Transferability for Graph Convolutional Networks

Christian Koke [1] **Abhishek Saroha** [1] **Yuesong Shen** [1] **Marvin Eisenberger** [1] **Michael Bronstein** [2]
Daniel Cremers [1]

Abstract

This work develops a novel and widely applicable transferability theory for graph convolutional networks; covering architectures based on undirected- and recently introduced directed convolutional filters. Experiments on real-world data validate the developed theory in practice.

1. Introduction

Graph Convolutional Networks (Kipf & Welling, 2017; Bruna et al., 2014; Defferrard et al., 2016) are a prominent class of machine learning architectures adapted to operating on graph structured data. Such GCNs (also known as spectral convolutional networks) continue to set the state of the art on a diverse selection of tasks (Bianchi et al., 2019; He et al., 2021; 2022; Wang & Zhang, 2022; Koke & Cremers, 2024). While for a long time thought to be only deployable on undirected graphs (Bronstein et al., 2021), they were recently successfully extended also to the directed setting (Koke & Cremers, 2024).

A key question for GCNs is that of **transferability** (Levie et al., 2019; Ruiz et al., 2023). This concept encodes the ability to train a GCN on one set of graphs and then successfully apply it to previously unseen 'similar' graphs. Typical examples of such similar graphs arise from coarse-graining an original graph, re-meshing (c.f. e.g. (Botsch & Kobbelt, 2004)) if the graph discretizes an underlying object such as manifold, or from re-sampling if the original graph in question is drawn from a statistical distribution.

In the literature, transferability has been investigated almost exclusively in the setting of (very) large and undirected graphs taken to faithfully discretise a common underlying "continuous" object such as fixed metric measure space (Levie et al., 2019), the same graphon (Ruiz et al., 2020; Maskey et al., 2021) or the same graphop (Le & Jegelka, 2023). Beyond the deterministic setting, transferability in (high) probability has also been investigated for graphs *statistically sampled* from the same underlying manifold (Wang et al., 2021; 2022) or drawn from the same statistical distribution (Keriven et al., 2020; Gao et al., 2021).

Contributions: Here we introduce a novel approach to transferability based on information-diffusion processes on graphs. This new approach also covers directed graphs, allows for transferability results beyond the asymptotic setting of large graphs and applies to typical examples such as transferability between graphs arising through coarsification or from edge-rewiring. Numerical experiments on real-world data validate our findings in practice.

2. Preliminaries

Weighted directed graphs: A (potentially) directed graph $G := (\mathcal{G}, \mathcal{E})$ is a collection of nodes $\mathcal{G}$ and edges $\mathcal{E} \subseteq \mathcal{G} \times \mathcal{G}$. We assume (real) edge-weights $w_{ij} \geqslant 0$ (with $w_{ij} \not\equiv w_{ji}$) and allow nodes $i \in \mathcal{G}$ to have individual **node-weights** $\mu_i > 0$. In a social network, a node weight $\mu_i = 1$ might e.g. signify that node i represents a single user, while a node with $\mu_j > 1$ represents a group of users.

Feature spaces: Given F-dimensional node features on a graph with $N = |\mathcal{G}|$ nodes, we collect individual node-feature vectors into a feature matrix X of dimension $N \times F$. Taking into account our node weights, we equip the space of such signals with an inner-product according to $\langle X, Y \rangle = \mathrm{Tr}(X^* M Y) = \sum_{i=1}^{N} \sum_{j=1}^{F} (\overline{X}_{ij} Y_{ij}) \mu_i$ with $M = \mathrm{diag}\left(\{\mu_i\}\right)$ the diagonal matrix of node-weights.

Characteristic Operators: Information about the geometry of a graph is encapsulated into the set of edge weights, collected into the (weighted) adjacency matrix A. Various characteristic operators such as Laplacians (e.g. (Hein et al., 2006; Maskey et al., 2023)) and (re-)normalized adjacency matrices (Kipf & Welling, 2017; Rossi et al., 2023; Koke & Cremers, 2024) may be then be derived. Our developed theory extends to all such choices. Results below will depend on the failure of characteristic operators to be unitarily diagonalizable, as measured via the characteristic operator L's **departure from normality** $\nu^2(L) = (\|L\|_F^2 - \sum_{\lambda_k \in \sigma(L)} |\lambda_k|^2)$. Since in the undirected

[1]Technical University Munich [2]University of Oxford. Correspondence to: Christian Koke <christian.koke@tum.de>.

Accepted as an extended abstract for the Geometry-grounded Representation Learning and Generative Modeling Workshop at the 41st International Conference on Machine Learning, ICML 2024, Vienna, Austria.

setting $\nu(L) = 0$, we may think of $\nu(L)$ as measuring the 'severity of the directedness' of the underlying graph.

Spectral Convolutional Filters: A spectral graph convolutional filter is constructed by applying a learnable function $h_\theta(\cdot)$ to an underlying characteristic operator L to build up the filter matrix $h_\theta(L)$. If the operator $L = V^{-1}\Lambda V$ is diagonalizable (e.g. on undirected graphs), spectral convolutional filters $h_\theta(L)$ are defined as $h_\theta(L) = V^{-1}h_\theta(\Lambda)V$, with h_θ applied to the eigenvalues Λ as $h_\theta(\Lambda) = \text{diag}\,(h_\theta(\lambda_1), ..., h_\theta(\lambda_N))$. If L is not diagonalizable, a more subtle definiton is used (c.f. Appendix B)

In practice one avoids working with the expensive eigendecomposition $h_\theta(L) = V^{-1}h(\Lambda)V$ by parametrizing a generic filter function $h_\theta(\cdot)$ as a weighted sum over 'simpler' basis functions $\{\psi_i\}_{i\in I} =: \Psi$ as $h_\theta(\cdot) := \sum_{i\in I}\theta_i \cdot \psi_i(\cdot)$. These simpler functions $\psi_i(\cdot)$ may then e.g. be chosen as polynomials (Defferrard et al., 2016; He et al., 2021; 2022; Koke & Cremers, 2024) or rational functions (Bianchi et al., 2019; Koke & Cremers, 2024), with $\psi_i(L)$ then simply given as a polynomial (or rational function) in the matrix L. Complete filters $h_\theta(L)$ are then parametrized via the learnable coefficients $\{\theta_i\}_{i\in I}$ as $h_\theta(L) := \sum_{i\in I}\theta_i \cdot \psi_i(L)$.

Graph Convolutional Networks: Learnable filters are then combined into a (K-layer) graph convolutional network mapping initial node-features $X \in \mathbb{C}^{N\times F}$ to final representations $X^K \in \mathbb{C}^{N\times F_K}$. With bias matrices $B^\ell \in \mathbb{C}^{N\times F_\ell}$ ($B_{:j} = b_j \cdot \mathbb{1}_G$) and weight matrices $W_i^\ell \in \mathbb{C}^{F_{\ell-1}\times F_\ell}$, layer-updates are then implemented as:

$$X^\ell = \rho\left(\sum_{i\in I}\psi_i(L)\cdot X^{\ell-1}\cdot W_i^\ell + B^\ell\right) \qquad (1)$$

Here ρ is a point-wise non-linearity, for which we assume $\rho(0) = 0$ and $|\rho(a) - \rho(b)| \leqslant |a - b|$ ($a, b \in \mathbb{C}$). With basis functions $\Psi = \{\psi_i\}_{i\in I}$ and weights and biases represented as $\mathscr{W}$ and $\mathscr{B}$, we denote the output of a graph neural network based on the characteristic operator L and applied to the node feature matrix X as $\Phi = \Phi_{\mathscr{W},\mathscr{B},\Psi}(L, X)$.

3. Transferability via Information-Diffusion

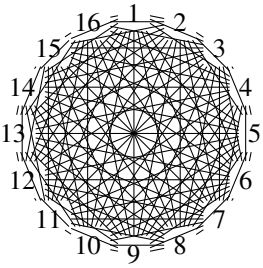


Figure 1. Perturbed K_{16}

We are interested in significant perturbations of graphs which nevertheless should be considered small: Consider e.g. the (unnormalized) graph Laplacian ($L = \Delta$) on the unweighted complete N-node graph K_N. Let $\tilde{L} = \tilde{\Delta}$ be the Laplacian obtained from deleting a single edge from K_N. Clearly $\|L - \tilde{L}\| = 2$ irrespective of N. For $N = 2$ deleting the only present edge in K_2 clearly amounts to a significant change in geometry. For larger N however, deleting a single edge intuitively corresponds to a comparatively small change in the graph's geometry (c.f. K_{16} with edge '$1 \leftrightarrow 5$' removed in Fig 1).

This intuition is related to the way information diffuses over the underlying graph: On a two node graph, deleting the only present edge clearly completely disrupts information flow. Deleting an edge within a *large* fully connected clique hardly modifies the way information is diffused. To quantify these considerations, we recall that the diffusion equation on a graph is given as (Veerman & Lyons, 2020; Gasteiger et al., 2019b) $dX(t)/dt = -L\cdot X(t)$ with solution $X(t) = e^{-Lt}\cdot X(0)$. Solving this for the same initial condition $X(0)$ but with diffusion implemented via L and $\tilde{L}$ respectively, we find $\|e^{-Lt}X(0) - e^{-\tilde{L}t}X(0)\| \lesssim e^{-(N-2)t}$ (c.f. Appendix G). Hence information indeed diffuses similarly over the distinct graph structures determined by L and $\tilde{L}$ if $N \gg 1$.

The observation that even if $\|L - \tilde{L}\| \geqslant 1$ information might still diffuse similarly over the corresponding graph structures $G, \tilde{G}$ provides the core of the transferability theory we develop here: We consider two graphs to be similar, if the information diffusion flows $e^{-tL}, e^{-t\tilde{L}}$ generated by their characteristic operators are similar.[1] We then desire that networks are transferable between such similar graphs.

If the two graphs share a common node set, similarity is captured by $\sup_{t\geqslant 0}\|e^{-Lt} - e^{-\tilde{L}t}\| \ll 1$. If the node sets are distinct, we facilitate contact between the two graphs via linear intertwining operators J and $\tilde{J}$, with J linearly mapping features from G to $\tilde{G}$ and $\tilde{J}$ mapping in the opposite direction. We may then consider two notions of comparing diffusion flows on graphs G and $\tilde{G}$ (c.f. also Appendix H):

Definition 3.1. Two graphs $G, \tilde{G}$ are **unidirectionally similar** under the identification J if $\sup_{t\geqslant 0}\|Je^{-Lt} - e^{-\tilde{L}t}J\| \ll 1$. They are **bidirectionally similar** if $\|e^{-Lt} - \tilde{J}e^{-\tilde{L}t}J\| \leqslant \eta(t)$ for some (fast decaying) function $\eta(t) \geqslant 0$ with $\lim_{t\to\infty}\eta(t) = 0$ and $\eta(0) = \|Id_G - \tilde{J}J\|$.

Single Filter Transferability: We now want to characterize the class of filters that are transferable between graphs which are similar in the sense of Definition 3.1. This class will turn out to consist of functions that arise as *Laplace transforms* (c.f. (Widder, 1941) or Appendix I.2):

Definition 3.2. Let $\hat{\psi}$ be a (generalized) function defined on $\mathbb{R}_{\geqslant 0} := [0, \infty)$ for which $\|\hat{\psi}\|_1 := \int_0^\infty |\hat{\psi}(t)|dt < \infty$. A **Laplace Transform Filter** ψ is any function defined as $\psi(z) := \int_0^\infty e^{-tz}\hat{\psi}(t)dt$.

Here a generalized function $\hat{\psi}$ is meant to be understood in a distributional sense: We e.g. allow $\hat{\psi}(t)$ to be given as a (complex multiple of) the dirac delta distribution $\hat{\psi}_{\delta_{t_0}}(t) := c\delta(t - t_0)$ with $c \in \mathbb{C}$ and $t_0 \geqslant 0$ (c.f. Appendix I).

[1] Appendix I.6 discusses implicit assumptions on $L, \tilde{L}$.

Example 3.3. Considering $\hat{\psi}_k = \delta(t - kt_0)$ for $t_0 > 0$ and $k \in \mathbb{N}$ yields $\psi_k(z) = e^{-(kt_0)z}$. Using this set of **exponential basis-functions** $\Psi^{\text{Exp}} = \{e^{-(kt_0)z}\}_{k\in\mathbb{N}}$ yields a wide class of filter functions (c.f. Appendix I.2).

Example 3.4. Defining $\hat{\psi}_k := (-t)^{k-1}e^{-\lambda t}$ yields $\psi_k(z) = (z+\lambda)^{-k}$. Using this set of **resolvent basis-functions** $\Psi^{\text{Res}} = \{(z+\lambda)^{-k}\}_{k\in\mathbb{N}}$ yields a broad function class $\{h_\theta(\cdot) := \sum_i \theta_i \cdot \psi_i(\cdot)\}$ as well (c.f. Appendix I.2). The name arises since corresponding filters $\psi_k(L) = \left[(L+\lambda Id)^{-1}\right]^k$ are given as powers of the **resolvent** $(L+\lambda Id)^{-1}$ of the operator L.

Since we can write $\psi(L) = \int_0^\infty \hat{\psi}(t)e^{-tL}dt$, we may think of Laplace transform filters as a weighted sum over diffusion processes that have progressed to various times $t \in [0,\infty)$. As we prove in Appendix I.3, this property endows such filters with transferability in the setting of Definition 3.2:

Theorem 3.5. We have $\|J\psi(L) - \psi(\tilde{L})J\| \leqslant \|\hat{\psi}\|_1 \cdot \sup_{t\geqslant 0}\|Je^{-Lt} - e^{-\tilde{L}t}J\|$ in the *unidirectional* setting. In the *bidirectional* setting $\|\psi(L) - \tilde{J}\psi(\tilde{L})J\| \leqslant \int_0^\infty |\hat{\psi}(t)|\eta(t)dt$ holds true.

In the unidirectional setting, $\|\hat{\psi}\|_1$ hence provides the single filter stability constant. In the bidirectional setting, we note that if $\hat{\psi}(t) = \delta(t)$, we have $\int_0^\infty |\hat{\psi}(t)|\eta(t)dt = \|Id_G - \tilde{J}J\| > 0$ irrespective of $L, \tilde{L}$. Hence for filters to be transferable in the bidirectional setting, we need to assume that the generalized function $\hat{\psi}$ contains no dirac-delta at $t = 0$; or equivalently (as we show in Appendix I.4):

Corollary 3.6. Consider a sequence of graphs G_n for which $\|e^{-L_n t} - \tilde{J}_n e^{-\tilde{L}t}J_n\| \to 0$. Then for a Laplace transform filter ψ, we have $\|\psi(L_n) - \tilde{J}_n\psi(\tilde{L})J_n\| \to 0$ if and only if $\lim_{r\to\infty}\psi(r) = 0$.

While there exist computational methods for evaluating the quantity $\sup_{t\geqslant 0}\|Je^{-Lt} - e^{-\tilde{L}t}J\|$ numerically (Braker Scott, 2021), this is sometimes cumbersome to do in practice. For this reason, we here provide estimates in terms of a different quantity, that is often more accessible in practice (c.f. also our example below). This quantity makes use of the concept of the resolvent $R_\lambda(L) := (L+\lambda Id)^{-1}$ of the operator L; introduced in Example 3.4 above:

Theorem 3.7. Let ψ be a Laplace transform filter. There exists a constant $C = C_{\psi,\nu(L),\nu(\tilde{L})} < \infty$ so that we have $\|J\psi(L) - \psi(\tilde{L})J\| \leqslant C \cdot \|J(L+\lambda Id)^{-1} - (\tilde{L}+\lambda Id)^{-1}J\|$.

If either $\tilde{J}J = Id_{\tilde{G}}$ or $J\tilde{J} = Id_G$ (as is e.g. the case in our coarse-graining example below, Theorem 3.7 directly translates to the bidirectional setting. If this is not the case, we still have the following bidirectional convergence result; proved together with Theorem 3.7 in Appendix I.4:

Theorem 3.8. Consider a graph sequence G_n with $\|(L_n + \lambda Id)^{-1} - \tilde{J}_n(\tilde{L}+\lambda Id)^{-1}J_n\| \to 0$. If the graphs are directed, assume eigenvalues of all L_ns lie within a cone of opening angle $\alpha < \pi$ symmetric about the real axis. Then $\|\psi(L_n) - \tilde{J}_n\psi(\tilde{L})J_n\| \to 0$ iff $\lim_{r\to\infty}\psi(r) = 0$.

Network Transferability: Building on this, we find the following for the transferability of *networks*:

Theorem 3.9. Let $\Phi_{\mathscr{W},\mathscr{B},\Psi}$ be a K-layer deep GCN. Assume that $\sum_{i\in I}\|W_i^\ell\| \leqslant W$ and $\|B^\ell\| \leqslant B$. Choose $C \geqslant \|\Psi_i(\tilde{L})\|$ $(i \in I)$ and w.l.o.g. assume $CW > 1$. Assume $\rho(J\tilde{X}) = J\rho(\tilde{X})$ and if biases are enabled, assume $J\mathbb{1}_G = \mathbb{1}_{\tilde{G}}$. With this, we have with $\delta = \max_{i\in I}\{\|J\psi_i(L) - \psi_i(\tilde{L})J\|\}$ that

$$\|J\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \Phi_{\mathscr{W},\mathscr{B},\Psi}(\tilde{L},JX)\| \leqslant \left[K\cdot C^K W^{K-1}\cdot\left(\|X\| + \frac{1}{CW-1}B\right)\right]\cdot\delta.$$

We prove Theorem 3.9 together with the corresponding result for the bidirectional setting in Appendix I.7. Extensions to graph level tasks are discussed in Appendix I.8.

Example: Coarse-Graining Graphs. Deferring additional examples to Appendix J, we here consider graphs containing clusters of nodes connected by significantly larger edge weights than those of edges outside of these clusters. This might for example arise for weighted graphs discretizing underlying continuous spaces: Here, edge weights are typically set to inverse discretization length ($w_{ij} \sim d_{ij}^{-1}$), which might vary over the graph (Post, 2012; Post & Simmer, 2021). Strongly connected sub-graphs then correspond to clusters of spatially closely co-located nodes. Alternatively, such different scales can occur in social networks; e.g. if edge-weights are set to number of exchanged messages.

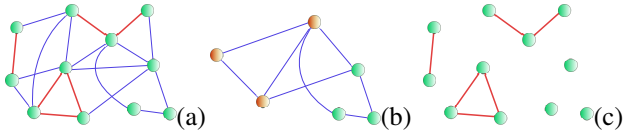


Figure 2. (a) G with red stong edges (b) Coarse grained $\underline{G}$ (c) G_{high}

From a diffusion perspective, information in a graph equalizes faster along edges with large weights. In the limit where edge-weights within certain sub-graphs tend to infinity, information within these clusters equalizes immediately and such sub-graphs should thus effectively behave as single nodes. We might thus consider a course grained graph $\underline{G}$ where these strongly connected clusters are indeed fused together and represented only via single nodes. The corresponding node set $\underline{\mathcal{G}}$ of $\underline{G}$ is then given by the set of connected components in G_{high} Edges $\underline{\mathcal{E}}$ are given by elements $(R,P) \in \underline{\mathcal{G}} \times \underline{\mathcal{G}}$ with non-zero accumulated edge weight $\underline{W}_{RP} = \sum_{r\in R}\sum_{p\in P} W_{rp}$. Node weights in $\underline{G}$ are defined accordingly by aggregating as $\underline{\mu}_R = \sum_{r\in R}\mu_r$. To compare signals on these two graphs $G, \underline{G}$, we define intertwining operators $J^\downarrow, J^\uparrow$ transferring information between

the two graphs: Let x be a scalar graph signal and let $\mathbb{1}_R$ be the vector that has 1 as entry for nodes $r \in R$ and is zero otherwise. Denote by u_R the entry of u at node $R \in \underline{\mathcal{G}}$. Projection $J^{\downarrow}$ is then defined component-wise by evaluation at node $R \in \underline{\mathcal{G}}$ as $(J^{\downarrow}x)_R = \langle \mathbb{1}_R, x\rangle / \underline{\mu}_R$. Interpolation is defined as $J^{\uparrow}u = \sum_{R\in\underline{\mathcal{G}}} u_R \cdot \mathbb{1}_R$.

