demonstrate that, in scenarios where GNNs are trained with general nonlinear activations that are not homogeneous, adjacency matrices that are not stochastic, or unbounded weights, the effective rank converges to a small value. This convergence indicates a collapse of the features into a low-dimensional space, offering a clear explanation for the poor performance of GNNs under these conditions. In contrast, existing norm-based energy metrics often remain unchanged in such scenarios and, therefore, fail to capture the occurrence of oversmoothing.

Background

Graph Convolutional Network

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with $\mathcal{V}$ denoting its set of vertices and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ its set of edges. Let $A \in \mathbb{R}^{N\times N}$ be the adjacency matrix, where $N = |\mathcal{V}|$ is the total number of nodes of $\mathcal{G}$. To construct a Graph Convolutional Network (GCN) (Kipf and Welling 2016), standard practice is to augment the adjacency matrix with self-loops $\widetilde{A} = A + I$ and then normalize it to obtain

$$P = \widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}, \tag{1}$$

where $\widetilde{D} = D + I$, D is the diagonal degree matrix of the graph $\mathcal{G}$, and I is the identity matrix. The rows of the feature matrix $X \in \mathbb{R}^{N\times d}$ are the concatenation of the d-dimensional feature vectors of all nodes in the graph. At each layer l, the GCN updates the node features as follows

$$X^{(l+1)} = \sigma(PX^{(l)}W^{(l)}) \tag{2}$$

where σ is a nonlinear activation function, $W^{(l)}$ is a trainable weight matrix.

Graph Attention Network

While GCNs use a fixed normalized adjacency matrix to perform graph convolutions at each layer, Graph Attention Networks (GATs) (Veličković et al. 2017; Brody, Alon, and Yahav 2021) perform graph convolution through a layer-dependent message-passing matrix $P^{(l)}$ learned through an attention mechanism as follows

$$P_{ij}^{(l)} = \text{softmax}_j(\sigma_a(a_1^{(l)\top}W^{(l)\top}X_i + a_2^{(l)\top}W^{(l)\top}X_j)) \tag{3}$$

where $a_i^{(l)}$ are learnable parameter vectors, X_i, X_j denote the feature of the ith and jth nodes respectively, the activation σ_a is typically chosen to be LeakyReLU, and softmax_j corresponds to the row-wise normalization

$$\text{softmax}_j(A_{ij}) = \frac{\exp(A_{ij})}{\sum_{j'}\exp(A_{ij'})}. \tag{4}$$

The corresponding feature update is

$$X^{(l+1)} = \sigma(P^{(l)}X^{(l)}W^{(l)})\,. \tag{5}$$

Graph Oversmoothing

Although the exact definition of oversmoothing differs depending on the sources, oversmoothing can be broadly considered as an increase in the similarity between the node features as the inputs are propagated through an increasing number of message passing layers, accompanied by an observable decay in GNN performance. We can easily understand this problem when considering a linear GCN,

$$X^{(l)} = PP\cdots PX^{(0)}W^{(0)}\ldots W^{(l)}. \tag{6}$$

Indeed, for a fully connected graph $\mathcal{G}$ with only one connected component, P has spectral radius equal to 1 with multiplicity 1, so P^l collapses towards the eigenspace spanned by the dominant eigenvector

$$P^l \to xy^\top \quad l \to \infty \tag{7}$$

where $Px = x$, and $P^\top y = y$.

As a consequence, if the features $X^{(l)}$ converge in the limit $l \to \infty$, they degenerate to a matrix having rank at most one, where all the features are aligned with the dominant eigenvector x.

Existing Oversmoothing Metrics

To quantify oversmoothing, different metrics have been introduced, and most of which measure the alignment of the features with the dominant eigenvector of the matrix P. Among these, the most prominent one is the Dirichlet Energy, which aims to measure the overall norm of the difference between neighbouring node features (Cai and Wang 2020)

$$E_{\text{Dir}}(X) = \sum_{i,j\in\mathcal{E}}\left\|\frac{X_i}{\sqrt{1+d_i}} - \frac{X_j}{\sqrt{1+d_j}}\right\|_2^2, \tag{8}$$