As proved in Appendix J.1, we have for Laplacians $\Delta, \underline{\Delta}$ on G and $\underline{G}$ that (with Δ_{high} the Laplacian for G_{high})

$$\|(\Delta + Id)^{-1} - J^{\uparrow}(\underline{\Delta} + Id)^{-1}J^{\downarrow}\| \lesssim 1/\lambda_1(\Delta_{\text{high}}). \quad (2)$$

Hence – as desired in view of Definition 3.1 – the transferability error for networks based on Laplace transform filters decreases inversely with increasing connectivity within G_{high} as measured via the first non-trivial eigenvalue λ_1 of Δ_{high} (scaling linearly with the edge-weights in G_{high}). In contrast other common, GNNs equip G and $\underline{G}$ with vastly different feature-vectors, as observed below.

4. Numerical Results

In this section we focus on showcasing the transferability properties established in Section 3 in a practical example. Following (Levie et al., 2019; Koke, 2023) in spirit, we consider transferability between original- and coarse grained versions of graphs. In order to evaluate on **real world data** we follow (Koke, 2023) and evaluate on the task of molecular property prediction, which allows to fairly compare properties of standard GNN architectures (Hu et al., 2020). Our dataset (QM7; (Rupp et al., 2012)) contains 7165 organic molecules; each containing both hydrogen and heavy atoms. Each molecule is represented by a weighted adjacency matrix, whose entries $A_{ij} = Z_iZ_j \cdot |\vec{x}_i - \vec{x}_j|^{-1}$ correspond to Coulomb repulsions between atoms i and j. Prediction target is molecular atomization energy. We also consider a coarsified version of QM7: Here we fuse together each heavy atom with its surrounding hydrogen atoms into super-nodes. Appendix K.1 provides exact details and discusses baselines. We might interpret this *low-resolution* QM7 dataset as a model for data obtained from a resolution-limited observation process unable to resolve positions of individual (small) hydrogen atoms and only providing information about how many are bound to a given heavy atom.

We then consider two architectures using Laplace transform filters, with the set of basis functions Ψ (c.f. Section 2) given respectively as the set of exponential basis functions Ψ^{Exp} introduced in Example 3.3 and the set of resolvent basis-functions Ψ^{Res} introduced in Example 3.4. We compare the transferability properties of these architectures (LTF-Ψ^{Res} and LTF-Ψ^{Exp}) against those of typical GNNs.

Using the high-resolution graphs $\{G\}$ of QM7 and the low-resolution graphs $\{\underline{G}\}$ in coarsified-QM7, we then investigate transferability by confronting models during inference with a resolution-scale different from the one they were trained on: Table 1 collects corresponding results; including reference results for inference on same-resolution data.

Table 1. Regression using high- and low-resolution QM7

Mean Absolute Error (↓) on QM7 [kcal/mol]				
Training	**High Resolution**		**Low Resolution**	
Inference	**Low Resolution**	High Resolution	Low Resolution	**High Resolution**
GCN	$125.34_{\pm 2.47}$	$63.17_{\pm 0.92}$	$67.75_{\pm 3.73}$	$380.51_{\pm 30.33}$
ARMA	$206.50_{\pm 18.68}$	$62.18_{\pm 3.24}$	$62.30_{\pm 4.70}$	$301.44_{\pm 38.29}$
GATv2	$415.09_{\pm 96.57}$	$48.41_{\pm 19.20}$	$60.01_{\pm 3.34}$	$245.03_{\pm 90.97}$
ChebNet	$568.47_{\pm 37.70}$	$64.63_{\pm 1.21}$	$64.90_{\pm 4.55}$	$339.64_{\pm 101.30}$
SAG	$542.16_{\pm 27.33}$	$68.43_{\pm 1.93}$	$104.20_{\pm 3.92}$	$506.75_{\pm 60.57}$
BernNet	$765.22_{\pm 495.28}$	$83.76_{\pm 21.75}$	$90.52_{\pm 37.17}$	$594.62_{\pm 341.55}$
SAG-M	$285.53_{\pm 95.54}$	$66.22_{\pm 4.51}$	$73.57_{\pm 14.57}$	$307.67_{\pm 77.24}$
UFGNet	$620.21_{\pm 4.80}$	$13.71_{\pm 1.05}$	$24.53_{\pm 4.80}$	$156.44_{\pm 156.44}$
Lanczos	$939.87_{\pm 16.35}$	$10.55_{\pm 3.22}$	$83.11_{\pm 5.27}$	$654.61_{\pm 529.13}$
PushNet	$2442.59_{\pm 303.27}$	$60.94_{\pm 1.83}$	$69.25_{\pm 3.11}$	$124.08_{\pm 3.94}$
LTF-Ψ^{Res}	$16.54_{\pm 3.01}$	$16.53_{\pm 3.03}$	$15.79_{\pm 0.98}$	$13.80_{\pm 1.34}$
LTF-Ψ^{Exp}	$16.37_{\pm 1.71}$	$16.36_{\pm 2.16}$	$16.25_{\pm 1.41}$	$16.25_{\pm 1.41}$

We first note that for all baselines, the mean-absolute-errors (MAEs) made during inference increase significantly when going from a same-resolution setting to a cross-resolution setting. This shows clearly that **standard architectures are not transferable**. Their errors increase by factors of $\mathcal{O}(2)$ for simple methods (e.g. GCN) up to $\mathcal{O}(100)$ for sophisticated ones (e.g. UFGNet and Lanczos). MAEs of LTF-Ψ^{Res} and LTF-Ψ^{Exp} do not increase when going from a same- to a cross-resolution setting: Thus we see that **networks based on Laplace transform filters *are* transferable**. In cross-resolution settings, MAEs of LTF-Ψ^{Res} and LTF-Ψ^{Exp} are lower than that of all baselines by a factor of at least $\mathcal{O}(10)$ but up to $\mathcal{O}(100)$. It is interesting to observe that LTF-Ψ^{Res}'s best performance is achieved when only low-resolution training data is available, but inference is performed on high resolution data; a setup is likely to occur in real-life settings without high-quality training-data.

5. Conclusion

We developed a widely applicable transferability theory for (potentially directed) graph convolutional networks based on the intrinsic notion of information diffusion on graphs. In an example, we saw how our theory enables the design of networks transferable between graphs arising from one another via coarse-graining. This was confirmed experimentally: Networks designed according to the principles laid out by our developed theory were seen to be transferable between graphs describing the same object at different resolutions. Other architectures proved not transferable.

References

Arendt, W. APPROXIMATION OF DEGENERATE SEMIGROUPS. *Taiwanese Journal of Mathematics*, 5(2):279 – 295, 2001. doi: 10.11650/twjm/1500407337. URL https://doi.org/10.11650/twjm/1500407337.

Bandtlow, O. F. Estimates for norms of resolvents and an application to the perturbation of spectra. *Mathematische Nachrichten*, 267(1):3–11, 2004. doi: https://doi.org/10.1002/mana.200310149. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mana.200310149.

Bianchi, F. M., Grattarola, D., Livi, L. F., and Alippi, C. Graph neural networks with convolutional arma filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:3496–3507, 2019.

Blum, L. C. and Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.*, 131:8732, 2009.

Botsch, M. and Kobbelt, L. A remeshing approach to multiresolution modeling. In *Eurographics Symposium on Geometry Processing*, 2004. URL https://api.semanticscholar.org/CorpusID:16924063.

Braker Scott, C. *Diffusion Distance: Efficient Computation and Applications*. PhD Thesis. UNIVERSITY OF CALIFORNIA IRVINE, 2021.

Bronstein, M. M., Bruna, J., Cohen, T., and Velikovi, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.

Bruna, J., Zaremba, W., Szlam, A., and Lecun, Y. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

Defferrard, M., Bresson, X., and Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.

Gao, Z., Isufi, E., and Ribeiro, A. Stability of graph convolutional neural networks to stochastic perturbations. *Signal Process.*, 188:108216, 2021. doi: 10.1016/j.sigpro.2021.108216. URL https://doi.org/10.1016/j.sigpro.2021.108216.

Gasteiger, J., Bojchevski, A., and Günnemann, S. Predict then propagate: Graph neural networks meet personalized pagerank. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019a. URL https://openreview.net/forum?id=H1gL-2A9Ym.

Gasteiger, J., Weißenberger, S., and Günnemann, S. Diffusion improves graph learning. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13333–13345, 2019b. URL https://proceedings.neurips.cc/paper/2019/hash/23c894276a2c5a16470e6a31f4618d73-Abstract.html.

Hamilton, W. L., Ying, Z., and Leskovec, J. Inductive representation learning on large graphs. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 1024–1034, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

He, M., Wei, Z., Huang, Z., and Xu, H. Bernnet: Learning arbitrary graph spectral filters via bernstein approximation. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 14239–14251, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/76f1cfd7754a6e4fc3281bcccb3d0902-Abstract.html.

He, M., Wei, Z., and Wen, J. Convolutional neural networks on graphs with chebyshev approximation, revisited. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/2f9b3ee2bcea04b327c09d7e3145bd1e-Abstract-Confere
html.

Hein, M., Audibert, J.-Y., and von Luxburg, U. Graph laplacians and their convergence on random neighborhood graphs. *J. Mach. Learn. Res.*, 8:1325–1368, 2006. URL https://api.semanticscholar.org/CorpusID:1355782.

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. ISSN 0378-8733. doi: https://doi.org/10.1016/0378-8733(83)90021-7. URL https://www.sciencedirect.com/science/article/pii/0378873383900217.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. Open graph benchmark: Datasets for machine learning on graphs. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/fb60d411a5c5b72b2e7d3527cfc84fd0-Abstract.html.

Kato, T. *Perturbation theory for linear operators; 2nd ed.* Grundlehren der mathematischen Wissenschaften : a series of comprehensive studies in mathematics. Springer, Berlin, 1976. URL https://cds.cern.ch/record/101545.

Keriven, N., Bietti, A., and Vaiter, S. Convergence and stability of graph convolutional networks on large random graphs. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5a14d4963acf488e3a24780a84ac96c-Abstract.html.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Koke, C. Limitless stability for graph convolutional networks. In *11th International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=XqcQhVUr2h0.

Koke, C. and Cremers, D. Holonets: Spectral convolutions do extend to directed graphs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=EhmEwfavOW.

Le, T. and Jegelka, S. Limits, approximation and size transferability for GNNs on sparse graphs via graphops. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=kDQwossJuI.

Levie, R., Bronstein, M. M., and Kutyniok, G. Transferability of spectral graph convolutional neural networks. *CoRR*, abs/1907.12972, 2019. URL http://arxiv.org/abs/1907.12972.

Liao, R., Zhao, Z., Urtasun, R., and Zemel, R. S. Lanczosnet: Multi-scale deep graph convolutional networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BkedznAqKQ.

Maskey, S., Levie, R., and Kutyniok, G. Transferability of graph neural networks: an extended graphon approach. *CoRR*, abs/2109.10096, 2021. URL https://arxiv.org/abs/2109.10096.

Maskey, S., Paolino, R., Bacho, A., and Kutyniok, G. A fractional graph laplacian approach to oversmoothing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=kS7ED7eE74.

McCallum, A., Nigam, K., Rennie, J., and Seymore, K. Automating the construction of internet portals with machine learning. *Inf. Retr.*, 3(2):127–163, 2000. doi: 10.1023/A:1009953814988. URL https://doi.org/10.1023/A:1009953814988.

Post, O. *Spectral Analysis on Graph-like Spaces / by Olaf Post.* Lecture Notes in Mathematics, 2039. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st ed. 2012. edition, 2012. ISBN 3-642-23840-8.

Post, O. and Simmer, J. Graph-like spaces approximated by discrete graphs and applications. *Mathematische Nachrichten*, 294(11):2237–2278, 2021. doi: https://doi.org/10.1002/mana.201900108.

Rossi, E., Charpentier, B., Giovanni, F. D., Frasca, F., Gnnemann, S., and Bronstein, M. Edge directionality improves learning on heterophilic graphs, 2023.

Ruiz, L., Chamon, L. F. O., and Ribeiro, A. Graphon neural networks and the transferability of graph neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/12bcd658ef0a540cabc36cdf2b1046fd-Abstract.html.

Ruiz, L., Chamon, L. F. O., and Ribeiro, A. Transferability properties of graph neural networks. *IEEE Transactions*

on Signal Processing, 71:3474–3489, 2023. doi: 10.1109/TSP.2023.3297848.

Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108:058301, 2012.

Sahi, S. Harmonic vectors and matrix tree theorems, 2013.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. Collective classification in network data. *AI Magazine*, 29 (3):93, Sep. 2008. doi: 10.1609/aimag.v29i3.2157. URL https://ojs.aaai.org/index.php/aimagazine/article/view/2157.

Tao, T. *An Introduction to Measure Theory*. Graduate studies in mathematics. American Mathematical Society, 2013. ISBN 9781470409227. URL https://books.google.de/books?id=SPGJjwEACAAJ.

Teschl, G. *Mathematical Methods in Quantum Mechanics*. American Mathematical Society, 2014.

Veerman, J. J. P. and Lyons, R. A primer on laplacian dynamics in directed graphs. *Nonlinear Phenomena in Complex Systems*, 2020. URL https://api.semanticscholar.org/CorpusID:211066395.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Wang, X. and Zhang, M. How powerful are spectral graph neural networks. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:248987544.

Wang, Z., Ruiz, L., and Ribeiro, A. Stability of neural networks on riemannian manifolds. *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 1845–1849, 2021. URL https://api.semanticscholar.org/CorpusID:232110514.

Wang, Z., Ruiz, L., and Ribeiro, A. Convolutional neural networks on manifolds: From graphs and back. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pp. 356–360, 2022. doi: 10.1109/IEEECONF56349.2022.10051964.

Widder, D. V. *The Laplace Transform*, volume vol. 6 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1941.

Wihler, T. On the hlder continuity of matrix functions for normal matrices. *Journal of inequalities in pure and applied mathematics*, 10(4), Dec 2009. ISSN 1443-5756. URL https://www.emis.de/journals/JIPAM/images/276_09_JIPAM/276_09_www.pdf.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

A. Notation

We provide a summary of employed notational conventions:

Table 2. Notational Conventions

Symbol	Meaning
$\|\cdot\|$	The standard 2-norm
G	a graph
$\mathcal{G}$	Nodes of the graph G
N	number of nodes $\|\mathcal{G}\|$ in G
$\underline{G}$	Coarse grained version of graph G
μ_i	weight of node i
M	weight matrix
$\langle\cdot,\cdot\rangle$	inner product
A	(weighted) adjacency matrix
$D^{\text{in/out}}$	in/out-degree matrix
L^{in}	in-degree graph Laplacian
L	generic characteristic operator
L^*	hermitian adjoint of L
$L^\top$	transpose of L (used if and only if L has only real entries)
U	change-of-basis matrix to a basis of orthogonal eigenvectors (used in the undirected setting only)
V	change-of-basis matrix to a basis of eigenvectors (used in the diagonalizable setting only)
$\kappa(V)$	condition number of V
$\nu(L)$	departure from normality of L
$\sigma(L)$	spectrum (i.e. collection of eigenvalues) of L
λ	an eigenvalue
h	a filter function
$h(L)$	function h applied to operator L
Ψ	a filter bank
ψ_i	an element of a filter-bank
z	a complex number
$J^\downarrow, J^\uparrow$	projection and interpolation operator
$J, \tilde{J}$	intertwining operators
Φ	map associated to a graph convolution network
Ω	graph-level aggregation mechanism
$\mathcal{M}$	a manifold
Z_i	atomic charge of atom corresponding to node i
$\vec{x}_i$	Cartesian position of atom corresponding to node i
$\frac{Z_i Z_j}{\|\vec{x}_i - \vec{x}_j\|}$	Coulomb interaction between atoms i and j
$\|\vec{x}_i - \vec{x}_j\|$	Euclidean distance between x_i and x_j

B. Additional Details on spectral convolutional filters on directed graphs

For a detailed discussion, the reader is referred to (Koke & Cremers, 2024); which this appendix follows closely.

On undirected graphs, one may apply generic functions $\{h\}$ to the a characteristic operator $L = U^\top \Lambda U$ employing the complete eigendecomposition of L as $h(L) := U^\top h_\theta(\Lambda) U$. On directed graphs L is generically not even diagonalizable. Here (Koke & Cremers, 2024) discussed a different approach to consistently defining the matrix $h(T)$: One restricts h to be a **holomorphic** function: For a given subset U of the complex plane, these are the complex valued functions $h : U \to \mathbb{C}$ for which the complex derivative $dh(z)/dz$ can be defined everywhere.

Any value $h(\lambda)$ of such a function can then be reproduced by calculating an integral of the function h along a path Γ encircling λ (c.f. Fig. 3) as

$$h(\lambda) = -\frac{1}{2\pi i} \oint_{\Gamma} h(z) \cdot (\lambda - z)^{-1} dz. \tag{3}$$

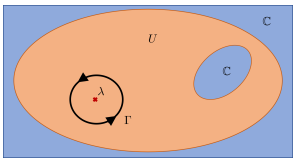


Figure 3. Cauchy Integral (3)

Fundamental Definition: When defining the matrix $h(L)$, the formal replacement $\lambda \mapsto L$ is then made on both sides of the Cauchy formula (3). The path Γ now not only encircles a single value λ but all eigenvalues $\lambda \in \sigma(L)$ in the spectrum of L (c.f. also Fig. 4):

$$g(L) := -\frac{1}{2\pi i} \oint_{\Gamma} g(z) \cdot (L - z \cdot Id)^{-1} dz \tag{4}$$

It can be shown that the precise choice of path Γ is not important (Koke & Cremers, 2024).

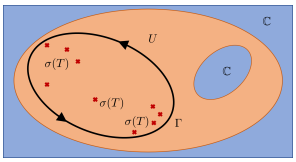


Figure 4. Operator Integral (4)

Compatibility with Algebraic relations This holomorphic functional calculus is compatible with algebraic relations (Kato, 1976): Applying the function $h(\lambda) = \lambda^k$ to L yields L^k and if y is not an eigenvalue of L, applying the function $h(\lambda) = \left(\frac{1}{\lambda - y}\right)^k$ to L yields $h(L) = [(L - y \cdot Id)^{-1}]^k$.

C. An additional perspective on Graph Convolutional Networks:

Learnable filters are combined into a (K-layer) graph convolutional network mapping initial node-features $X \in \mathbb{C}^{N \times F}$ to final representations $X^K \in \mathbb{C}^{N \times F_K}$. With bias matrices $B^\ell \in \mathbb{C}^{N \times F_\ell}$ ($B_{:j} = b_j \cdot \mathbb{1}_G$) and weight matrices $W_i^\ell \in \mathbb{C}^{F_{\ell-1} \times F_\ell}$, layer-updates are then implemented as:

$$X_{i:}^\ell = \rho\left(\sum_{j=1}^{F_{\ell-1}} h_{\theta_{ij}}^\ell(L)(X_{j:}^{\ell-1}) + B_{i:}^\ell\right) \tag{5} \quad \Leftrightarrow \quad X^\ell = \rho\left(\sum_{i \in I} \psi_i(L) \cdot X^{\ell-1} \cdot W_i^\ell + B^\ell\right) \tag{6}$$

Here ρ is a point-wise non-linearity, for which we assume $\rho(0) = 0$ and $|\rho(a) - \rho(b)| \leqslant |a - b|$ ($a, b \in \mathbb{C}$). The connection between the scalar viewpoint (5) and the matrix formulation (6) is given via the identity $h_{\theta_{ij}}(L) \equiv \sum_k (W_k)_{ij} \psi_k(L)$. With the set of basis functions denoted as $\Psi = \{\psi_i\}_{i \in I}$, and weights and biases represented as $\mathscr{W}$ and $\mathscr{B}$, we denote the output of a graph neural network based on the characteristic operator L and applied to the node feature matrix X as $\Phi = \Phi_{\mathscr{W},\mathscr{B},\Psi}(L, X)$.

D. Additional Result I: Stability to Node Level Perturbations

In real world settings, node-features are generically only known up to a certain level of precision. Our first result (proved below) bounds GCN output variations in terms of input-uncertainty.

Theorem D.1. Let $\Phi_{\mathscr{W},\mathscr{B},\Psi}$ be a K-layer GCN. We have that

$$\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L, X) - \Phi_{\mathscr{W},\mathscr{B},\Psi}(L, Y)\| \leqslant \left(\prod_{\ell=1}^{K} C_\ell\right) \cdot \|X - Y\|$$

with $C_\ell = \frac{1}{2\pi} \oint_\Gamma \frac{\sqrt{e}}{d(z,\sigma(L))} \exp\left(\frac{1}{2} \frac{\nu(L)}{d(z,\sigma(L))}\right) \sqrt{\sum_{j \in F_{\ell-1}} \sum_{i \in F_\ell} |h_{\theta_{ij}}^\ell(z)|^2} d|z|$ using (4) and (5). Alternatively, we may set $C_\ell = \sum_{i \in I} \|W_i^\ell\| \cdot \|\psi_i(L)\|$ using the the formulation of (1).

Our estimate of C_ℓ using (4) extends preliminary results in (Koke, 2023) to generic complex differentiable filters and provides an explicit expression for C_ℓ: We see that a failure of L to be unitarily diagonalizable (i.e. $\nu(L) > 0$) negatively influences stability. The smallest stability constants correspond to the undirected setting ($\nu(L) = 0$). We also note that in the formulation (1) the magnitude of weight matrices $W_k \in \mathbb{C}^{F_{\ell-1} \times F_\ell}$ is estimated in spectral norm $\|\cdot\|$ and not – say

– Frobenius norm $\|\cdot\|_F$. This yields reasonable stability constants and allows to retain predictive power even if $F_{\ell-1}, F_\ell \gg 1$.

Hence let us prove the above result:

Proof. Given input signals $X, Y \in \mathbb{C}^{N\times F}$, let us denote the intermediate signal representations in the intermediate layers by $X^\ell, Y^\ell \in \mathbb{C}^{N\times F_\ell}$. With the update rule described in Appendix C, we then have

$$\begin{aligned}
&\|X^{\ell+1} - Y^{\ell+1}\|^2 \\
&= \sum_{i=1}^{F_{\ell+1}} \left\| \rho\left(\sum_{j=1}^{F_n} h_{ij}^{n+1}(L) X_{:j}^\ell\right) - \rho\left(\sum_{j=1}^{F_n} h_{ij}^{n+1}(L) Y_{:j}^\ell\right)\right\|^2 \\
&\leqslant \sum_{i=1}^{F_{\ell+1}} \left\| \sum_{j=1}^{F_\ell} h_{ij}^{n+1}(L) X_{:j}^\ell - \sum_{j=1}^{F_n} h_{ij}^{n+1}(L) Y_{:j}^\ell \right\|^2 \\
&= \sum_{i=1}^{F_{\ell+1}} \left\| \sum_{j=1}^{F_\ell} h_{ij}^{\ell+1}(L) \left[X_{:j}^\ell - Y_{:j}^\ell\right] \right\|^2,
\end{aligned}$$

which follows from $\rho(0) = 0$ and $|\rho(a) - \rho(b)| \leqslant |a-b|$ $(a, b \in \mathbb{C})$. We next note

$$\begin{aligned}
&\sum_{i=1}^{F_{\ell+1}} \left\| \sum_{j=1}^{F_\ell} h_{ij}^{\ell+1}(L) \left[X_{:j}^\ell - Y_{:j}^\ell\right] \right\|^2 \\
&\leqslant \sum_{i=1}^{F_{\ell+1}} \left(\sum_{j=1}^{F_\ell} \|h_{ij}^{\ell+1}(L)\| \cdot \| \left[X_{:j}^\ell - Y_{:j}^\ell\right] \| \right)^2 \\
&\leqslant \left(\sum_{i=1}^{F_{\ell+1}} \sum_{j=1}^{F_\ell} \|h_{ij}^{\ell+1}(L)\|^2 \right) \sum_{j=1}^{F_\ell} \| \left[X_{:j}^\ell - Y_{:j}^\ell\right] \|^2 \\
&= \left(\sum_{i=1}^{F_{\ell+1}} \sum_{j=1}^{F_\ell} \|h_{ij}^{\ell+1}(L)\|^2 \right) \|X^\ell - Y^\ell\|^2
\end{aligned}$$

where the second to last step is an application of the Cauchy Schwarz inequality.

For generic L and holomorphic h, we note

Lemma D.2. For holomorphic g and generic T we have

$$\|h(L)\| \leqslant \frac{1}{2\pi} \oint_\Gamma |h(z)| \frac{\sqrt{e}}{d(z, \sigma(L))} \exp\left(\frac{1}{2} \frac{\nu(L)}{d(z, \sigma(L))}\right) d|z|.$$

Proof. We first note

$$\begin{aligned}
\left\| \frac{1}{2\pi i} \oint_\Gamma h(z) \cdot (zId - L)^{-1} dz \right\| &\leqslant \left\| \frac{1}{2\pi i} \oint_\Gamma h(z) \cdot (zId - L)^{-1} dz \right\| \\
&\leqslant \frac{1}{2\pi} \oint_\Gamma |h(z)| \left\| \cdot (zId - L)^{-1} \right\| d|z|.
\end{aligned}$$

The claim thus follows from (c.f. (Bandtlow, 2004))

$$\|(zId - L)^{-1}\| \leqslant \frac{\sqrt{e}}{d(z, \sigma(L))} \exp\left(\frac{1}{2} \frac{\nu(L)}{d(z, \sigma(L))}\right).$$

□

An application of the triangle inequality together with the above Lemma then yields

$$\left(\sum_{i=1}^{F_{\ell+1}}\sum_{j=1}^{F_\ell}\|h_{ij}^{\ell+1}(L)\|^2\right)^{\frac{1}{2}} \leqslant \frac{1}{2\pi}\oint_\Gamma \frac{\sqrt{e}}{d(z,\sigma(T))}\exp\left(\frac{1}{2}\frac{\nu(T)}{d(z,\sigma(T))}\right)\sqrt{\sum_{j\in F_{\ell-1}}\sum_{i\in F_\ell}|h_{ij}^{\ell+1}(z)|^2 d|z|}.$$

Which hence establishes the characterization of C_ℓ. Iterating through the Layers yields the total claim.

To establish our second characterization of C_ℓ we note

$$\begin{aligned}
&\|X^{\ell+1}-Y^{\ell+1}\|\\
=&\left\|\rho\left(\sum_{i\in I}\psi_i(L)X^\ell W_i^\ell + B^\ell\right)-\rho\left(\sum_{i\in I}\psi_i(L)Y^\ell W_i^\ell + B^\ell\right)\right\|\\
\leqslant&\left\|\sum_{i\in I}\psi_i(L)X^\ell W_i^\ell + B^\ell - \left(\sum_{i\in I}\psi_i(L)Y^\ell W_i^\ell + B^\ell\right)\right\|\\
=&\left\|\sum_{i\in I}\psi_i(L)(X^\ell - Y^\ell)W_i^\ell\right\|\\
\leqslant&\sum_{i\in I}\|\psi_i(L)\|\cdot\|X^\ell-Y^\ell\|\cdot\|W_i^\ell\|\\
=&\left(\sum_{i\in I}\|\psi_i(L)\|\|W_i^\ell\|\right)\cdot\|X^\ell-Y^\ell\|.
\end{aligned}$$

Iterating through the layers then yields the claim.

□

E. Stability to Graph Level Perturbations

Beyond node features, also edge weights of graphs (entering the architecture via L) are generically only known approximately. Stability under variations in these weights is captured by our next result:

Theorem E.1. Let $\Phi_{\mathscr{W},\mathscr{B},\Psi}$ be a K-layer deep graph convolutional architecture. Assume in each layer $1\leqslant\ell\leqslant K$ that $\sum_{i\in I}\|W_i^\ell\|\leqslant W$ and $\|B^\ell\|\leqslant B$. Choose $C\geqslant\|\Psi_i(L)\|$ $(\forall i\in I)$ and w.l.o.g. assume $CW>1$. With this, we have with $\delta=\max_{i\in I}\{\|\Psi_i(L)-\Psi_i(\widetilde{L})\|\}$ that

$$\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X)-\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},X)\|\leqslant\left[K\cdot C^{K-1}W^K\cdot\left(\|X\|+\frac{1}{CW-1}B\right)\right]\cdot\delta.$$

To ease presentation, we here have chosen the stability constant larger than is strictly necessary. The proof below contains additional results (e.g. for $CW\leqslant 1$). Contrary to previous results our result also applies to networks containing biases.

Theorem E.1 reduces the question of stability of entire networks to the question of *single filter stability* of the basis elements ψ_i in $\Psi=\{\psi_i\}_{i\in I}$. In practice, the difference "$\|\psi_i(L)-\psi_i(\widetilde{L})\|$" may of course be evaluated numerically if the basis Ψ is already given.

When *designing* new architectures, it is however important to know in advance how the choice of basis functions affects the stability properties of the network. To this end, bounds of the form $\|\psi_i(L)-\psi_i(\widetilde{L})\|\leqslant C_{\psi_i}\cdot\|L-\widetilde{L}\|$ are desirable. Previous works have derived such bounds for specific classes of filter functions (c.f. e.g. (Koke, 2023)). Here we provide two new useful characterizations (proved in Appendix F) of C_ψ in the most general (potentially directed) setting without assuming specific forms of the underlying filter functions:

Theorem E.2. Let $L,\widetilde{L}$ be characteristic operators. We have $\|\psi(L)-\psi(\widetilde{L})\|\leqslant C_\psi\cdot\|L-\tilde{L}\|$, with $C_\psi=\frac{1}{2\pi}\oint_\Gamma\frac{e}{|z|\cdot d(z,\sigma(L))\cdot d(z,\sigma(\tilde{T}))}\exp\left(\frac{1}{2}\frac{\nu(L)}{d(z,\sigma(L))}+\frac{1}{2}\frac{\nu(\tilde{L})}{d(z,\sigma(\tilde{T}))}\right)|\psi(z)|d|z|$ using Γ as in (4). If $L,\tilde{L}$ are additionally diago-

nalizable, we have with the Frobenius norm denoted by $\|\cdot\|_F$ that $\|\psi(\widetilde{L}) - \psi(L)\| \leqslant \kappa(V_L) \cdot \kappa(V_{\widetilde{L}}) \cdot L_\psi \cdot \|\widetilde{L} - L\|_F$. Here L_{ψ_i} is the Lipschitz constant of ψ_i.

Here we made used of the **condition number** $\kappa(V_L) = \|V_L\| \cdot \|V_L^{-1}\|$ of the change-of-basis matrix V_L (with $\kappa(V_L) = 1$ if V_L is unitary).

For both characterisations of C_ψ, we hence see that the departure of L from being unitarily diagonalizable (either measured via the condition number $\kappa(V_L)$ or the departure from normality $\nu(L)$) increases the stability constant. Thus the same difference $\delta = \|L - \tilde{L}\|$ in characteristic operators has a bigger effect on GCN output variations if the operators $L, \tilde{L}$ correspond to directed graphs.

Hence let us prove Theorem E.1

Proof. For simplicity in notation, let us denote the hidden representations in the network corresponding to $\tilde{L}$ by X^ℓ. With this, we note:

$$\begin{aligned}
\|X^K - \tilde{X}^K\| &\leqslant \sum_{i\in I} \|\psi_i(L) - \psi_i(\tilde{L})\| \cdot \|X^{K-1}\| \cdot \|W_i^K\| + \sum_{i\in I} \|\psi_i(\tilde{L})\| \cdot \|\tilde{X}^{K-1} - X^{K-1}\| \cdot \|W_i^K\| \\
&\leqslant \delta W \|X^{K-1}\| + CW\|\tilde{X}^{K-1} - X^{K-1}\| \\
&\leqslant \delta W \|X^{K-1}\| + CW\delta\|X^{K-2}\| + (CW)^2\|\tilde{X}^{K-1} - X^{K-1}\| \\
&\leqslant \frac{\delta}{C} \cdot \left(\sum_{\ell=1}^{K} (CW)^\ell \|X^{K-\ell}\| \right) \\
&= \frac{\delta}{C} \cdot \left(\sum_{j=0}^{K-1} (CW)^{K-j} \|X^j\| \right)
\end{aligned}$$

Hence we need to bound the quantity $\|X^j\|$ in terms of C, W, B and X.

We have

$$\begin{aligned}
\|X^j\| &\leqslant \sum_i \|\psi_i(L)\| \cdot \|X^{j-1}\| \cdot \|W_i^j|\ + \|B^J\| \\
&\leqslant CW\|X^{j-1}\| + B \\
&\leqslant (CW)^2\|X^{j-2}\| + CWB + B \\
&\leqslant B\left(\sum_{k=0}^{j-1} (CW)^k \right) + (CW)^j\|X\| \\
&= \begin{cases} B\frac{(CW)^j - 1}{CW - 1} + (CW)^j\|X\| & ; CW \neq 1 \\ jB + \|X\| & ; CW = 1 \end{cases}.
\end{aligned}$$

For the case $CW = 1$, we thus find

$$\begin{aligned}
\|X^K - \tilde{X}^K\| &\leqslant \frac{\delta}{C} \cdot \left(\sum_{j=0}^{K-1} (jB + \|X\|) \right) \\
&= \frac{\delta}{C} \cdot \left(K\|X\| + B\frac{K(K-1)}{2} \right).
\end{aligned}$$

For the case $CW \neq 1$, we find

$$\|X^K - \tilde{X}^K\| \leqslant \frac{\delta}{C} \cdot \left(\sum_{j=0}^{K-1} (CW)^{K-j} \left[B\frac{(CW)^j - 1}{CW - 1} + (CW)^j\|X\| \right] \right)$$

For $CW > 1$, we may further estimate this as

$$\begin{aligned}\|X^K - \tilde{X}^K\| &\leqslant \frac{\delta}{C} \cdot \left(\sum_{j=0}^{K-1} (CW)^{K-j} \left[B \frac{(CW)^j - 1}{CW - 1} + (CW)^j \|X\| \right] \right) \\ &\leqslant \delta \cdot \frac{K(CW)^K}{C} \left[\frac{B}{CW-1} + \|X\| \right].\end{aligned}$$

This proves the claim. □

F. Proof of Theorem E.2

In this section, we prove Theorem E.2 stated here again for convenience:

Theorem F.1. Let $L, \widetilde{L}$ be characteristic operators. We have $\|\psi(L) - \psi(\widetilde{L})\| \leqslant C_\psi \cdot \|L - \tilde{L}\|$, with $C_\psi = \frac{1}{2\pi} \oint_\Gamma \frac{e}{|z| \cdot d(z,\sigma(L)) \cdot d(z,\sigma(\tilde{T}))} \exp\left(\frac{1}{2} \frac{\nu(L)}{d(z,\sigma(L))} + \frac{1}{2} \frac{\nu(\tilde{L})}{d(z,\sigma(\tilde{T}))} \right) |\psi(z)| d|z|$. If $T, \tilde{T}$ are additionally diagonalizable, we have with the Frobenius norm denoted by $\|\cdot\|_F$ that $\|\psi(\widetilde{L}) - \psi(L)\| \leqslant \kappa(V_L) \cdot \kappa(V_{\tilde{L}}) \cdot L_\psi \cdot \|\widetilde{L} - L\|_F$. Here L_{ψ_k} is the Lipschitz constant of Ψ_k.

We split this proof into proving two Lemmata:

Lemma F.2. Let $g : \mathbb{C} \to \mathbb{C}$ be Lipschitz continuous with Lipschitz constant D_g. Let X and Y satisfy

$$\begin{aligned}V^{-1} X V &= \text{diag}(\lambda_1, ... \lambda_{d_2}) =: D(X) \\ W^{-1} Y W &= \text{diag}(\mu_1, ... \mu_{d_1}) =: D(Y).\end{aligned}$$

This implies

$$\|g(X) - g(Y)\|_F \leqslant \|V^{-1}\| \|V\| \|W^{-1}\| \|W\| \cdot D_g \cdot \|X - Y\|_F.$$

Proof. This proof builds on the proof idea in (Wihler, 2009). We find:

$$\begin{aligned}\|g(X) - g(Y)\|_F^2 &= ||g(VD(X)V^{-1}) - g(WD(Y)W^{-1})\|_F^2 \\ &= \|Vg(D(X))V^{-1} - Wg(D(Y))W^{-1}\|_F^2 \\ &\leqslant \|V\| \|W^{-1}\| \cdot \|g(D(X))V^{-1}W - V^{-1}Wg(D(Y))\|_F^2 \\ &= \|V\| \|W^{-1}\| \cdot \sum_{i,j} \left| (g(D(X))V^{-1}W - V^{-1}Wg(D(Y)))_{ij} \right|^2 \\ &= \|V\| \|W^{-1}\| \cdot \sum_{i,j} \left| \sum_k [g(D(X))]_{ik} [V^{-1}W]_{kj} - [V^{-1}W]_{ik} [g(D(Y))]_{kj} \right|^2 \\ &= \|V\| \|W^{-1}\| \cdot \sum_{i,j} \left| [V^{-1}W]_{ij} \right|^2 |g(\lambda_j) - g(\mu_i)|^2 \\ &\leqslant \|V\| \|W^{-1}\| \cdot \sum_{i,j} \left| [V^{-1}W]_{ij} \right|^2 D_g^2 |\lambda_j - \mu_i|^2 \\ &= \|V\| \|W^{-1}\| \cdot D_g^2 \|D(X)V^{-1}W - V^{-1}WD(Y)\|_F^2 \\ &\leqslant \|V\| \|V^{-1}\| \|W^{-1}\| \|W\| \cdot D_g^2 \|X - Y\|_F^2.\end{aligned}$$

□

Next we want to prove the following:

Lemma F.3. Let $L, \widetilde{L}$ be operators . With

$$K_g = \frac{1}{2\pi} \oint_\Gamma \frac{1}{|z|} \frac{\sqrt{e}}{d(z,\sigma(L))} \exp\left(\frac{1}{2} \frac{\nu(L)}{d(z,\sigma(L))} \right) \frac{\sqrt{e}}{d(z,\sigma(\tilde{L}))} \exp\left(\frac{1}{2} \frac{\nu(\tilde{L})}{d(z,\sigma(\tilde{L}))} \right) |g(z)| d|z|$$

for g holomorphic, we have

$$\|g(L) - g(\widetilde{L})\| \leqslant K_g \cdot \|L - \widetilde{L}\|$$

Proof. We first note the following:

$$\begin{aligned}
&\frac{1}{\widetilde{L} - z}(\widetilde{L} - L)\frac{1}{L - z} \\
=&\frac{1}{\widetilde{L} - z}\widetilde{L}J\frac{1}{T - z} - \frac{1}{\widetilde{L} - z}L\frac{1}{L - z} \\
=&\left[\frac{1}{\widetilde{L} - z}(\widetilde{L} - z)J + \frac{z}{\widetilde{L} - z}\right]\frac{1}{L - z} - \frac{1}{\widetilde{L} - z}\left[\frac{1}{L - z}(L - z)J + \frac{z}{T - z}\right] \\
=&z\left(\frac{1}{L - z} - \frac{1}{\widetilde{L} - z}J\right).
\end{aligned}$$