where d_i is the degree of node i, X_i is the ith row of the features matrix X. Note that the vector $v = \widetilde{D}^{1/2}\mathbf{1}$, with entries $v_i = \sqrt{1+d_i}$, is the dominant eigenvector of the message-passing matrix in (1) as $Pv = \widetilde{D}^{-1/2}\widetilde{A}\mathbf{1} = \widetilde{D}^{-1/2}(\mathbf{1}+d) = \widetilde{D}^{-1/2}\widetilde{D}\mathbf{1} = v$. It thus immediately follows from our discussion on the linear setting that $E_{\text{Dir}}(X^{(l)})$ converges to zero as $l \to \infty$ for a linear GCN. The intuition suggests that a similar behaviour may occur for "smooth-enough" nonlinearities. In particular, when the nonlinear activation function used in the model is LeakyReLU, the authors of (Cai and Wang 2020) have proved that $E_{\text{Dir}}(X^{(l+1)}) \leq s_l\bar{\lambda}E_{\text{Dir}}(X^{(l)})$, where $s_l = \|W^{(l)}\|_2$ is the largest singular value of the weight matrix $W^{(l)}$, and $\bar{\lambda} = (1 - \min_i \lambda_i)^2$, where $\lambda_i \in (0, 2]$ varies among the nonzero eigenvalues of the normalized graph Laplacian $\widetilde{\Delta} = I - P = I - \widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}$. However, from our experience, the condition to obtain exponential decaying energy, $(\bar{\lambda}\max_l s_l) < 1$, is rarely satisfied in practice, yet oversmoothing still occurs.

There are several variants of the Dirichlet energy for quantifying oversmoothing, for example (Rusch, Bronstein, and Mishra 2023)

$$E'_{\text{Dir}}(X) = \frac{1}{N}\sum_{i,j\in\mathcal{E}}\|X_i - X_j\|_2^2\,. \tag{9}$$

This formulation is applicable to the cases with a graph convolution matrix $P' = \widetilde{D}^{-1/2}P\widetilde{D}^{1/2} = \widetilde{D}^{-1}\widetilde{A}$. In that case, P' is row-stochastic, and the dominant eigenvector becomes the constant vector.

Being absolute measures, these notions of Dirichlet energy may fail to measure the norm of the difference between neighbouring node features. This can occur because the Dirichlet energy may decay to zero even when node features remain far apart, but their norm decays to zero. To address this limitation, normalised versions of the Dirichlet energy have been proposed. For example, (Di Giovanni et al. 2023) introduces a normalized Dirichlet energy, $E_{\text{Dir}}(X)/\|X\|_F^2$, where $\|\cdot\|_F$ denotes the Frobenius norm.

Architecture	$E_{\text{Dir}}(X)$		$\mu(X)$		Erank(X)	$\|X\|_F$	Accuracy ratio
	standard	normalized	standard	normalized			
LeakyReLU + GCN (DAD)	✗	✗	✗	✗	✓*	✗	0.2406
Tanh + GCN (DAD)	✗	✗	✗	✗	✓*	✗	0.1937
LeakyReLU + GCN (DAD) + bias	✗	✗	✗	✗	✓*	✗	0.2547
Tanh + GCN (DAD) + bias	✗	✗	✗	✗	✓*	✗	0.1995
LeakyReLU + GCN (DAD) + layernorm	✗	✗	✗	✗	✓*	✗	0.1825
Tanh + GCN (DAD) + layernorm	✗	✗	✗	✗	✓*	✗	0.1807
LeakyReLU + GAT	✗	✗	✗	✗	✓*	✗	0.2534
Tanh + GAT	✗	✗	✗	✗	✓*	✗	0.1418

Table 1: **Trained cases**. Experimental results on networks trained for node classification on the Cora dataset. ✓ indicates a clear decay of the metric to zero. For Erank(X), we measure Erank$(X) - r^*$ for some $r^* > 0$ and use ✓for the case $r^* = 1$ and ✓* for $r^* > 1$ but $r^* \ll \min\{N, d\}$, typically less than 1.5. ✗ indicates the metric stays roughly constant, or $r^* \approx \min\{N, d\}$ for Erank(X). Accuracy ratio denotes the ratio between the classification accuracies of 2-layered GNN and 24-layered GNN. DAD and DA denote the message-passing matrices $\widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}$ (non-stochastic) and $\widetilde{D}^{-1}\widetilde{A}$ (stochastic), respectively.