Thus we have

$$\begin{aligned}
&\|g(\widetilde{L}) - g(L)\| \leqslant \frac{1}{2\pi}\oint_\Gamma \frac{1}{|z|}\|R_z(L)\|\|R_z(\widetilde{L})\||g(z)|d|z| \\
\leqslant&\frac{1}{2\pi}\oint_\Gamma \frac{1}{|z|}\frac{\sqrt{e}}{d(z, \sigma(L))}\exp\left(\frac{1}{2}\frac{\nu(L)}{d(z, \sigma(L))}\right)\frac{\sqrt{e}}{d(z, \sigma(\tilde{L}))}\exp\left(\frac{1}{2}\frac{\nu(\tilde{L})}{d(z, \sigma(\tilde{L}))}\right)|g(z)|d|z|.
\end{aligned}$$

Here we estimated (using (Bandtlow, 2004))

$$\|(L - zId)^{-1}\| \equiv \|R_z(L)\| \leqslant \frac{\sqrt{e}}{d(z, \sigma(L))}\exp\left(\frac{1}{2}\frac{\nu(L)}{d(z, \sigma(L))}\right).$$

□

G. Comparison of Diffusion Flows for edge-rewiring in K_N

We are interested in establishing that in the setting of Section 3, we have

$$\|e^{-Lt} - e^{-\tilde{L}t}\| \lesssim e^{-(N-2)t}.$$

To this end, we first note that both Laplacians $L, \tilde{L}$ correspond to graphs that are connected. Hence the kernel of both Laplacians is spanned by the vector of $\mathbb{1}$ of all ones. Denote by P the orthogonal projection onto $\mathbb{1}$ and set $Q = Id - P$. We then have

$$\|e^{-Lt} - e^{-\tilde{L}t}\| = \|Qe^{-Lt}Q - Qe^{-\tilde{L}t}Q\|.$$

Next we note for the Laplacian L on K_N that

$$L = N \cdot Q,$$

and hence

$$\|e^{-Lt} - e^{-\tilde{L}t}\| = \|Qe^{-Nt} - Qe^{-\tilde{L}t}Q\|.$$

From perturbation theory, we note that for the eigenvalues of symmetric matrices $A, (A + B)$ ordered in decreasing order, we have (c.f. e.g. (Kato, 1976))

$$|\lambda_i(A + B) - \lambda_i(A)| \leqslant \|B\|.$$

Since $\tilde{L}$ arises from L by deleting a single edge and the Laplacian defined on an unweighted connected two-node graph has operator norm equal to two, we find

$$|\lambda - N| \leqslant 2$$

for any $\lambda \in \sigma(\tilde{L})$. Thus with spectral projection P_λ of $\tilde{L}$, we find

$$\|e^{-Lt} - e^{-\tilde{L}t}\| \leqslant e^{-Nt}\left\|\sum_{0 \neq \lambda \in \sigma(\tilde{L})} Q(1 - e^{(N-\lambda)t}P_\lambda Q\right\| \lesssim e^{-(N-2)t}.$$

H. Further discussion of unidirectional similarity

In the *unidirectional* setting, we can transfer the diffusion process from G to $\tilde{G}$ without producing a large deviation, but not vice versa. Such a setting might e.g. occur if G is a subgraph of $\tilde{G}$ (c.f. the example in Fig. 5 further discussed in Appendix H. *Bidirectional* similarity is a stronger measure of similarity. In this setting, diffusing features on G is approximately the same as first projecting them to $\tilde{G}$ via J,

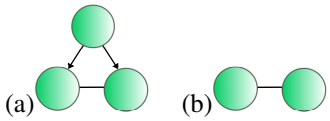


Figure 5. Example of unidirectionally similar graphs

then diffusing on $\tilde{G}$ and finally interpolating back to G with $\tilde{J}$. Since G and $\tilde{G}$ generically have different numbers of nodes, we can not demand $J\tilde{J} = Id_{\tilde{G}}$ while $\tilde{J}J = Id_G$, as at least one of the products $\{J\tilde{J}, \tilde{J}J\}$ can not have full rank. Hence for $t = 0$ we have $\|e^{-Lt} - \tilde{J}e^{-\tilde{L}t}J\|_{t=0} = \|Id_G - \tilde{J}J\| > 0$ irrespective of $L, \tilde{L}$. In this setting, similarity between the two graphs is then measured by how fast the difference between the respective diffusion processes on G and $\tilde{G}$ becomes negligible as diffusion time t increases beyond the initial $t = 0$; i.e. by how fast $\eta(t)$ decays to zero.

Hence let us further discuss the example of unidirectionally similar graphs introduced in Fig. 5. Let us denote the graph of Fig. 5 (a) by $\tilde{G}$ and the graph of Fig. 5 (b) by G. On both these graphs let us consider the out-degree Laplacian

$$L^{\text{out}} := D^{\text{out}} - W$$

as characteristic operator on both G and $\tilde{G}$. Here D^{out} denotes the diagonal out-degree matrix $D^{\text{out}}_{jj} = \sum_i A_{ij}$.

The diffusion process e^{-tL} arises as the solution operator of the differential equation

$$\frac{dx(t)}{dt} = -Lx(t).$$

Using this, we see that no information flows from the 'top' node of $\tilde{G}$ to either of the two bottom nodes in Fig. 5 (a). Chosing as J the obvious inclusion operator mapping from $\tilde{G}$ to G and assigning the value '0' to the top node in $\tilde{G}$, we easily find $\|e^{-tL}J - e^{-t\tilde{L}}J\| = 0$. The diffusion on $\tilde{G}$ (i.e. the graph in Fig. 5 (a)) however is dependent on the top node in $\tilde{G}$ as well if this node carries a non-zero initial value. Hence we can not transfer it to G.

I. Laplace Transform Filters

In this section we provide an overview of the concept of Laplace transforms. We begin with a recapitulation of complex measures.

I.1. Complex measures on $\mathbb{R}_{\geqslant 0}$ and their Theory of Integration

As reference for this section (Tao, 2013) might serve.

In mathematics, a measure is a formal generalization of concepts such as length, area and volume. We are interested in assigning a generalized notion of length (or mass) to subsets of the real half-line

$$\mathbb{R}_{\geqslant 0} = [0, \infty).$$

The set will turn out to be a so called σ-Algebra; i.e. a set Σ of sets for which

- $\varnothing, \mathbb{R}_{\geqslant 0} \in \Sigma$
- $A, B \in \sigma \Rightarrow A \cap B \in \Sigma$
- $A, B \in \Sigma \Rightarrow A \backslash B \in \Sigma$
- $A, B \in \Sigma \Rightarrow A \cup B \in \Sigma$.

We now take $\Sigma_{\mathbb{R}_{\geqslant 0}}$ to be the smallest such set of sets Σ that contains all open intervals.

A complex measure then is a set-function that assigns to each set in $\Sigma_{\mathbb{R}_{\geqslant 0}}$ a complex number in a certain way:

Definition I.1. A complex measure μ on $\mathbb{R}_{\geqslant 0}$ is a complex valued function $\mu : \Sigma_{\mathbb{R}_{\geqslant 0}} \to \mathbb{C}$ satisfying

$$\mu\left(\bigcup_n A_n\right) = \sum_n \mu\left(A_n\right)$$

for any countable (potentially infinite) collection of sets in $\Sigma_{\mathbb{R}_{\geqslant 0}}$ which are pairwise disjoint.

Let us provide some examples:

Example I.2. The prototypical example of a measure is the standard Lebesgue measure that assigns to any interval (a, b) the length $\mu_{\text{Leb}}((a, b)) = |a - b|$ $(a, b \in \mathbb{R}_{\geqslant 0})$.

Example I.3. Alternatively, we might consider the Dirac measure $\mu_{\delta_{t_0}}$, which assigns the value $\mu_{\delta_{t_0}}((a, b)) = 1$ to any interval (a, b) containing t_0 (i.e. $t_0 \in (a, b)$). Otherwise it assigns the value $\mu_{\delta_{t_0}}((a, b)) = 0$ if $t_0 \notin (a, b)$.

Example I.4. Every integrable function $\hat{\psi} : \mathbb{R}_{\geqslant 0} \to \mathbb{C}$ defines a complex measure via $\mu_{\hat{\psi}}((a, b)) = \int_a^b \hat{\psi}(t)dt$.

Any given measure on $\mathbb{R}_{\geqslant 0}$ defines a unique way of integrating (known as Lebesgue integration) a function f defined on $\mathbb{R}_{\geqslant 0}$. This proceeds by approximating any function f via a weighted sequence of indicator functions (with $A \in \Sigma_{\mathbb{R}_{\geqslant 0}}$ a set)

$$\chi_A(t) = \begin{cases} 1 & ; t \in A \\ 0 & ; t \notin A \end{cases}.$$

as

$$f(t) \approx f_n(t) := \sum_k a_k^n \chi_{A_k}(t).$$

with $a_k \in \mathbb{C}$. For these functions, one then sets

$$\int_{\mathbb{R}_{\geqslant 0}} f_n d\mu \equiv \sum_k a_k^n \cdot \mu(A_k).$$

Since we have $\lim_{n\to\infty} f_n = f$, one then simply sets

$$\int_{\mathbb{R}_{\geqslant 0}} f d\mu \equiv \lim_{n\to\infty} \int_{\mathbb{R}_{\geqslant 0}} f_n d\mu.$$

Example I.5. For the prototypical example of the standard Lebesgue measure, this process simply yields

$$\int_{\mathbb{R}_{\geqslant 0}} f(t) d\mu_{\text{Leb}}(t) = \int_0^\infty f(t)dt.$$

Example I.6. For the Dirac measure $\mu_{\delta_{t_0}}$, the above process yields

$$\int_{\mathbb{R}_{\geqslant 0}} f(t) d\mu_{\delta_{t_0}}(t) = f(t_0)$$

Example I.7. For measures arising from integrable functions $\hat{\psi} : \mathbb{R}_{\geqslant 0} \to \mathbb{C}$ as $\mu_{\hat{\psi}}((a, b)) = \int_a^b \hat{\psi}(t)dt$, we find

$$\int_{\mathbb{R}_{\geqslant 0}} f(t) d\mu_{\hat{\psi}} = \int_0^\infty \hat{\psi}(t) f(t) dt.$$

I.2. Laplace Transforms

We say complex valued measure μ is finite if we have

$$\int_{\mathbb{R}_{\geqslant 0}} d|\mu|(t) < \infty.$$

Here the measure $|\mu|$ arises from the original measure μ via

$$|\mu|((a,b)) \equiv |\mu((a,b))|.$$

For any such finite measure μ we may define its Laplace transform as

$$\psi_\mu(z) := \int_{\mathbb{R}_{\geqslant 0}} e^{-tz} d\mu(t).$$

This function f_μ is well defined for z in the right hemisphere

$$\mathbb{C}_R := \{z \in \mathbb{C} : \mathrm{Re}(z) \geqslant 0\}.$$

of the complex plane $\mathbb{C}$, since there we have

$$\begin{aligned} |\psi_\mu(z)| &= \left| \int_{\mathbb{R}_{\geqslant 0}} e^{-tz} d\mu(t) \right| \\ &\leqslant \int_{\mathbb{R}_{\geqslant 0}} |e^{-tz}| d|\mu|(t) \\ &\leqslant \int_{\mathbb{R}_{\geqslant 0}} d|\mu|(t) < \infty. \end{aligned}$$

Example I.8. For the Dirac measure $\mu_{\delta_{t_0}}$, we have

$$\psi_{\mu_{\delta_{t_0}}}(z) = e^{-t_0 z}.$$

Example I.9. For any integrable function $\hat{\psi}$, we have

$$\psi(z) \equiv \int_{\mathbb{R}_{\geqslant 0}} e^{-tz} d\mu_{\hat{\psi}} = \int_0^\infty \hat{\psi}(t) e^{-tz} dt.$$

More specifically, if the integrable function is given as $\hat{\psi}_k := (-t)^{k-1} e^{-\lambda t}$ (with $\mathrm{Re}(\lambda) > 0$), then $\psi_k(z) = (z+\lambda)^{-k}$:
Example I.10. If $\hat{\psi}_k := (-t)^{k-1} e^{-\lambda t}$ yields $\psi_k(z) = (z+\lambda)^{-k}$, then

$$\psi_k(z) = (z+\lambda)^{-k}.$$

For $k = 1$, this can be seen from

$$\int_0^\infty e^{-tz} e^{-\lambda t} dt = -\frac{1}{z+\lambda} e^{-(z+\lambda)} \Big|_0^\infty.$$

For $k > 1$, the claim follows from differentiating the above expression with respect to z Note that the functions $\psi_k(z) = (z+\lambda)^{-k}$ are also defined if $\mathrm{Re}(z) \leqslant 0$, as long as $z \neq -\lambda$.

Using the function ψ_k of the examples above, a wide class of functions may be parametrized
Theorem I.11. Let $f : \mathbb{R}_{\geqslant 0} \to 0$ be any function with $\lim_{x\to\infty} f(x) = 0$. Then for any $\epsilon > 0$, there is a function

$$h(x) = \sum_k \theta_k \psi_k(x)$$

for which

$$\sup_{x \in [0,\infty)} |f(x) - h(x)| < \epsilon.$$

Here the basis functions $\{\psi_k\}$ may either be chosen as $\psi_k(z) = (z+\lambda)^{-k}$ or $\psi_k(x) = e^{-(kt_0)x}$ for any $t_0 > 0$.

Proof. This is a direct consequence of the Weierstrass approximation theorem. □

I.3. Proof of Theorem 3.5

In this section, we prove Theorem 3.5, which we restate here for convenience:

Theorem I.12. We have $\|J\psi(L) - \psi(\tilde{L})J\| \leqslant \|\hat{\psi}\|_1 \cdot \sup_{t\geqslant 0} \|Je^{-Lt} - e^{-\tilde{L}t}J\|$ in the *unidirectional* setting. In the *bidirectional* setting $\|\psi(L) - \tilde{J}\psi(\tilde{L})J\| \leqslant \int_0^\infty |\hat{\psi}(t)|\eta(t)dt$ holds true.

Proof. We start by proving the first claim. To this end, we note

$$\begin{aligned}\|J\psi(L) - \psi(\tilde{L})J\| &= \left\| \int_{\mathbb{R}_{\geqslant 0}} \left[Je^{-tL} - e^{-t\tilde{L}}J \right] d\mu_{\hat{\psi}} \right\| \\ &\leqslant \int_{\mathbb{R}_{\geqslant 0}} \left\| \left[Je^{-tL} - e^{-t\tilde{L}}J \right] \right\| d|\mu|_{\hat{\psi}} \\ &\leqslant \sup_{t\geqslant 0} \|Je^{-Lt} - e^{-\tilde{L}t}J\| \cdot \int_{\mathbb{R}_{\geqslant 0}} d|\mu|_{\hat{\psi}}\end{aligned}$$

Observing that in the notation of Section 3 we precisely have

$$\|\hat{\psi}\|_1 \equiv \int_{\mathbb{R}_{\geqslant 0}} d|\mu|_{\hat{\psi}}$$

the claim follows.
Proceeding as above, we note

$$\|\psi(L) - \tilde{J}\psi(\tilde{L})J\| \leqslant \int_0^\infty \left\| \left[e^{-tL} - \tilde{J}e^{-t\tilde{L}}J \right] \right\| d|\mu|_{\hat{\psi}},$$

from which the second claim follow.

□

I.4. Proof of Corollary 3.6

Here we prove Corollary 3.6; restated here for convenience:

Corollary I.13. Consider a sequence of graphs G_n for which $\|e^{-L_n t} - \tilde{J}_n e^{-\tilde{L}t} J_n\| \to 0$. Then for a Laplace transform filter ψ, we have $\|\psi(L_n) - \tilde{J}_n \psi(\tilde{L}) J_n\|| \to 0$ if and only if $\lim_{r\to\infty} \psi(r) = 0$.

Proof. Let us first prove that the condition is sufficient. To this end assume that $\lim_{r\to\infty} \psi(r) = 0$. This implies that $\mu_{\hat{\psi}}(\{0\}) = 0$. Hence we have

$$\begin{aligned}\|\psi(L_n) - \tilde{J}_n\psi(\tilde{L})J_n\| &= \left\| \int_0^\infty \left[e^{-Lt} - \tilde{J}e^{-\tilde{L}t}J \right] d\mu_{\hat{\psi}}(t) \right\| \\ &\leqslant \int_0^\infty \left\| e^{-Lt} - \tilde{J}e^{-\tilde{L}t}J \right\| d|\mu|_{\hat{\psi}}(t)\end{aligned}$$

The integrand $\left\| e^{-Lt} - \tilde{J}e^{-\tilde{L}t}J \right\|$ converges to zero everywhere except on a set of measure zero (i.e. the set $\{t|t=0\} = \{0\}$). The dominated convergence theorem then yields the claim. □

I.5. Proof of Theorem 3.7 and Corollary 3.8

We begin by proving Theorem 3.7; restated here for convenience:

Theorem I.14. Let ψ be a Laplace transform filter. There exists a constant $C = C_{\psi,\nu(L),\nu(\tilde{L})} < \infty$ so that we have $\|J\psi(L) - \psi(\tilde{L})J\| \leqslant C \cdot \|J(L + \lambda Id)^{-1} - (\tilde{L} + \lambda\tilde{Id})^{-1}J\|$.

Proof. We make use of the characterization (4)

$$\psi(L) := -\frac{1}{2\pi i}\oint_{\Gamma}\psi(z)\cdot(L-z\cdot Id)^{-1}dz$$

to arrive at

$$\|J\psi(L)-\psi(\tilde{L})J\| \leqslant \frac{1}{2\pi}\oint_{\Gamma}|\psi(z)|\cdot\|J(L-zId)^{-1}-(\tilde{L}-zId)^{-1}J\|d|z|.$$

Combining results of (Post, 2012) and (Bandtlow, 2004) yields

$$\begin{aligned}&\|J(L-zId)^{-1}-(\tilde{L}-zId)^{-1}J\|\\ \leqslant&\left(1+|\lambda+z|\frac{\sqrt{e}}{d(z,\sigma(\tilde{L}))}\exp\left(\frac{1}{2}\frac{\nu(\tilde{L})}{d(z,\sigma(\tilde{L}))}\right)\right)\cdot\left(1+|\lambda+z|\frac{\sqrt{e}}{d(z,\sigma(\tilde{L}))}\exp\left(\frac{1}{2}\frac{\nu(\tilde{L})}{d(z,\sigma(\tilde{L}))}\right)\right)\\ &\times\|J(L+\lambda Id)^{-1}-(\tilde{L}+\lambda Id)^{-1}J\|.\end{aligned}$$

Hence we may set

$$C=\frac{1}{2\pi}\oint_{\Gamma}|\psi(z)|\cdot p_{\nu(L),\nu(\tilde{L})}(z)d|z|$$

with

$$\begin{aligned}&p_{\nu(L),\nu(\tilde{L})}(z)\\ \equiv&\left(1+|\lambda+z|\frac{\sqrt{e}}{d(z,\sigma(\tilde{L}))}\exp\left(\frac{1}{2}\frac{\nu(\tilde{L})}{d(z,\sigma(\tilde{L}))}\right)\right)\cdot\left(1+|\lambda+z|\frac{\sqrt{e}}{d(z,\sigma(\tilde{L}))}\exp\left(\frac{1}{2}\frac{\nu(\tilde{L})}{d(z,\sigma(\tilde{L}))}\right)\right)\end{aligned}$$

□

Next we prove Theorem 3.8; restated below.

Theorem I.15. Consider a graph sequence G_n with $\|(L_n+\lambda Id)^{-1}-\tilde{J}_n(\tilde{L}+\lambda Id)^{-1}J_n\|\to 0$. If the graphs are directed, assume eigenvalues of all L_ns lie within a cone of opening angle $\alpha<\pi$ symmetric about the real axis. Then we have $\|\psi(L_n)-\tilde{J}_n\psi(\tilde{L})J_n\|\to 0$ if and only if $\lim_{r\to\infty}\psi(r)=0$.

Proof. As in the proof of Theorem 3.7 above, we arrive at

$$\|\psi(L)-\tilde{J}\psi(\tilde{L})J\| \leqslant \frac{1}{2\pi}\oint_{\Gamma}|\psi(z)|\cdot\|(L-zId)^{-1}-\tilde{J}(\tilde{L}-zId)^{-1}J\|d|z|.$$

Since $\|(L_n+\lambda Id)^{-1}-\tilde{J}_n(\tilde{L}+\lambda Id)^{-1}J_n\|\to 0$ implies $\|(L_n-zId)^{-1}-\tilde{J}_n(\tilde{L}-zId)^{-1}J_n\|\to 0$ uniformly (in z) on compact sets (c.f. e.g. (Arendt, 2001)), we can apply dominated convergence as in the proof of Corollary 3.6 in Appendix I.4; if we find an majorizing function that is integrable on Γ. But this is ensured by the decay of ψ and the possibility to choose Γ to lie within in a cone of opening angle $\alpha\lneq\pi$ about the real axis of opening angle less than π. □

I.6. Discussion of extension beyond spectral assumptions

Above, we have assumed that all appearing eigenvalues $\lambda\in\mathbb{C}$ in the spectrum $\sigma(L)$ have real part $\mathrm{Re}(\lambda)\geqslant 0$. This guarantees that

$$\limsup_{t\to\infty}\|e^{-Lt}\|<\infty.$$

From this we find that

$$\|\psi(L)\|=\left\|\int_{\mathbb{R}_{\geqslant 0}}e^{-tL}d\mu(t)\right\|\leqslant\left(\limsup_{t\to\infty}\|e^{-Lt}\|\right)\cdot\int_{\mathbb{R}_{\geqslant 0}}d|\mu|(t)<\infty,$$

so that the filter $\psi(L)$ is indeed well-defined. If we want to allow $\mathrm{Re}(\lambda)<0$ as well, we have two options:

The set $\{\mathbf{Re}(\lambda)\}$ is bounded from below: In this setting we have a guarantee that there is $c_- > 0$ so that for all appearing eigenvalues in the spectra of L and $\tilde{L}$ we have

$$-c_- \leqslant \mathrm{Re}(\lambda).$$

This implies that

$$\limsup_{t\to\infty} \|e^{-Lt}e^{-c_-t}\| < \infty.$$

Using

$$\left\|\int_{\mathbb{R}_{\geqslant 0}} e^{-tL}d\mu(t)\right\| = \left\|\int_{\mathbb{R}_{\geqslant 0}} e^{-tL}e^{-c_-t}e^{c_-t}d\mu(t)\right\|$$
$$\leqslant \left(\limsup_{t\to\infty} \|e^{-Lt}e^{-c_-t}\|\right)\cdot \int_{\mathbb{R}_{\geqslant 0}} e^{c_-t}d|\mu|(t),$$

the developed theory above is still applicable in this setting, as long as we assume that the measure μ defining the Laplace transform filter ψ satisfies

$$\int_{\mathbb{R}_{\geqslant 0}} e^{c_-t}d|\mu|(t) < \infty.$$

Note that this is stronger than the demand

$$\int_{\mathbb{R}_{\geqslant 0}} d|\mu|(t) < \infty.$$

made in Definition 3.2.

The set $\{\mathbf{Re}(\lambda)\}$ is not bounded from below: In this setting, we pick a $\mu \in \mathbb{C}$ with $\mathrm{Re}(\mu) < 0$ and $\mu \notin \sigma(L) \cup \sigma(\tilde{L})$. We then restrict the class of filters to those determined by Example 3.4: There we chose $\hat{\psi}_k := (-t)^{k-1}e^{-\mu t}$, which yielded filters of the form $\{h_\theta(\cdot) := \sum_i \theta_i \cdot \psi_i(\cdot)\}$, with $\psi_k(L) = \left[(L+\mu Id)^{-1}\right]^k$. Such filters hence remain defined as long as $\mu \notin \sigma(L)$.

I.7. Proof of Theorem 3.9

Theorem I.16. Let $\Phi_{\mathscr{W},\mathscr{B},\Psi}$ be an L-layer deep GCN. Assume that $\sum_k \|W_k^\ell\| \leqslant W$ and $\|B^\ell\| \leqslant B$. Choose $C \geqslant \|\Psi_k(T)\|$ $(\forall k)$ and w.l.o.g. assume $CW > 1$. Assume $\rho(J\tilde{X}) = J\rho(\tilde{X})$ and if biases are enabled, assume $J\mathbb{1}_G = \mathbb{1}_{\tilde{G}}$. With this, we have with $\delta = \max_{i\in I}\{\|J\psi_i(L) - \psi_i(\tilde{L})J\|\}$ that

$$\|J\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \Phi_{\mathscr{W},\mathscr{B},\Psi}(\tilde{L},JX)\| \leqslant \left[K\cdot C^K W^{K-1}\cdot\left(\|X\| + \frac{1}{CW-1}B\right)\right]\cdot\delta.$$

Proof. Let us define

$$\tilde{X} := JX.$$

Let us further use the notation $\tilde{\psi}_i := \psi_i(\tilde{L})$ and $\psi_i := \psi_i(L)$.

Denote by X^ℓ and $\tilde{X}^\ell$ the (hidden) feature matrices generated in layer ℓ for networks based on ψ_i and $\tilde{\psi}_i$ respectively: I.e. we have

$$X^\ell = \rho\left(\sum_{i\in I}\psi_i X^{\ell-1}W_i^\ell + B^\ell\right)$$

and

$$\tilde{X}^\ell = \rho\left(\sum_{i\in I}\tilde{\psi}_i \tilde{X}^{\ell-1}W_i^\ell + \tilde{B}^\ell\right).$$

We then have

$$
\begin{aligned}
&\|J\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X)-\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\|\\
=&\|JX^K-\widetilde{X}^K\|\\
=&\left\|J\rho\left(\sum_{i\in I}\psi_i X^{L-1}W_i^K+B^K\right)-\rho\left(\sum_{i\in I}\tilde{\psi}_i\widetilde{X}^{K-1}W_i^K+\tilde{B}^L\right)\right\|\\
=&\left\|\rho\left(J\sum_{i\in I}\psi_i X^{L-1}W_i^K+\tilde{B}^K\right)-\rho\left(\sum_{i\in I}\tilde{\psi}_i\widetilde{X}^{K-1}W_i^K+B^L\right)\right\|
\end{aligned}
$$

Here we used the assumption that ρ and J commute. We also made use of the assumption $J\mathbb{1}_G=\mathbb{1}_{\tilde{G}}$ when dealing with biases .
Using the fact that $\rho(\cdot)$ is 1-Lipschitz-continuous (c.f. Section C), we can establish

$$
\begin{aligned}
&\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X)-\tilde{J}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\|\\
\leqslant&\left\|\left(J\sum_{i\in I}\psi_i X^{L-1}W_i^K+\tilde{B}^K\right)-\left(\sum_{i\in I}\tilde{\psi}_i\widetilde{X}^{K-1}W_i^K+\tilde{B}^K\right)\right\|.
\end{aligned}
$$

We then have

$$
\begin{aligned}
&\|J\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X)-\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\|\\
\leqslant&\left\|\sum_{i\in I}J\psi_i X^{K-1}W_i^K-\sum_{i\in I}\tilde{\psi}_i\widetilde{X}^{K-1}W_i^K\right\|.
\end{aligned}
$$

From this, we find (inserting a zero), that

$$
\begin{aligned}
&\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X)-\tilde{J}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\|\\
\leqslant&\left\|\sum_{i\in I}J\psi_i X^{K-1}W_i^K-\sum_{i\in I}\tilde{\psi}_i\widetilde{X}^{K-1}W_i^K\right\|\\
\leqslant&\left\|\sum_{i\in I}(J\psi_i-\tilde{\psi}_i J)X^{K-1}W_i^K\right\|+\sum_{i\in I}\|\tilde{\psi}_i\|\cdot\|\widetilde{X}^{K-1}-JX^{K-1}\|\cdot\|W_i^K\|\\
\leqslant&\left\|\sum_{i\in I}(J\psi_i-\tilde{\psi}_i J)X^{K-1}W_i^K\right\|+CW\cdot\|\widetilde{X}^{K-1}-JX^{K-1}\|\\
\leqslant&\sum_{i\in I}\left\|(J\psi_i-J\tilde{\psi}_i J)\right\|\cdot\left\|X^{K-1}\right\|\cdot\left\|W_i^K\right\|+CW\cdot\|\widetilde{X}^{K-1}-JX^{K-1}\|\\
\leqslant&\sum_{i\in I}\delta\cdot\left\|X^{K-1}\right\|W+CW\cdot\|\tilde{J}\widetilde{X}^{K-1}-X^{K-1}\|
\end{aligned}
$$

Arguing as in the proof of Theorem E.1 in Appendix E then yields the claim.

□

For the bidirectional setting we find the following:

Theorem I.17. Let $\Phi_{\mathscr{W},\mathscr{B},\Psi}$ be an L-layer deep GCN. Assume that $\sum_k\|W_k^\ell\|\leqslant W$ and $\|B^\ell\|\leqslant B$. Choose $C\geqslant\|\Psi_k(T)\|$ $(\forall k)$ and w.l.o.g. assume $CW>1$. Assume $\rho(\tilde{J}X)=\tilde{J}\rho(X)$ and if biases are enabled, assume $\tilde{J}\mathbb{1}_{\tilde{G}}=\mathbb{1}_G$. Further assume $J\tilde{J}=Id_{\tilde{G}}$ With this, we have with $\delta=\max_{i\in I}\{\|\psi_i(L)-\tilde{J}\psi_i(\widetilde{L})J\|\}$ that

$$
\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X)-\tilde{J}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\|\leqslant\left[K\cdot C^K W^{K-1}\cdot\left(\|X\|+\frac{1}{CW-1}B\right)\right]\cdot\delta.
$$

Proof. Let us define

$$\tilde{X} := JX.$$

Let us further use the notation $\tilde{\psi}_i := \psi_i(\tilde{L})$ and $\psi_i := \psi_i(L)$.

Denote by X^ℓ and $\widetilde{X}^\ell$ the (hidden) feature matrices generated in layer ℓ for networks based on ψ_i and $\tilde{\psi}_i$ respectively: I.e. we have

$$X^\ell = \rho\left(\sum_{i\in I} \psi_i X^{\ell-1} W_i^\ell + B^\ell\right)$$

and

$$\widetilde{X}^\ell = \rho\left(\sum_{i\in I} \tilde{\psi}_i \widetilde{X}^{\ell-1} W_i^\ell + \tilde{B}^\ell\right).$$

We then have

$$\begin{aligned}
&\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \tilde{J}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\| \\
=&\|X^K - \tilde{J}\widetilde{X}^K\| \\
=&\left\|\rho\left(\sum_{i\in I}\psi_i X^{L-1}W_i^K + B^K\right) - \tilde{J}\rho\left(\sum_{i\in I}\tilde{\psi}_i\widetilde{X}^{K-1}W_i^K + \tilde{B}^L\right)\right\| \\
=&\left\|\rho\left(\sum_{i\in I}\psi_i X^{L-1}W_i^K + B^K\right) - \rho\left(\tilde{J}\sum_{i\in I}\tilde{\psi}_i\widetilde{X}^{K-1}W_i^K + B^L\right)\right\|
\end{aligned}$$

Here we used the assumption that ρ and $\tilde{J}$ commute. fact that since ReLU$(\cdot)$ maps positive entries to positive entries and acts pointwise, it commutes with $J^\uparrow$. We also made use of the assumption $\tilde{J}\mathbb{1}_{\tilde{G}} = \mathbb{1}_G$ when dealing with biases .
Using the fact that $\rho(\cdot)$ is 1-Lipschitz-continuous (c.f. Section C), we can establish

$$\begin{aligned}
&\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \tilde{J}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\| \\
\leqslant&\left\|\rho\left(\sum_{i\in I}\psi_i X^{L-1}W_i^K + B^K\right) - \rho\left(\tilde{J}\sum_{i\in I}\tilde{\psi}_i\widetilde{X}^{K-1}W_i^K + B^L\right)\right\|.
\end{aligned}$$

Using the assumption that that $J\tilde{J} = Id_{\tilde{G}}$, we have

$$\begin{aligned}
&\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \tilde{J}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\| \\
\leqslant&\left\|\sum_{i\in I}\psi_i X^{K-1}W_i^K - \sum_{i\in I}(\tilde{J}\tilde{\psi}_i J)\tilde{J}\widetilde{X}^{K-1}W_i^K\right\|.
\end{aligned}$$

From this, we find (assuming $\|\tilde{J}\|, \|J\| \leqslant 1$), that

$$
\begin{aligned}
&\|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \tilde{J}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\| \\
\leqslant &\left\|\sum_{i\in I}\psi_i X^{K-1}W_i^K - \sum_{i\in I}(\tilde{J}\tilde{\psi}_i J)\tilde{J}\widetilde{X}^{K-1}W_i^K\right\| \\
\leqslant &\left\|\sum_{i\in I}(\psi_i - \tilde{J}\tilde{\psi}_i J)X^{K-1}W_i^K\right\| + \sum_{i\in I}\|\tilde{J}\tilde{\psi}_i J\|\cdot\|\tilde{J}\widetilde{X}^{K-1} - X^{K-1}\|\cdot\|W_i^K\| \\
\leqslant &\left\|\sum_{i\in I}(\psi_i - \tilde{J}\tilde{\psi}_i J)X^{K-1}W_i^K\right\| + CW\cdot\|\tilde{J}\widetilde{X}^{K-1} - X^{K-1}\| \\
\leqslant &\sum_{i\in I}\left\|(\psi_i - \tilde{J}\tilde{\psi}_i J)\right\|\cdot\left\|X^{K-1}\right\|\cdot\left\|W_i^K\right\| + CW\cdot\|\tilde{J}\widetilde{X}^{K-1} - X^{K-1}\| \\
\leqslant &\sum_{i\in I}\delta\cdot\left\|X^{K-1}\right\|W + CW\cdot\|\tilde{J}\widetilde{X}^{K-1} - X^{K-1}\|
\end{aligned}
$$

Arguing as in the proof of Theorem E.1 in Appendix E then yields the claim.

□

I.8. Graph Level Transferability

Aggegating node features $X \in \mathbb{C}^{N\times F}$ to graph-level features $\Omega(X) \in \mathbb{C}^F$ via $\Omega(X)_j = \sum_{i=1}^N |X_{ij}|\cdot\mu_i$ for graph level property prediction, we have :

Theorem I.18. Assuming $\Omega(JX) = \Omega(X)$, we have in the setting of Theorem 3.9 that
$\|\Omega\circ\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \Omega\circ\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\| \leqslant \|J\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\|$.

The assumption $\Omega(JX) = \Omega(X)$ clearly need only be satisfied on the potential output of the node-level network Φ (which might e.g. be limited to tensors with positive entries). Such a consistency assumption is for example satisfied when coarse graining graphs.

Let us now prove Theorem I.18:

Proof. We note

$$
\begin{aligned}
&\|\Omega\circ\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \Omega\circ\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\| \\
=&\|\Omega(\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X)) - \Omega(\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX))\| \\
=&\|\Omega(J\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X)) - \Omega(\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX))\|.
\end{aligned}
$$

To prove the claim from here, we only have to note that the aggregation method Ω is 1-Lipschitz (as a consequence of the reverse triangle inequality). □

A similar proof shows the following for the bidirectional setting:

Theorem I.19. Assuming $\Omega(X) = \Omega(\tilde{J}X)$, we have in the setting of Theorem I.17 that
$\|\Omega\circ\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \Omega\circ\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\| \leqslant \|\Phi_{\mathscr{W},\mathscr{B},\Psi}(L,X) - \tilde{J}\Phi_{\mathscr{W},\mathscr{B},\Psi}(\widetilde{L},JX)\|$.

J. Further Discussion for Examples of Transferability Settings

J.1. Further Discussion of the example of Coarse-Graining Graphs

In this appendix, we prove (2):

$$\|(\Delta + Id)^{-1} - J^{\uparrow}(\underline{\Delta} + Id)^{-1}J^{\downarrow}\| \lesssim 1/\lambda_1(\Delta_{\text{high}}).$$

For convenience, we restate the definitions leading up to this result again:

Definition J.1. Denote by $\underline{\mathcal{G}}$ the set of connected components in G_{high}. We give this set a graph structure as follows: Let R and P be elements of $\underline{\mathcal{G}}$ (i.e. connected components in G_{high}). We define the real number

$$\underline{W}_{RP} = \sum_{r\in R}\sum_{p\in P} W_{rp},$$

with r and p nodes in the original graph G. We define the set of edges $\underline{\mathcal{E}}$ on $\underline{G}$ as

$$\underline{\mathcal{E}} = \{(R,P) \in \underline{\mathcal{G}} \times \underline{\mathcal{G}} : \underline{W}_{RP} > 0\}$$

and assign $\underline{W}_{RP}$ as weight to such edges. Node weights of limit nodes are defined similarly as aggregated weights of all nodes r (in G) contained in the component R as

$$\underline{\mu}_R = \sum_{r\in R} \mu_r.$$

In order to translate signals between the original graph G and the limit description $\underline{G}$, we need translation operators mapping signals from one graph to the other:

Definition J.2. Denote by $\mathbb{1}_R$ the vector that has 1 as entries on nodes r belonging to the connected (in G_{high}) component R and has entry zero for all nodes not in R. We define the down-projection operator $J^{\downarrow}$ component-wise via evaluating at node R in $\underline{\mathcal{G}}$ as

$$(J^{\downarrow}x)_R = \langle \mathbb{1}_R, x\rangle / \underline{\mu}_R.$$

The upsampling operator $J^{\uparrow}$ is defined as

$$J^{\uparrow}u = \sum_R u_R \cdot \mathbb{1}_R;$$

where u_R is a scalar value (the component entry of u at $R \in \underline{\mathcal{G}}$) and the sum is taken over all connected components in G_{high}.

The result we then prove is the following:

Theorem J.3. We have

$$\left\|R_z(\Delta) - J^{\uparrow}R_z(\underline{\Delta})J^{\downarrow}\right\| = \mathcal{O}\left(\frac{\|\Delta_{\text{reg.}}\|}{\lambda_1(\Delta_{\text{high}})}\right)$$

holds; with $\lambda_1(\Delta_{\text{high}})$ denoting the first non-zero eigenvalue of Δ_{high}.

$$\lambda_{\max}(\Delta_{\text{reg.}}) = \|\Delta_{\text{reg.}}\|.$$

Proof. We will split the proof of this result into multiple steps. For $z < 0$ Let us denote by

$$\begin{aligned} R_z(\Delta) &= (\Delta - zId)^{-1}, \\ R_z(\Delta_{high}) &= (\Delta_{high} - zId)^{-1} \\ R_z(\Delta_{reg.}) &= (\Delta_{reg.} - zId)^{-1} \end{aligned}$$

the resolvents correspodning to Δ, Δ_{high} and $\Delta_{reg.}$ respectively.
Our first goal is establishing that we may write

$$R_z(\Delta) = \left[Id + R_z(\Delta_{high})\Delta_{reg.}\right]^{-1} \cdot R_z(\Delta_{high})$$

This will follow as a consequence of what is called the second resolvent formula (Teschl, 2014):

"Given self-adjoint operators A, B, we may write

$$R_z(A+B) - R_z(A) = -R_z(A)BR_z(A+B).$$"

In our case, this translates to

$$R_z(\Delta) - R_z(\Delta_{high}) = -R_z(\Delta_{high})\Delta_{\text{reg.}}R_z(\Delta)$$

or equivalently

$$\left[Id + R_z(\Delta_{high})\Delta_{\text{reg.}}\right] R_z(\Delta) = R_z(\Delta_{high}).$$

Multiplying with $\left[Id + R_z(\Delta_{high})\Delta_{\text{reg.}}\right]^{-1}$ from the left then yields

$$R_z(\Delta) = \left[Id + R_z(\Delta_{high})\Delta_{reg.}\right]^{-1} \cdot R_z(\Delta_{high})$$

as desired.
Hence we need to establish that $\left[Id + R_z(\Delta_{high})\Delta_{reg.}\right]$ is invertible for $z < 0$.