Architecture	$E_{\text{Dir}}(X)$		$\mu(X)$		Erank(X)	$\|X\|_F$
	standard	normalized	standard	normalized		
LeakyReLU + GCN (DAD) + identity weights	✓	✓	✗	✓	✓	✗
Tanh + GCN (DAD) + identity weights	✓	✓	✗	✓	✓	✗
LeakyReLU + GCN (DA) + identity weights	✗	✗	✓	✓	✓	✗
Tanh + GCN (DA) + identity weights	✗	✗	✓	✓	✓	✗
LeakyReLU + GCN (DAD) + small weights	✓	✓	✓	✓	✓	✓
Tanh + GCN (DAD) + small weights	✓	✓	✓	✓	✓	✓
LeakyReLU + GCN (DA) + small weights	✓	✗	✓	✓	✓	✓
Tanh + GCN (DA) + small weights	✓	✗	✓	✓	✓	✓
LeakyReLU + GCN (DAD) + large weights	✗	✓	✗	✓	✓	✗
Tanh + GCN (DAD) + large weights	✗	✗	✗	✗	✗	✗
LeakyReLU + GCN (DA) + large weights	✗	✗	✗	✓	✓	✗
Tanh + GCN (DA) + large weights	✗	✗	✗	✗	✗	✗

Table 2: **Synthetic cases**. Experimental results of all major metrics on networks with 200 layers with random weights. Notations have the same meaning as in table 1.

Similarly, Wu et al. (2023) analyse GATs as in (5), i.e., architectures with row-stochastic message-passing matrix $P^{(l)}$ that vary at each layer, and propose the following metric:

$$\mu(X) = \|X - \mathbf{1}\gamma\|_F, \qquad \gamma = \frac{\mathbf{1}^\top X}{N}. \tag{10}$$

$\mu(X)$ measures the alignment of each feature with the dominant eigenvector of the message-passing matrix, rather than the alignment of features with one another as in the case of the Dirichlet energy. For GATs, all $P^{(l)}$ share the same all-ones dominant eigenvector $\mathbf{1}$.

A major result from (Wu et al. 2023) is that if $\|\prod_{l=1}^{L} |W|^{(l)}\|$ is bounded, then $\mu(X)$ provably decays to zero for a wide range of ReLU-like activation functions.

Obviously, μ can also be used to measure the oversmoothing of GCN architectures. However, for it to be meaningful, a minor adjustment is required when GCNs are employed with a message-passing matrix P that is not stochastic. In such cases, the dominant eigenvector x_D of P may differ from $\mathbf{1}$. To account for this, μ should be modified to

$$\mu(X) = \|X/x_D - \mathbf{1}\gamma\|_F, \quad \gamma = \frac{\mathbf{1}^\top (X/x_D)}{N},$$

with the division done point-wise on each column of X.

It is important to note that, similar to the Dirichlet energy, this measure has the drawback that it can be zero when the feature norms are small, thereby failing to capture the distance from the dominant eigenspace in such cases. To address this limitation, a more appropriate version of this metric is its normalized form, $\mu(X)/\|X\|_F$. Interestingly, despite its intuitive appeal, we have not encountered this normalized version used in any existing literature.

It is evident that the primary purpose of all these energy functions is to measure the alignment of features with a specific rank-one embedding space spanned by the dominant eigenvector of the message-passing matrix P. However, these metrics are heavily dependent on the architecture being used and, in particular, on prior knowledge of the dominant

eigenvector of P. Furthermore, they generally quantify the convergence of the neural network towards a very specific rank-one fixed point. While this approach is valid in certain scenarios, we argue that it oversimplifies the oversmoothing phenomenon. Oversmoothing is not merely about convergence to a rank-one space but is fundamentally characterized by the reduction of the rank of the feature representation, which may collapse into a low-dimensional space without necessarily being rank-one. To address this limitation, we propose the use of the effective rank as a more general and powerful tool for quantifying oversmoothing.

Effective Rank

The rank of a matrix is defined as the number of its non-zero singular values. However, this notion is highly sensitive to noise. To mitigate this sensitivity, one possible approach is to set a minimum threshold for the singular values (Feng et al. 2022). Alternatively, continuous relaxations of the rank can be considered to provide a more robust measure.

In this work, we focus on the effective rank (Roy and Vetterli 2007). To define the effective rank of a matrix X, let $\sigma_1 > \sigma_2 > \cdots > \sigma_{\min\{N,d\}}$ denote the singular values of X. Define the normalized singular value probabilities as $p_k = \sigma_k / \sum_i \sigma_i$, and set the effective rank as

$$\text{Erank}(X) = \exp\left(-\sum_k p_k \log p_k\right). \tag{11}$$

For completeness, we note that other continuous relaxations of the rank exist. Examples include $\|X\|_*/\|X\|_2$, the stable rank $\|X\|_*^2/\|X\|_F^2$, and $\|X\|_F^2/\|X\|_2^2$, where $\|X\|_* = \sum_i \sigma_i$ is the nuclear norm (Rudelson and Vershynin 2006; Arora et al. 2019).