To establish a contradiction, assume it is not invertible. Then there is a signal x such that

$$\left[Id + R_z(\Delta_{high})\Delta_{reg.}\right] x = 0.$$

Multiplying with $(\Delta_{\text{high}} - zId)$ from the left yields

$$(\Delta_{\text{high}} + \Delta_{\text{reg.}} - zId)x = 0$$

which is precisely to say that

$$(\Delta - zId)x = 0$$

But since Δ is a graph Laplacian, it only has non-negative eigenvalues. Hence we have reached our contradiction and established

$$R_z(\Delta) = \left[Id + R_z(\Delta_{high})\Delta_{reg.}\right]^{-1} R_z(\Delta_{high}).$$

Our next step is to establish that

$$R_z(\Delta_{high}) \to \frac{P_0^{\text{high}}}{-z},$$

where P_0^{high} is the spectral projection onto the eigenspace corresponding to the lowest lying eigenvalue $\lambda_0(\Delta_{high}) = 0$ of Δ_{high}. Indeed, by the spectral theorem for finite dimensional operators (c.f. e.g. (Teschl, 2014)), we may write

$$R_z(\Delta_{high}) \equiv (\Delta_{high} - zId)^{-1} = \sum_{\lambda \in \sigma(\Delta_{high})} \frac{1}{\lambda - z} \cdot P_\lambda^{high}.$$

Here $\sigma(\Delta_{high})$ denotes the spectrum (i.e. the collection of eigenvalues) of Δ_{high} and the $\{P_\lambda^{high}\}_{\lambda \in \sigma(\Delta_{high})}$ are the corresponding (orthogonal) eigenprojections onto the eigenspaces of the respective eigenvalues. Thus we find

$$\left\| R_z(\Delta_{high}) - \frac{P_0^{high}}{-z} \right\| = \left\| \sum_{0 < \lambda \in \sigma(\Delta_{high})} \frac{1}{\lambda - z} \cdot P_\lambda^{high} \right\|;$$

where the sum on the right hand side now excludes the eigenvalue $\lambda = 0$.

Using orthonormality of the spectral projections, the fact that $z < 0$ and monotonicity of $1/(\cdot + |z|)$ we find

$$\left\| R_z(\Delta_{high}) - \frac{P_0^{high}}{-z} \right\| = \frac{1}{\lambda_1(\Delta_{high}) + |z|}.$$

Here $\lambda_1(\Delta_{high})$ is the firt non-zero eigenvalue of (Δ_{high}).
Non-zero eigenvalues scale linearly with the weight scale since we have

$$\lambda(S \cdot \Delta) = S \cdot \lambda(\Delta)$$

for any graph Laplacian (in fact any matrix) Δ with eigenvalue λ. Thus we have

$$\left\| R_z(\Delta_{high}) - \frac{P_0^{high}}{-z} \right\| = \frac{1}{\lambda_1(\Delta_{high}) + |z|} \leqslant \frac{1}{\lambda_1(\Delta_{high})} \longrightarrow 0$$

as $\lambda_1(\Delta_{high}) \to \infty$.

Our next task is to use this result in order to bound the difference

$$I := \left\| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} \frac{P_0^{high}}{-z} - \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} R_z(\Delta_{high}) \right\|.$$

To this end we first note that the relation

$$[A + B - zId]^{-1} = [Id + R_z(A)B]^{-1} R_z(A)$$

provided to us by the second resolvent formula, implies

$$[Id + R_z(A)B]^{-1} = Id - B[A + B - zId]^{-1}.$$

Thus we have

$$\begin{aligned} \left\| \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} \right\| &\leqslant 1 + \|\Delta_{\text{reg.}}\| \cdot \|R_z(\Delta)\| \\ &\leqslant 1 + \frac{\|\Delta_{\text{reg.}}\|}{|z|}. \end{aligned}$$

With this, we have

$$\begin{aligned} &\left\| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} \cdot \frac{P_0^{high}}{-z} - R_z(\Delta) \right\| \\ =& \left\| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} \cdot \frac{P_0^{high}}{-z} - \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} \cdot R_z(\Delta_{high}) \right\| \\ \leqslant& \left\| \frac{P_0^{high}}{-z} \right\| \cdot \left\| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} - \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} \right\| + \left\| \frac{P_0^{high}}{-z} - R_z(\Delta_{\text{high}}) \right\| \cdot \left\| \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} \right\| \\ \leqslant& \frac{1}{|z|} \left\| \left[Id + \frac{P_0^{high}}{-z} \Delta_{reg.} \right]^{-1} - \left[Id + R_z(\Delta_{high}) \Delta_{reg.} \right]^{-1} \right\| + \left(1 + \frac{\|\Delta_{\text{reg.}}\|}{|z|} \right) \cdot \frac{1}{\lambda_1(\Delta_{high})}. \end{aligned}$$

Hence it remains to bound the left hand summand. For this we use the following fact (c.f. (Horn & Johnson, 2012), Section 5.8. "Condition numbers: inverses and linear systems"):

Given square matrices A, B, C with $C = B - A$ and $\|A^{-1}C\| < 1$, we have

$$\|A^{-1} - B^{-1}\| \leqslant \frac{\|A^{-1}\| \cdot \|A^{-1}C\|}{1 - \|A^{-1}C\|}.$$

In our case, this yields (together with $\|P_0^{high}\| = 1$) that

$$\begin{aligned}&\left\|\left[Id + P_0^{high}/(-z)\cdot\Delta_{reg.}\right]^{-1} - \left[Id + R_z(\Delta_{high})\Delta_{reg.}\right]^{-1}\right\| \\ \leqslant &\frac{(1+\|\Delta_{\text{reg.}}\|/|z|)^2\cdot\|\Delta_{\text{reg.}}\|\cdot\|\frac{P_0^{\text{high}}}{-z} - R_z(\Delta_{\text{high}})\|}{1-(1+\|\Delta_{\text{reg.}}\|/|z|)\cdot\|\Delta_{\text{reg.}}\|\cdot\|\frac{P_0^{\text{high}}}{-z} - R_z(\Delta_{\text{high}})\|}\end{aligned}$$

For S_{high} sufficiently large, we have

$$\| - P_0^{\text{high}}/z - R_z(\Delta_{\text{high}})\| \leqslant \frac{1}{2\,(1+\|\Delta_{\text{reg.}}\|/|z|)}$$

so that we may estimate

$$\begin{aligned}&\left\|\left[Id + \Delta_{reg.}\frac{P_0^{high}}{-z}\right]^{-1} - \left[Id + \Delta_{reg.}R_z(\Delta_{high})\right]^{-1}\right\| \\ \leqslant &2\cdot(1+\|\Delta_{\text{reg.}}\|)\cdot\|\frac{P_0^{\text{high}}}{-z} - R_z(\Delta_{\text{high}})\| \\ = &2\frac{1+\|\Delta_{\text{reg.}}\|/|z|}{\lambda_1(\Delta_{\text{high}})}\end{aligned}$$

Thus we have now established

$$\left\|\left[Id + \frac{P_0^{high}}{-z}\Delta_{reg.}\right]^{-1}\cdot\frac{P_0^{high}}{-z} - R_z(\Delta)\right| = \mathcal{O}\left(\frac{\|\Delta_{\text{reg.}}\|}{\lambda_1(\Delta_{\text{high}})}\right).$$

Hence we are done with the proof, as soon as we can establish

$$\left[-zId + P_0^{high}\Delta_{reg.}\right]^{-1}P_0^{high} = J^{\uparrow}R_z(\underline{\Delta})J^{\downarrow},$$

with $J^{\uparrow}, \underline{\Delta}, J^{\downarrow}$ as defined above. To this end, we first note that

$$J^{\uparrow}\cdot J^{\downarrow} = P_0^{high} \tag{7}$$

and

$$J^{\downarrow}\cdot J^{\uparrow} = Id_{\underline{G}}. \tag{8}$$

Indeed,the relation (7) follows from the fact that the eigenspace corresponding to the eignvalue zero is spanned by the vectors $\{\mathbb{1}_R\}_R$, with $\{R\}$ the connected components of G_{high}. Equation (8) follows from the fact that

$$\langle \mathbb{1}_R, \mathbb{1}_R\rangle = \underline{\mu}_R.$$

With this we have

$$\left[Id + P_0^{high}\Delta_{reg.}\right]^{-1}P_0^{high} = \left[Id + J^{\uparrow}J^{\downarrow}\Delta_{reg.}\right]^{-1}J^{\uparrow}J^{\downarrow}.$$

To proceed, set

$$\underline{x} := F^{\downarrow}x$$

and

$$\mathscr{X} = \left[P_0^{high}\Delta_{reg.} - zId\right]^{-1}P_0^{high}x.$$

Then

$$\left[P_0^{high}\Delta_{reg.} - zId\right]\mathscr{X} = P_0^{high}x$$

and hence $\mathscr{X} \in \text{Ran}(P_0^{high})$. Thus we have

$$J^{\uparrow}J^{\downarrow}(\Delta_{\text{reg.}} - zId)J^{\uparrow}J^{\downarrow}\mathscr{X} = J^{\uparrow}J^{\downarrow}x.$$

Multiplying with $J^{\downarrow}$ from the left yields

$$J^{\downarrow}(\Delta_{\text{reg.}} - zId)J^{\uparrow}J^{\downarrow}\mathscr{X} = J^{\downarrow}x.$$

Thus we have

$$(J^{\downarrow}\Delta_{\text{reg.}}J^{\uparrow} - zId)J^{\uparrow}J^{\downarrow}\mathscr{X} = J^{\downarrow}x.$$

This – in turn – implies

$$J^{\uparrow}J^{\downarrow}\mathscr{X} = \left[J^{\downarrow}\Delta_{\text{reg.}}J^{\uparrow} - zId\right]^{-1}J^{\downarrow}x.$$

Using

$$P_0^{high}\mathscr{X} = \mathscr{X},$$

we then have

$$\mathscr{X} = J^{\uparrow}\left[J^{\downarrow}\Delta_{\text{reg.}}J^{\uparrow} - zId\right]^{-1}J^{\downarrow}x.$$

We have thus concluded the proof if we can prove that $J^{\downarrow}\Delta_{\text{reg.}}J^{\uparrow}$ is the Laplacian corresponding to the graph $\underline{G}$ defined in Definition J.1. But this is a straightforward calculation. □

As a corollary, we find

Corollary J.4. We have

$$R_z(\Delta)^k \to J^{\uparrow}R^k(\underline{\Delta})J^{\downarrow}$$

Proof. This follows directly from the fact that

$$J^{\downarrow}J^{\uparrow} = Id_{\underline{G}}.$$

□

J.2. Example II: Graph Rewiring

In real world unweighted graph datasets, the presence of edges is often determined by arbitrary thresholds (Gasteiger et al., 2019b). Thus node embeddings should not depend too strongly on the presence of any given edge. At the beginning of Section 3, we already observed that deleting an edge in a large fully connected clique corresponds to a minor change in geometry from the perspective of information diffusion. Here we generalize this finding to arbitrary rewiring operations within high connectivity areas. To this end, let G be a graph with adjacency matrix A. Let us split the adjacency matrix as $A = A_c + A_{rw}$ (c.f. Figure 6). We will keep the summand A_c constant and perform rewiring operations within the graph structure determined by A_{rw} (depicted in Figure 6 (c)).

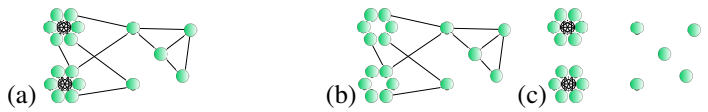


Figure 6. figure
(a) Graph G with adjacency matrix A, (b) G_c corresponding to A_c, (c) G_{rw} corresponding to A_{rw}

Let us denote the hence obtained modified partial adjacency matrix by $\tilde{A}_{rw}$. For the total modified graph structure let us write $\tilde{A} = A_c + \tilde{A}_{rw}$. Below, we then prove for the graph Laplacians $\Delta, \tilde{\Delta}$ corresponding to the graph structures determined by A and $\tilde{A}$, that we have

$$\|(\Delta + Id)^{-1} - (\tilde{\Delta} + Id)^{-1}\| \leqslant C_{A_c} \cdot (1/\lambda_1(\Delta_{rw}) + 1/\lambda_1(\tilde{\Delta}_{rw})). \tag{9}$$

Here $\lambda_1(\Delta_{rw})$ is the first non-zero eigenvalue of the Laplacian Δ_{rw} corresponding to the graph structure A_{rw} (c.f. Fig. 6 (c)). It is a well known fact in spectral graph theory, that much information about the connectivity of a graph G_{rw} is encoded into the first non-zero eigenvalue $\lambda_1(\Delta_{rw})$ of its graph Laplacian Δ_{rw}. For an unweighted graph G on N nodes, $\lambda_1(\Delta_{rw})$ is for example maximized if every node is connected to all other nodes in which case we have $\lambda_1(\Delta_{rw}) = N$.

Combining this result with Theorems 3.7 and 3.9, we see that the transferability error for networks confronted with the graph structures A and $\tilde{A}$ decreases inversely with the connectivity within G_{rw}.

Hence let us, prove (9):

$$\|(\Delta + Id)^{-1} - (\tilde{\Delta} + Id)^{-1}\| \leqslant C_{A_c} \cdot (1/\lambda_1(\Delta_{rw}) + 1/\lambda_1(\tilde{\Delta}_{rw})). \tag{10}$$

With the work that we have already done in Appendix J.1, this is now straightforward. We note that an application of the triangle inequality yields

$$\|(\Delta + Id)^{-1} - (\tilde{\Delta} + Id)^{-1}\| \leqslant \|(\Delta + Id)^{-1} - J^{\uparrow}(\underline{\Delta} + Id)^{-1}J^{\downarrow}\| + \|(\tilde{\Delta} + Id)^{-1} - J^{\uparrow}(\underline{\Delta} + Id)^{-1}J^{\downarrow}\|$$

Here $\underline{\Delta}$ is the Laplacian arising from collapsing the connected clusters of G_{rw} to single nodes. We may then simply use (2).

J.3. Example III: Graphs discretizing an Ambient Space.

The concept of characteristic operators capturing the geometry of the space on which they are defined is not limited to graphs: Similar considerations also apply to continuous spaces such as manifolds $\mathcal{M}$, where the Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ can be thought of as a continuous analogue of the graph Laplacian (Hein et al., 2006). Motivated by this observation, we consider the setting of two graphs G_1, G_2 discretely approximating the same ambient space (c.f. e.g. Fig. 7). Mathematically, this notion can be made precise using the concept of generalized norm resolvent convergence (Post, 2012; Post & Simmer, 2021): Making use of projection operators $J_i^{\downarrow}$ mapping from $\mathcal{M}$ to G_i and interpolation operators $J_i^{\uparrow}$ mapping from G_i to $\mathcal{M}$, one then measures the difference $\|(\Delta_i + Id)^{-1} - J_i^{\uparrow}(\Delta_{\mathcal{M}} + Id)^{-1}J_i^{\downarrow}\| \leqslant \delta$. The fidelity of the discrete approximation is then determined by the size of $\delta \ll 1$ (Post, 2012; Post & Simmer, 2021). As we discuss in detail below, we have in this setting of two graphs discretizing the same ambient space that $\|(\Delta_1 + Id)^{-1} - (J_1^{\downarrow}J_2^{\uparrow})(\Delta_2 + Id)^{-1}(J_2^{\downarrow}J_1^{\uparrow})\| \lesssim 2\delta$. Thus Theorem 3.7 directly applies and networks may be transferred between G_1 and G_2.

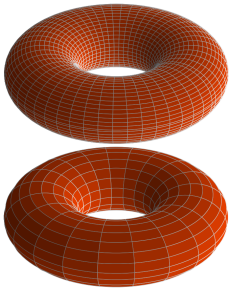


Figure 7. figure
Distinct Torus Discretizations

Somewhat similar transferability settings of graphs discretizing manifolds have been considered in other works: In (Levie et al., 2019), transferability for bandlimited signals sampled from manifolds are considered and a stability constant that grows linearly with the number of allowed eigenvalues is derived. The setup in (Wang et al., 2022) considers graphs that are statistically sampled from a manifold and yields probabilistic transferability statements. In contrast, our framework provides results beyond the probabilistic setting and without stability constants depending linearly on the bandwidth of band-limited features.

Hence let us further discuss the setting of two graphs discretizing the same ambient space $\mathcal{M}$ in the sense of

$$\|(\Delta_i + Id)^{-1} - J_i^{\uparrow}(\Delta_{\mathcal{M}} + Id)^{-1}J_i^{\downarrow}\| \leqslant \delta.$$

We will assume $J_i^{\downarrow}J_i^{\uparrow} = Id_{G_i}$, which is a justified assumption, as Example J.5 below elucidates. In this setting, we then have

$$\begin{aligned}
&\|(\Delta_1 + Id)^{-1} - (J_1^{\downarrow}J_2^{\uparrow})(\Delta_2 + Id)^{-1}(J_2^{\downarrow}J_1^{\uparrow})\| \\
=&\|(\Delta_1 + Id)^{-1} - J_1^{\downarrow}(\Delta_{\mathcal{M}} + Id)^{-1}J_1^{\uparrow} + J_1^{\downarrow}(\Delta_{\mathcal{M}} + Id)^{-1}J_1^{\uparrow} - (J_1^{\downarrow}J_2^{\uparrow})(\Delta_2 + Id)^{-1}(J_2^{\downarrow}J_1^{\uparrow})\| \\
\leqslant&\|(\Delta_1 + Id)^{-1} - J_1^{\downarrow}(\Delta_{\mathcal{M}} + Id)^{-1}J_1^{\uparrow}\| + \|J_1^{\downarrow}(\Delta_{\mathcal{M}} + Id)^{-1}J_1^{\uparrow} - (J_1^{\downarrow}J_2^{\uparrow})(\Delta_2 + Id)^{-1}(J_2^{\downarrow}J_1^{\uparrow})\|
\end{aligned}$$

We note

$$\begin{aligned}&\|(\Delta_1 + Id)^{-1} - J_1^{\downarrow}(\Delta_{\mathcal{M}} + Id)^{-1}J_1^{\uparrow}\|\\ =&\|J_1^{\downarrow}J_1^{\uparrow}(\Delta_1 + Id)^{-1}J_1^{\downarrow}J_1^{\uparrow} - J_1^{\downarrow}(\Delta_{\mathcal{M}} + Id)^{-1}J_1^{\uparrow}\|\\ \leqslant&\|J_1^{\downarrow}\|\|J_1^{\uparrow}\| \cdot \|(\Delta_1 + Id)^{-1} - J_1^{\uparrow}(\Delta_{\mathcal{M}} + Id)^{-1}J_1^{\downarrow}\| \lesssim \delta.\end{aligned}$$

We consider:

$$\begin{aligned}&\|(\Delta_{\mathcal{M}} + Id)^{-1} - (J_1^{\downarrow}J_2^{\uparrow})(\Delta_2 + Id)^{-1}(J_2^{\downarrow}J_1^{\uparrow})\|\\ \leqslant&\|J_1^{\downarrow}\|\|J_1^{\uparrow}\| \cdot \|(\Delta_{\mathcal{M}} + Id)^{-1} - J_2^{\uparrow}(\Delta_2 + Id)^{-1}J_2^{\downarrow}\|\\ \lesssim&\|(\Delta_{\mathcal{M}} + Id)^{-1} - J_2^{\uparrow}(\Delta_2 + Id)^{-1}J_2^{\downarrow}\| \leqslant \delta.\end{aligned}$$

Hence we have indeed established

$$\|(\Delta_1 + Id)^{-1} - (J_1^{\downarrow}J_2^{\uparrow})(\Delta_2 + Id)^{-1}(J_2^{\downarrow}J_1^{\uparrow})\| \lesssim 2\delta.$$

Next let us consider an explicit example.

Example J.5. To this end, let us revisit the torus-setting introduced in Fig. 7.

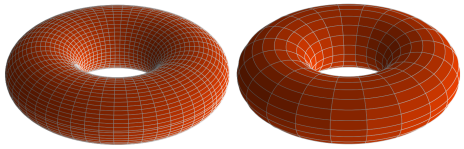


Figure 8. Distinct Torus Discretizations

In fact, instead of bounding the resolvent distances (10) after one which might apply Theorem 3.7 to quantify filter transferability, we directly bound the diffusion distances as originally proposed in Definition 3.1.

We begin by recalling that the standard torus $\mathbb{T}$ arises as the cartesian product of two circles S_1 of circumference 2π:

$$\mathbb{T} = S^1 \times S^1.$$

Let us parametrize these circles via angles $0 \leqslant \theta_1, \theta_1 \leqslant 2\pi$. The Laplacian on $\mathbb{T}$ can then be written as

$$\Delta_{\mathbb{T}} = -\partial_{\theta_1}^2 - \partial_{\theta_2}^2.$$

A set of corresponding normalized eigenfunctions are given as

$$\phi_{k_1,k_2} = \frac{1}{2\pi}e^{-ik_1\theta_1}e^{-ik_2\theta_2}$$

with corresponding eigenvalues

$$\lambda_{k_1,k_2} = k_1^2 + k_2^2$$

and $k_1, k_2 \in \mathbb{Z}$.

We now consider a regular discretization of $\mathbb{T}$ using N^2 nodes. This mesh can be thought of as arising from regular discretizations of each S^1 factor; with a node being placed at angles $\phi = \frac{2\pi}{N}k$ with $0 \leqslant k \leqslant N$. The individual node weight of each node in the mesh discretization of $\mathbb{T}$ is set to $\mu = \frac{(2\pi)^2}{N^2}$. We might think of this discretization $\mathbb{T}_N$ pf $\mathbb{T}$

as arising via a cartesian product of the group $\mathbb{Z}/N\mathbb{Z}$ (i.e. the group of integers modulo N) with itself. Each node of $\mathbb{T}_N = \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}$ is then specified by a tuple $(a, b) \in \mathbb{T}_N$, with $a \in \mathbb{Z}/N\mathbb{Z}$ and $b \in \mathbb{Z}/N\mathbb{Z}$.

The graph Laplacian Δ_N on $\mathbb{T}_N = \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}$ then acts on a scalar node signal x_{ab} as

$$(\Delta_N x)_{ab} = \frac{N^2}{(2\pi)^2}\left(4x_{ab} - x_{(a+1)b} - x_{(a-1)b} - x_{a(b+1)} - x_{a(b-1)}\right).$$

Henceforth we will adopt the notation $x(a, b) \equiv x_{ab}$.
Normalized eigenvectors for this Laplacian Δ_N on $\mathbb{T}_N$ are given as

$$\phi^N_{k_1,k_2} = \frac{1}{2\pi} e^{-i\frac{2\pi k_1}{N}a} e^{-i\frac{2\pi k_1}{N}b}$$

with $0 \leqslant k_1, k_2 \leqslant (N-1)$. Corresponding eigenvalues are found to be

$$\lambda^N_{k_1,k_2} = \frac{N^2}{\pi^2}\left[\sin^2\left(\frac{\pi}{N}\cdot k_1\right) + \sin^2\left(\frac{\pi}{N}\cdot k_2\right)\right].$$

To facilitate contact between $\mathbb{T}$ and its graph approximation $\mathbb{T}_N$, we define an interpolation operator $J^{\uparrow}_N$ that maps a graph signal $f(a, b)$ defined on $\mathbb{T} = \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}$ to a function $\overline{f}$ defined on $\mathbb{T}$ by defining

$$\overline{f}(\theta_1, \theta_2) = f(a, b)$$

whenever $\frac{2\pi}{N}(a-1) \leqslant \theta_1 \leqslant \frac{2\pi}{N}a$ and $\frac{2\pi}{N}(b-1) \leqslant \theta_2 \leqslant \frac{2\pi}{N}b$.
We then take $J^{\downarrow}$ to be the adjoint of $J^{\uparrow}$ (i.e. $J^{\downarrow} = (J^{\uparrow})^*$. It is not hard to see that $J^{\downarrow}J^{\uparrow} = Id_{\mathbb{T}_N}$.
We now want to show that (for $t > 0$)

$$\|e^{-t\Delta_{\mathbb{T}}} - J^{\uparrow}e^{-t\Delta_N}J^{\downarrow}\| \to 0 \tag{11}$$

as $N \to \infty$. To this end, denote by P_{k_1,K_2} the orthogonal projection onto ϕ_{k_1,k_2}. Denote by $P^N_{k_1,K_2}$ the orthogonal projection onto $\overline{\phi^N_{k_1,k_2}}$. We note

$$\|e^{-t\Delta_{\mathbb{T}}} - J^{\uparrow}e^{-t\Delta_N}J^{\downarrow}\| = \left\| \sum_{k_1,k_2 \in \mathbb{Z}} e^{-\lambda_{k_1,k_2}t} P_{k_1,k_2} - \sum_{--\frac{N-1}{2} \leqslant p_1,p_2 \leqslant \frac{N-1}{2}} e^{-\lambda_{k_1,k_2}t} P^N_{p_1,p_2} \right\|.$$

From this we observe

$$\|e^{-t\Delta_{\mathbb{T}}} - J^{\uparrow}e^{-t\Delta_N}J^{\downarrow}\| = \left\| \sum_{k_1,k_2 \in \mathbb{Z}} e^{-\lambda_{k_1,k_2}t} P_{k_1,k_2} - \sum_{--\frac{N-1}{2} \leqslant p_1,p_2 \leqslant \frac{N-1}{2}} e^{-\lambda^N_{p_1,p_2}t} P^N_{p_1,p_2} \right\|$$

$$\leqslant \left\| \sum_{\frac{N-1}{2} < |k_1|,|k_2|} e^{-\lambda_{k_1,k_2}t} P_{k_1,k_2} \right\| + \left\| \sum_{--\frac{N-1}{2} \leqslant k_1,k_2 \leqslant \frac{N-1}{2}} \left(e^{-\lambda_{k_1,k_2}t} P_{k_1,k_2} - e^{-\lambda^N_{k_1,k_2}t} P^N_{k_1,k_2}\right) \right\|$$

For the first summand, we already have

$$\left\| \sum_{\frac{N-1}{2} < |k_1|,|k_2|} e^{-\lambda_{k_1,k_2}t} P_{k_1,k_2} \right\| \leqslant e^{-t\frac{(N-1)^2}{2}}.$$

Hence let us investigate the second summand. We note

$$\left\| \sum_{-\frac{N-1}{2} \leqslant k_1,k_2 \leqslant \frac{N-1}{2}} \left(e^{-\lambda_{k_1,k_2}t} P_{k_1,k_2} - e^{-\lambda^N_{k_1,k_2}t} P^N_{k_1,k_2}\right) \right\| \tag{12}$$

$$\leqslant \left\| \sum_{-\frac{N-1}{2} \leqslant k_1,k_2 \leqslant \frac{N-1}{2}} \left(e^{-\lambda_{k_1,k_2}t} - e^{-\lambda^N_{k_1,k_2}t}\right) P^N_{k_1,k_2} \right\| + \left\| \sum_{-\frac{N-1}{2} \leqslant k_1,k_2 \leqslant \frac{N-1}{2}} e^{-\lambda_{k_1,k_2}t}(P_{k_1,k_2} - P^N_{k_1,k_2}) \right\|$$

For the first summand we note

$$\begin{aligned}
&\left\| \sum_{-\frac{N-1}{2} \leqslant k_1,k_2 \leqslant \frac{N-1}{2}} \left(e^{-\lambda_{k_1,k_2} t} - e^{-\lambda^N_{k_1,k_2} t} \right) P^N_{k_1,k_2} \right\| \\
= &\sup_{-\frac{N-1}{2} \leqslant k_1,k_2 \leqslant \frac{N-1}{2}} \left| e^{-\lambda_{k_1,k_2} t} - e^{-\lambda^N_{k_1,k_2} t} \right| \\
= &\sup_{-\frac{N-1}{2} \leqslant k_1,k_2 \leqslant \frac{N-1}{2}} e^{-t(k_1^2+k_2^2)} \left| 1 - e^{-t\left(\frac{N^2}{\pi^2}\sin^2\left(\frac{\pi}{N}k_1\right)-k_1^2\right)} e^{-t\left(\frac{N^2}{\pi^2}\sin^2\left(\frac{\pi}{N}k_2\right)-k_2^2\right)} \right|
\end{aligned}$$

We note

$$\left(\frac{N^2}{\pi^2} \sin^2\left(\frac{\pi}{N}k\right) - k^2 \right) = \mathcal{O}\left(\frac{k^4}{N^2}\right).$$

Using

$$\frac{N^2}{\pi^2} \sin^2\left(\frac{\pi}{N} N^{\frac{1}{3}}\right) \lesssim N^{\frac{2}{3}}$$

we note

$$\begin{aligned}
&\sup_{-\frac{N-1}{2} \leqslant k_1,k_2 \leqslant \frac{N-1}{2}} e^{-t(k_1^2+k_2^2)} \left| 1 - e^{-t\left(\frac{N^2}{\pi^2}\sin^2\left(\frac{\pi}{N}k_1\right)-k_1^2\right)} e^{-t\left(\frac{N^2}{\pi^2}\sin^2\left(\frac{\pi}{N}k_2\right)-k_2^2\right)} \right| \\
\leqslant &\sup_{|k_1|,|k_2| \leqslant N^{\frac{1}{3}}} e^{-t(k_1^2+k_2^2)} \left| 1 - e^{-t\left(\frac{N^2}{\pi^2}\sin^2\left(\frac{\pi}{N}k_1\right)-k_1^2\right)} e^{-t\left(\frac{N^2}{\pi^2}\sin^2\left(\frac{\pi}{N}k_2\right)-k_2^2\right)} \right| \\
+ &\sup_{|k_1|,|k_2| > N^{\frac{1}{3}}} e^{-t(k_1^2+k_2^2)} \left| 1 - e^{-t\left(\frac{N^2}{\pi^2}\sin^2\left(\frac{\pi}{N}k_1\right)-k_1^2\right)} e^{-t\left(\frac{N^2}{\pi^2}\sin^2\left(\frac{\pi}{N}k_2\right)-k_2^2\right)} \right| \\
&\leqslant e^{-t(2N^{\frac{2}{3}})} + e^{-t(2N^{\frac{2}{3}})} + e^{-t(N^{\frac{2}{3}})}.
\end{aligned}$$

Hence it remains to bound the second summand in (12). We note

$$\begin{aligned}
&\left\| \sum_{-\frac{N-1}{2} \leqslant k_1,k_2 \leqslant \frac{N-1}{2}} e^{-\lambda_{k_1,k_2} t} (P_{k_1,k_2} - P^N_{k_1,k_2}) \right\| \\
\leqslant &\sum_{|k_1|,|k_2| \leqslant \frac{N-1}{2}} e^{-(k_1^2+k_2^2)t} \|P_{k_1,k_2} - P^N_{k_1,k_2}\|.
\end{aligned}$$

Next we note

$$\|P_{k_1,k_2} - P^N_{k_1,k_2}\| \leqslant 2 \left\|\phi_{k_1,k_2} - \phi_{k_1,k_2}\right\|.$$

It is not hard to see that

$$\left\| \phi_{k_1,k_2} - \overline{\phi^N_{k_1,k_2}} \right\| \leqslant 2C(|k_1| + |k|_2)\frac{2\pi}{N}$$

for some appropriately chosen $C > 0$. Hence we have

$$\begin{aligned}
&\left\| \sum_{-\frac{N-1}{2} \leqslant k_1,k_2 \leqslant \frac{N-1}{2}} e^{-\lambda_{k_1,k_2} t} (P_{k_1,k_2} - P^N_{k_1,k_2}) \right\| \\
\leqslant &\sum_{|k_1|,|k_2| \leqslant \frac{N-1}{2}} e^{-(k_1^2+k_2^2)t} \cdot 2C(|k_1| + |k|_2)\frac{2\pi}{N} \\
= &\mathcal{O}(1/N).
\end{aligned}$$

Where the lass claim follows from summability in k_1, k_2. Thus we have in total indeed established that (11) holds.

J.4. Example IV: Coarse graining weighted directed Graphs

In this section we consider a graph G with directed weighted adjacency matrix A^s which we (disjointly) decompose as

$$A^s \equiv A^c + s \cdot A^m$$

into a weighted directed (partial) adjacency matrix A_C which we keep constant and a weighted directed (partial) adjacency matrix $s \cdot A^m$. Both adjacency matrices determine directed graph structures on the same common node set $\mathcal{G}$. Similar to the setting of Appendix J.1, we are then interested in establishing that when $s \to \infty$ this graph is similar (from a diffusion perspective) to a coarse grained graph $\underline{G}$. In the proof of (2) in Appendix J.1, we saw that the the coarse grained "limit graph" $\underline{G}$ was determined by the structure of the kernel of the operator Δ_{high}; which encoded the connected components of the graph G_{high} into its vectors. We expect that this also persists in the directed setting.

In this directed setting, we are faced with the choice of whether to make use of the in-degree Laplacian

$$L^{\text{in}} = M^{-1}\left[D^{\text{in}} - A\right]$$

or the out-degree Laplacian

$$L^{\text{out}} = M^{-1}\left[D^{\text{out}} - A\right].$$

The following is known about the kernels of these operators (c.f. (Veerman & Lyons, 2020; Sahi, 2013)):

In-degree Laplacian: To understand the kernel of directed in-degree Laplacians, we need the concept of reaches. Reaches generalize the concept of connected components of undirected graphs (Veerman & Lyons, 2020): A subgraph $R \subseteq G$ is called reach, if for any two vertices $a, b \in R$ there is a directed path in R along which the (directed) edge weights do not vanish, and R simultaneously possesses no outgoing connections (i.e. for any $c \in G$ with $c \notin R$: $w_{ca} = 0$). We here limit ourselves to the setting where all reaches within a given graph are disjoint (c.f. (Veerman & Lyons, 2020) for the general setting).

Consider now a graph G with adjacency matrix A^m The dimensionality of the kernel of L^{in} on this graph is then given as the number of reaches N_{Reach} present in A^m. The right-kernel of L^{in} is spanned by the vectors $\{v_i\}_{1 \leqslant R \leqslant N_{\text{Reach}}}$ which have entry 1 at all nodes in reach R and are zero outside of R. By definition these vectors satisfy

$$L^{\text{in}} \cdot v_i = 0.$$

The left-kernel is spanned by vectors $\{w_R\}_{1 \leqslant R \leqslant N_{\text{Reach}}}$ so that w_R has non-zero entries only for nodes in reach R and is zero elsewhere. As can be derived from results in (Sahi, 2013), we may write $w_R = M\hat{w}_R$ with M the matrix of node weights and the entry $(\hat{w}_R)_i$ (for i a node in the reach R) given as

$$(\hat{w}_R)_i = \sum_{\tau_i \in \mathcal{T}_i^R} \prod_{(ab) \in \tau_i} A^m_{ab}.$$

Here $\mathcal{T}_i^R$ is the set of all spanning trees of the reach R that are rooted at node $i \in R$. τ_i is such a spanning tree beginning at node i. The quantity $\prod_{(ab)\in\tau_i} A^m_{ab}$ then multiplies all (directed) edge weights along the spanning tree τ_i. From this, we can derive that we may write the (not necessarily orthogonal) projection P projecting onto the kernel of L^{in} as

$$P = \sum_{R \in \text{Reaches of } A^m} \frac{v_R \cdot (M\hat{w}_R)^\intercal}{(M\hat{w}_R)^\intercal \cdot V_R}.$$

We might write this as

$$P = J^{\uparrow} J^{\downarrow}$$

with $J^{\downarrow}$ mapping (similarly to the setting in Appendix J.1) to a coarsified graph $\underline{G}$, whose node set consists of the reaches in the original graph structure determined by A:

$$\underline{\mathcal{G}} = \{R\}_{R \in \{\text{Reaches of } A^m\}}.$$

Similarly to Definition J.2, we then have for x a signal defined on the original graph G, that $(J^{\downarrow}x)$ is a signal on the coarsified graph $\underline{G}$. It is defined by specifying it on each node $R \in \underline{\mathcal{G}}$ as

$$(J^{\downarrow}x)_R = \frac{1}{(M\hat{w}_R)^{\intercal} \cdot V_R} \cdot (M\hat{w}_R)^{\intercal} \cdot x.$$

Similarly interpolation back up to G is defined as

$$J^{\uparrow}\underline{x} := \sum_{R \in \underline{\mathcal{G}}} \underline{x}_R \cdot v_R.$$

Out-degree Laplacian: For the out-degree Laplacian L^{out}, the roles of left- and right kernels above are essentially reversed. Instead of reaches R determined by the adjacency matrix A^m, one considers reaches $\tilde{R}$ determined by the transpose $(A^m)^{\intercal}$ of the adjacency matrix. The left kernel of the out-degree Laplacian is given as the set of vectors $\{\tilde{v}_{\tilde{R}}\}$ given as $\tilde{v}_{\tilde{R}} = Mv_{\tilde{R}}$, with

$v_{\tilde{R}}$ again the vector with entry 1 at all nodes in reach $\tilde{R}$ and zero outside of $\tilde{R}$. The right kernel is spanned by vectors $\{\tilde{w}_{\tilde{R}}\}$ whose ith entry is given by

$$(\tilde{w}_{\tilde{R}})_i = \sum_{\tilde{\tau}_i \in \mathcal{T}_i^{\tilde{R}}} \prod_{(ab) \in \tilde{\tau}_i} A^{\intercal}_{ab}.$$

Here $\mathcal{T}_i^{\tilde{R}}$ is the set of all spanning trees of the reach $\tilde{R}$ (as determined by the connectivity structure of the transposed adjacency matrix $(A^m)^{\intercal}$).

We then note for the projection $\tilde{P}$ onto the kernel of L^{out}, that we may write

$$\tilde{P} = \sum_{\tilde{R} \in \text{Reaches of } (A^m)^{\intercal}} \frac{\tilde{w}_{\tilde{R}} \cdot (Mv_{\tilde{R}})^{\intercal}}{(Mv_{\tilde{R}})^{\intercal} \cdot \tilde{w}_{\tilde{R}}}.$$

We may again write this as

$$P = \tilde{J}^{\uparrow}\tilde{J}^{\downarrow}$$

with $J^{\downarrow}$ mapping (similarly to the setting in Appendix J.1) to a coarsified graph $\underline{G}$, whose node set consists of the reaches in the adjacency structure determined by $(A^m)^{\intercal}$:

Similarly to above, we then have for x a signal defined on the original graph G, that $(\tilde{J}^{\downarrow}x)$ is a signal on the coarsified graph $\underline{G}$. It is defined by specifying it on each node $\tilde{R} \in \underline{\mathcal{G}}$ as

$$(\tilde{J}^{\downarrow}x)_{\tilde{R}} = \frac{1}{(Mv_{\tilde{R}})^{\intercal} \cdot \tilde{w}_{\tilde{R}}} \cdot (Mv_{\tilde{R}})^{\intercal} \cdot x$$

Similarly interpolation back up to G is defined as

$$\tilde{J}^{\uparrow}\underline{x} := \sum_{\tilde{R} \in \underline{\mathcal{G}}} \underline{x}_R \cdot \tilde{w}_{\tilde{R}}.$$

In the setting

$$A_s \equiv A_c + s \cdot A^m$$

we may then prove (exactly as done in Appendix J.1) that – with L_s^{in}, L_s^{out} the in-and out-degree Laplacians corresponding to A_s – we have

$$\|(L_s^{\text{in}} + Id)^{-1} - J^{\downarrow}(\underline{L}^{\text{in}} + Id)^{-1}J^{\uparrow}\| = \mathcal{O}\left(\frac{1}{s}\right)$$

and

$$\|(L_s^{\text{out}} + Id)^{-1} - \tilde{J}^{\downarrow}(\underline{L}^{\text{out}} + Id)^{-1}\tilde{J}^{\uparrow}\| = \mathcal{O}\left(\frac{1}{s}\right).$$

This extends results of (Koke & Cremers, 2024), which still needed graphs to have the same in- and out-degee at every node.