It is important to note that $E_{\text{Dir}}(X)$ and $\mu(X)$ are only informative for confirming oversmoothing when they decay to zero. In contrast, a small $\text{Erank}(X)$ directly corresponds to the collapse of features onto a low-dimensional space, capturing oversmoothing that can be correlated with a drop in network performance (this is illustrated in figure 1).

We note that prior work has explored connections between effective rank and graph oversmoothing (Guo et al. 2023), particularly as a means to justify a proposed normalization scheme. This paper has a different goal: emphasize strong empirical evidence supporting the use of effective rank to quantify oversmoothing. Through extensive numerical investigations, we demonstrate the superiority of effective rank over norm-based energy metrics for this purpose.

Experiments

In this section, we validate the robustness of the effective rank in quantifying oversmoothing in GNNs against the available metrics in the literature. The experiments examine the behaviours of different metrics with homogeneous (LeakyReLU) and subhomogeneous (Tanh) activation functions and consider the following two scenarios:

- **Trained cases**: where the weights are obtained following the normal training dynamics of a GNN. We use the standard setup of GCN and GAT as stated in equation (2) and (5) with a fixed hidden dimension of 32. For each setup, eight GNNs of different depths ranging from 2 to 24 are trained separately, and the oversmoothing metrics are computed at the last hidden layer.
- **Synthetic cases**: where the weights are either the identity or randomly sampled at each layer from a distribution $\mathcal{N}(0, s^2)$. s is the standard deviation that can be varied depending on the setting: small weights ($s = 0.1$) lead to an exponentially decaying $\|X\|_F$, and large weights ($s = 1$) lead to an exploding $\|X\|_F$ for uncapped activation functions. For identity weights, $\|X\|_F$ is roughly constant. The adjacency matrix A is obtained from a synthetic 10-node Barabasi Albert Graph that guarantees connectivity between all nodes. The feature initialization $X^{(0)}$ is sampled from $\mathcal{N}(0, \mathbf{1})$.

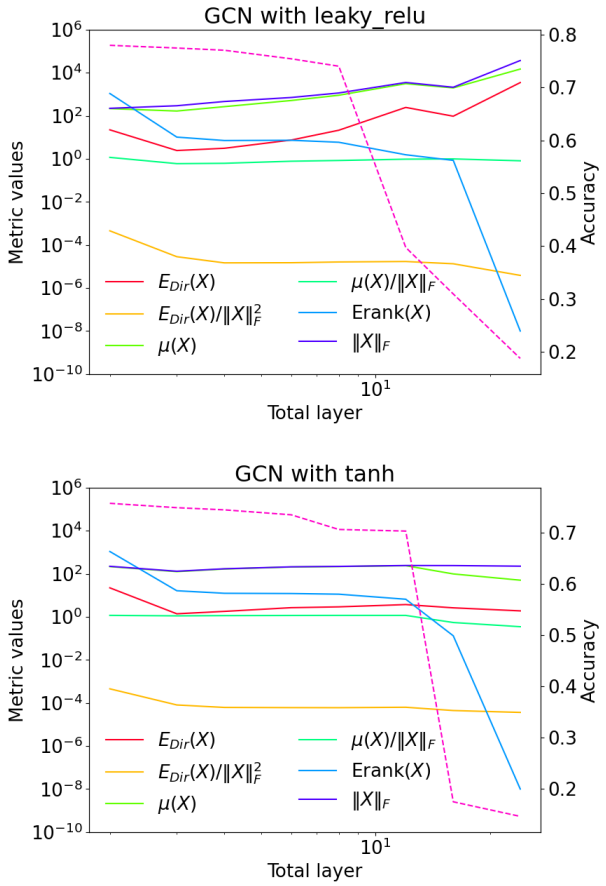

Figure 1: Two examples of GCNs trained with different activations, corresponding to the first two rows in table 1. The pink dashed line indicates the classification accuracy of the trained GNNs. In this particular case, $r^* = 1.52211511$ and 1.551852401 for the LeakyReLU (top) and tanh (bottom) plots, respectively. Note that the effective rank of the input features $\text{Erank}(X^{(0)})$ is about 1084.

The full results are shown in table 1 and 2.

In the synthetic settings, we observe that the standard $E_{\text{Dir}}(X)$ and $\mu(X)$ only decays for $\widetilde{D}^{-1/2}\widetilde{A}\widetilde{D}^{-1/2}$- or $\widetilde{D}^{-1}\widetilde{A}$-normalized P, which agrees with our analysis from

previous sections. In comparison, $\mu(X)$ normalized by both x_D and $\|X\|_F$ as well as $\text{Erank}(X)$ can more generally capture a wider range of oversmoothing settings.

In the trained settings, $\text{Erank}(X)$ always decays towards a small value close to one but not exactly one, while $E_{\text{Dir}}(X)$ and $\mu(X)$ are still far from converging to zero, as that would correspond to the effective rank being one.

Acknowledgments

PD was supported by the MUR-PRO3 grant STANDS. PD and FT are members of the Gruppo Nazionale Calcolo Scientifico - Istituto Nazionale di Alta Matematica (GNCS-INdAM).

References

Arora, S.; Cohen, N.; Hu, W.; and Luo, Y. 2019. Implicit Regularization in Deep Matrix Factorization. In *NeurIPS*.

Brody, S.; Alon, U.; and Yahav, E. 2021. How Attentive Are Graph Attention Networks? In *ICLR*.

Cai, C.; and Wang, Y. 2020. A Note on Over-Smoothing for Graph Neural Networks. arXiv:2006.13318.

Chen, G.; Zhang, J.; Xiao, X.; and Li, Y. 2022. Preventing Over-Smoothing for Hypergraph Neural Networks. arXiv:2203.17159.

Damianou, A.; Fabbri, F.; Gigioli, P.; De Nadai, M.; Wang, A.; Palumbo, E.; and Lalmas, M. 2024. Towards Graph Foundation Models for Personalization. In *WWW*.

Di Giovanni, F.; Rowbottom, J.; Chamberlain, B. P.; Markovich, T.; and Bronstein, M. M. 2023. Understanding Convolution on Graphs via Energies. *Transactions on Machine Learning Research*.

Epping, B.; René, A.; Helias, M.; and Schaub, M. T. 2024. Graph Neural Networks Do Not Always Oversmooth. arXiv:2406.02269.

Fan, W.; Ma, Y.; Li, Q.; He, Y.; Zhao, E.; Tang, J.; and Yin, D. 2019. Graph Neural Networks for Social Recommendation. In *WWW*.

Feng, R.; Zheng, K.; Huang, Y.; Zhao, D.; Jordan, M.; and Zha, Z.-J. 2022. Rank Diminishing in Deep Neural Networks. In *NeurIPS*.

Guo, X.; Wang, Y.; Du, T.; and Wang, Y. 2023. ContraNorm: A Contrastive Learning Perspective on Oversmoothing and Beyond. In *ICLR*.

Kipf, T. N.; and Welling, M. 2016. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Li, Q.; Han, Z.; and Wu, X.-M. 2018. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. In *AAAI*.

Oono, K.; and Suzuki, T. 2019. Graph Neural Networks Exponentially Lose Expressive Power for Node Classification. In *ICLR*.

Peng, C.; Xia, F.; Naseriparsa, M.; and Osborne, F. 2023. Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review*, 1–32.

Peng, S.; Sugiyama, K.; and Mine, T. 2022. SVD-GCN: A Simplified Graph Convolution Paradigm for Recommendation. In *CIKM*.

Roy, O.; and Vetterli, M. 2007. The Effective Rank: A Measure of Effective Dimensionality. In *EUSIPCO*.

Rudelson, M.; and Vershynin, R. 2006. Sampling from Large Matrices: An Approach through Geometric Functional Analysis. arXiv:math/0503442.

Rusch, T. K.; Bronstein, M. M.; and Mishra, S. 2023. A Survey on Oversmoothing in Graph Neural Networks. arXiv:2303.10993.

Rusch, T. K.; Chamberlain, B. P.; Rowbottom, J.; Mishra, S.; and Bronstein, M. M. 2022. Graph-Coupled Oscillator Networks. In *ICML*.

Shi, W.; and Rajkumar, R. 2020. Point-GNN: Graph Neural Network for 3D Object Detection in a Point Cloud. In *CVPR*. Seattle, WA, USA. ISBN 978-1-7281-7168-5.

Tian, L.; Zhou, X.; Wu, Y.-P.; Zhou, W.-T.; Zhang, J.-H.; and Zhang, T.-S. 2022. Knowledge Graph and Knowledge Reasoning: A Systematic Review. *Journal of Electronic Science and Technology*, 20(2): 100159.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2017. Graph Attention Networks. In *ICLR*.

Wu, X.; Ajorlou, A.; Wu, Z.; and Jadbabaie, A. 2023. Demystifying Oversmoothing in Attention-Based Graph Neural Networks. In *NeurIPS*.