Investigating the operators $J^{\uparrow}$ and $\tilde{J}^{\uparrow}$, we see that we have

$$J^{\uparrow}\mathbb{1}_{\underline{G}} = \mathbb{1}_{G}$$

$$\tilde{J}^{\uparrow}\mathbb{1}_{\underline{G}} \neq \mathbb{1}_{G}.$$

In view of Theorem I.17 we hence find:

Proposition J.6. In the directed setting, using the in-degree Laplacian allows for networks to be transferable between a graph G and its coarse grained version $\underline{G}$ even if biases are enabled. This is not true when using the out-degree Laplacian.

K. Additional Experimental Considerations

K.1. Additional details on Section 4

Dataset: The dataset we consider is the **QM7** dataset, introduced in (Blum & Reymond, 2009; Rupp et al., 2012). This dataset contains descriptions of 7165 organic molecules, each with up to seven heavy atoms, with all non-hydrogen atoms being considered heavy. A molecule is represented by its Coulomb matrix C^{Clmb}, whose off-diagonal elements

$$C^{\text{Clmb}}_{ij} = \frac{Z_i Z_j}{|R_i - R_j|}$$

correspond to the Coulomb-repulsion between atoms i and j. We discard diagonal entries of Coulomb matrices; which would encode a polynomial fit of atomic energies to nuclear charge (Rupp et al., 2012).

For each atom in any given molecular graph, the individual Cartesian coordinates R_i and the atomic charge Z_i are (in principle) also accessible individually. To each molecule an atomization energy - calculated via density functional theory - is associated. The objective is to predict this quantity. The performance metric is mean absolute error. Numerically, atomization energies are negative numbers in the range -600 to -2200. The associated unit is $[\textit{kcal/mol}]$.

Details on collapsing procedure: Again, we make use of the QM7 dataset (Rupp et al., 2012) and its Coulomb matrix description

$$C^{\text{Clmb}}_{ij} = \frac{Z_i Z_j}{|R_i - R_j|} \tag{13}$$

of molecules. We modify (all) molecular graphs in QM7 by deflecting hydrogen atoms (H) out of their equilibrium positions towards the respective nearest heavy atom. This is possible since the QM7 dataset also contains the Cartesian coordinates of individual atoms. Edge weights between heavy atoms then remain the same, while Coulomb repulsions between H-atoms and respective nearest heavy atom increasingly diverge; as is evident from (13).

Given an original molecular graph G with node weights $\mu_i = Z_i$, the corresponding limit graph $\underline{G}$ corresponds to a coarse grained description, where heavy atoms and surrounding H-atoms are aggregated into single super-nodes.

Mathematically, $\underline{G}$ is obtained by removing all nodes corresponding to H-atoms from G, while adding the corresponding charges $Z_H = 1$ to the node-weights of the respective nearest heavy atom. Charges in (13) are modified similarly to generate the weight matrix $\underline{W}$.

On original molecular graphs, atomic charges are provided via one-hot encodings. For the graph of methane – consisting of one carbon atom with charge $Z_C = 6$ and four hydrogen atoms of charges $Z_H = 1$ – the corresponding node-feature-matrix is e.g. given as

$$X = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 0\cdots \\ 1 & 0 & \cdots & 0 & 0 & 0\cdots \\ 1 & 0 & \cdots & 0 & 0 & 0\cdots \\ 1 & 0 & \cdots & 0 & 0 & 0\cdots \\ 1 & 0 & \cdots & 0 & 0 & 0\cdots \end{pmatrix}$$

with the non-zero entry in the first row being in the 6^{th} column, in order to encode the charge $Z_C = 6$ for carbon.

The feature vector of an aggregated node represents charges of the heavy atom and its neighbouring H-atoms jointly.

Node feature matrices are translated as $\underline{X} = J^{\downarrow}X$. Applying $J^{\downarrow}$ to one-hot encoded atomic charges yields (normalized) bag-of-word embeddings on $\underline{G}$: Individual entries of feature vectors encode how much of the total charge of the super-node is contributed by individual atom-types. In the example of methane, the limit graph $\underline{G}$ consists of a single node with node-weight

$$\mu = 6 + 1 + 1 + 1 + 1 = 10.$$

The feature matrix

$$\underline{X} = J^{\downarrow}X$$

is a single row-vector given as

$$\underline{X} = \left(\frac{4}{10}, 0, \cdots, 0, \frac{6}{10}, 0, \cdots\right).$$

Experimental Setup: We randomly select 1500 molecules for testing and train on the remaining graphs. On QM7 we run experiments for 23 different random random seeds and report mean and standard deviation. All experiments were performed on a single NVIDIA Quadro RTX 8000 graphics card.

Additional details on baselines: Baselines divided into **standard** architectures (GCN (Kipf & Welling, 2017), ChebNet (Defferrard et al., 2016), ARMA (Bianchi et al., 2019), BernNet (He et al., 2021), GATv2 (**?**)) and **multi- scale** architectures (PushNet (**?**), UFGNet (**?**), Lanczos (Liao et al., 2019)). Apart from UFGNet (already acting as a **pooling** layer) we also consider self-attention-pooling (**?**); both acting on the final layer (SAG) and as acting on the output of each indivifual layer, with resulting layer-wise features concatenated to produce the final embedding (SAG-M).

Additional details on training and models: All considered convolutional layers are incorporated into a two layer deep and fully connected graph convolutional architecture. In each hidden layer, we set the width (i.e. the hidden feature dimension) to

$$F_1 = F_2 = 64.$$

For BernNet, we set the polynomial order to $K = 3$ to combat appearing numerical instabilities. ARMA is set to $K = 2$ and $T = 1$. ChebNet uses $K = 2$. Lnaczos uses 20 Lanczos iterations, as proposed in the original paper (Liao et al., 2019). UFGNet uses Haar wavelets. For all baselines, the standard mean-aggregation scheme is employed after the graph-convolutional layers to generate graph level features. Finally, predictions are generated via an MLP.

LTF-Ψ^{Res} architecture, we set $\lambda = 1$ and and build filters using the $k = 1$ and $= 2$ atoms in $\Psi^{\text{Res}} = \{(z + \lambda)^{-k}\}_{k\in\mathbb{N}}$.

For the LTF-Ψ^{Exp} architecture, we set $t = 1$ and and build filters using the $k = 1$ and $= 2$ atoms in $\Psi^{\text{Exp}} = \{e^{-(kt_0)z}\}_{k\in\mathbb{N}}$.

As aggregation, we employ the graph level feature aggregation scheme discussed in Appendix I.8 with node weights set to atomic charges of individual atoms. Predictions are then generated via a final MLP with the same specifications as the one used for baselines.

L. Additional Experiment using the coarse graining setting of Section 3

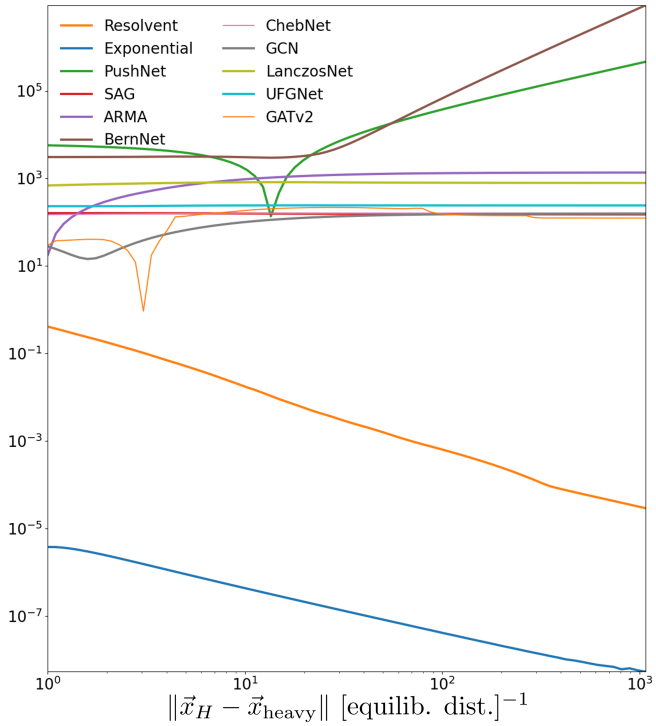


Figure 9. Transferability error $\|\tilde{F} - \underline{F}\|$

Equation (2) predicts in combination with Theorem I.18 that in the setting of the example of Section 3 the transferability error decreases with increasing edge-weights within the components of G_{high} that are being collapsed into single nodes. This is of course desirable: The stronger the connectivity within the connected components of G_{high}, the more it is justified to treat them as the (super-)nodes making up $\underline{G}$. To numerically verify that transferability errors indeed decrease with increasing connectivity within G_{high}, we modify the molecular graphs of QM7 again. We now deflect hydrogen atoms (H) out of their equilibrium positions towards the respective nearest heavy atom. This then introduces a setting precisely as discussed in in the example of Section 3: Edge-weights $A_{ij} = Z_i Z_j \cdot |\vec{x}_i - \vec{x}_j|^{-1}$ between heavy atoms remain the same, while those between H-atoms and nearest heavy atom increasingly diverge. We then compare embeddings $\{\underline{F}\}$ generated for coarsified graphs $\{\underline{G}\}$, with embeddings $\{\tilde{F}\}$ of graphs $\{\tilde{G}\}$ where hydrogen atoms have been deflected. As is evident from Figure 9, the transferability error of LTF-Ψ^{Res} and LTF-Ψ^{Exp} converges towards zero as the connectivity with G_{high} increases. Transferability errors of baselines remain large.

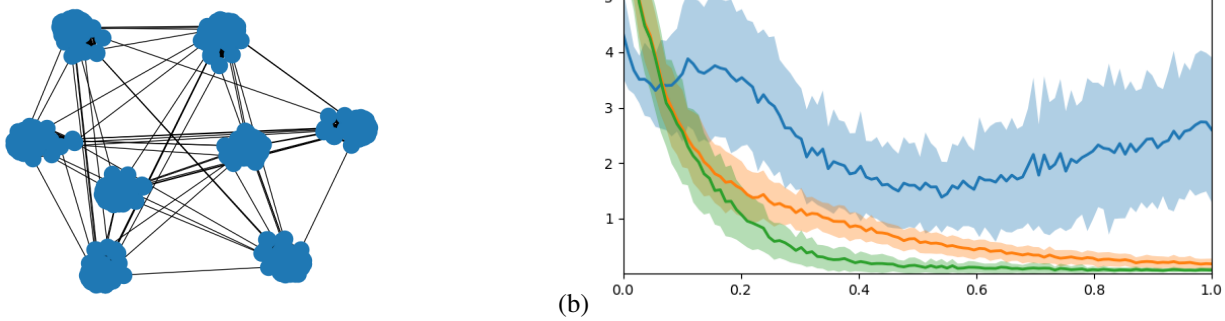
L.1. Transferability on Graphs generated via Stochastic Block Models

Stochastic Block Models: Stochastic block models (Holland et al., 1983) are generative models for random graphs that produce graphs containing strongly connected communities. In our experiments in this section, we consider a stochastic block model whose distributions is characterized by four parameters: The number of communities c_{number} determine how many (strongly connected) communities are present in the graph that is to be generated. The community size c_{size} determines the number of nodes belonging to each (strongly connected) community. The probability p_{connect} determines the probability that two nodes within the same community are connected by an edge. The probability p_{inter} determines the probabilities that two nodes in *different* communities are connected by an edge.

Experimental Setup: Since stochastic block models do not generate node-features, we equip each node with a randomly-generated unit-norm feature vector. Given such a graph G drawn from a stochastic block model, we then compute a version $\underline{G}$ of this graph, where all communities are collapsed to single nodes as described in Definition J.2. We then compare the feature vectors generated for G and $\underline{G}$. All experiments were performed on a single NVIDIA Quadro RTX 8000 graphics card. As before, we then consider the LTF-Ψ^{Res} and LTF-Ψ^{Exp} together with GCN as a baseline when investigating

transferability.

Experiment: Varying the Connectivity within the Communities: As discussed in detail in Appendix J.1 and Appendix J.2, we desire that networks assign similar feature vectors to graphs with strongly connected communities and coarse-grained versions of these graphs, where these communities are collapsed to aggregate nodes. The higher the connectivity within these communities, the more similar should the feature vector of the original graph G and its coarsified version $\underline{G}$ be, as Appendix J.1 established. In order to verify this experimentally, we fix the parameters $c_{\text{number}}, c_{\text{size}}$ and p_{inter} in our stochastic block model. We then vary the probability p_{connect} that two nodes within the same community are connected by an edge from $p_{\text{connect}} = 0$ to $p_{\text{connect}} = 1$. This corresponds to varying the connectivity within the communities from very sparse (or in fact no connectivity) to full connectivity (i.e. the community being a clique). In Figure 10 below, we then plot the difference of feature vectors generated by ResolvNet and baselines for for G and $\underline{G}$ respectively. For each $p_{\text{connect}} \in [0, 1]$, results are averaged over 100 graphs randomly drawn from the same stochastic block model.



(a) (b)

Figure 10. (a) Example Graph (b) Varying the parameter $p_{\text{connect}} \in [0, 1]$ for fixed $c_{\text{size}} = 20$, $p_{\text{inter}} = 2/c_{\text{size}}^2$ and $c_{\text{number}} = 10$.

We have chosen $p_{\text{inter}} = 2/c_{\text{size}}^2$ so that – on average – *clusters* are connected by two edges. The choice of two edges (as opposed to $1, 3, 4, 5, ...$) between clusters is not important; any arbitrary choice of p_{inter} ensures a decay behavior for ResolvNet as in Figure 10. A corresponding ablation study is provided below.

As can be inferred from Fig. 10, LTF-Ψ^{Res} and LTF-Ψ^{Exp} produce more and more similar feature-vectors for G and its coarse-grained version $\underline{G}$, as the connectivity within the clusters is increased. As a reference, we plot GCN for which such a transferability result clearly does not hold.

L.2. Implications for graphs with imbalanced Geometry

In the preceding experiments, baselines proved not transferable. Here we show that this lack of transferability can be harmful also for node-level tasks on a single graph that has an imbalanced geometry in the sense that it contains strongly connected subgraphs with weaker connectivity between such subgraphs.

To this end, we duplicated individual nodes on popular node-classification datasets (CITESEER & CORA (Sen et al., 2008; McCallum et al., 2000)) k-times to form (fully connected) k-cliques, while keeping the train-val-test partition constant. Models were then trained on the same (k-fold expanded) train-set and asked to classify nodes on the (k-fold expanded) test-partition. Baselines were chosen to form a representative selection of common information-propagation methods and in addition to previous baselines from Section 4 include GIN (Xu et al., 2019) and SAGE (Hamilton et al., 2017) (which could not handle weighted edges).

We then compare against an architecture using Laplace transform filters based on the atoms Ψ^{Res} introduced in Example 3.4. Contrary to earlier experiments, we also include the $k = 0$ term in $\Psi^{\text{Res}} = \{(z + \lambda)^{-k}\}_{k \in \mathbb{N}}$. To distinguish this architecture from previously employed networks that did not include the $k = 0$ atom ($\psi_0(L) = Id$), we do not refer to the architecture

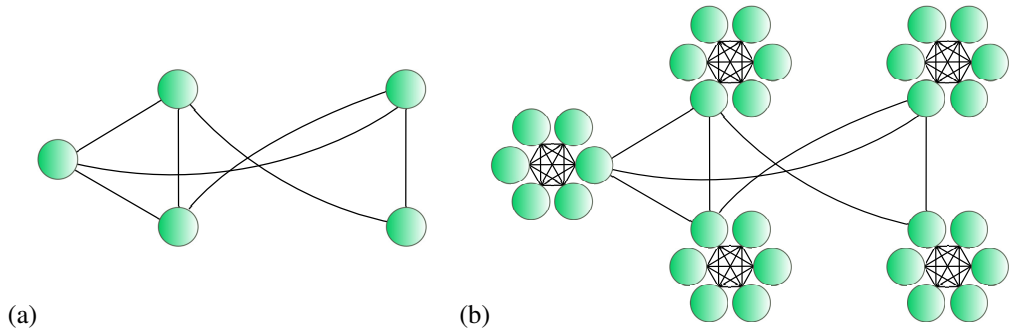


Figure 11. Individual nodes (a) replaced by k-cliques (b)

built here as LTF-Ψ^{Res}, but instead refer to it as ResolvNet.

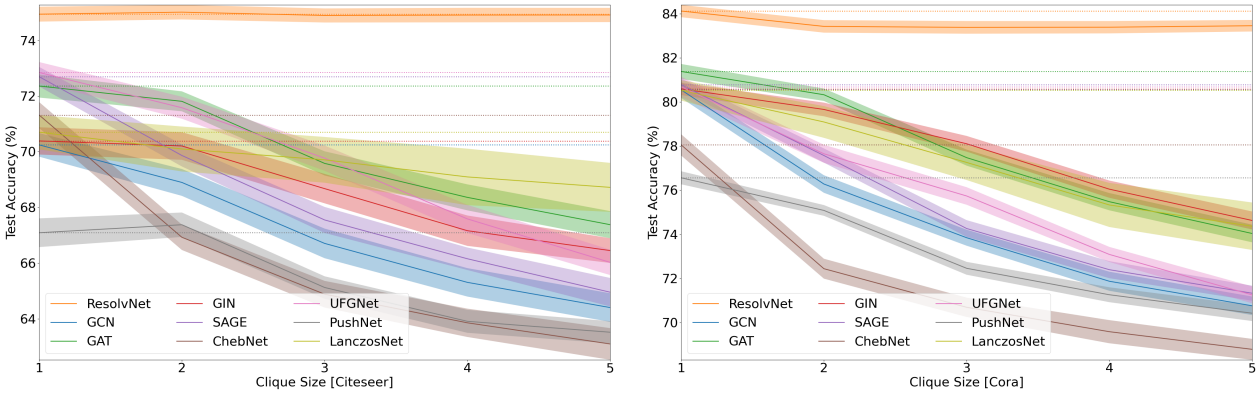


Figure 12. Node-Classification-Accuracy (↑) and uncertainty (for 100 runs) vs. clique size. ResolvNet with its Laplace transform filters remains stable while the performance of other architectures deteriorate significantly as geometry becomes more and more challenging with increasing clique-size.

As can be inferred from Fig. 12, the classification accuracy of all methods not employing Laplace transform filters decreases drastically as the geometry becomes more and more complex as k increases. We can understand the underlying reason for this considering the update rule implementing message passing in GCN as an example. There, a node feature matrix X is updated as $X \mapsto \hat{A}XW$, with the renormalized adjacency adjacency $\hat{A}$ determined by

$$\hat{A}_{ij} \sim \left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}\right)_{ij} = \frac{1}{\sqrt{d_i}} \cdot A_{ij} \cdot \frac{1}{\sqrt{d_j}}. \tag{14}$$

As the degree of each node increases (linearly) with increasing clique-size k, we see that the message-strength $\hat{A}_{ij}$ between the respective cliques decreases as $\hat{A}_{ij} \sim \frac{1}{k}$. Hence information propagation between the cliques becomes more and more challenging as k increases.

In principle however increasing the clique size does not increase the complexity of the classification task at hand as nodes are simply duplicated in the respective train-, val.- and test-sets.

What *does* become more challenging is the specific graph-geometry underlying the task. The considered ResolvNet architecture is able to handle this somewhat more challenging geometry; it consistently provides high classification accuracies even as k increases. This can be understood from the viewpoint of the considerations in Appendix J.1 and Appendix J.2: As the connectivity within the cliques increases (linearly as k becomes larger), the information flow over the graph G in Figure 11 (b) as implemented by an architecture using Laplace transform filters is more and more the same as an architecture that would first average information over the cliques of G; then project to a coarse grained graph $\underline{G}$ where the cliques are fused together to single nodes and subsequently propagate information there (c.f. the discussion of Appendix J.1

or Theorem I.7 together with its extended discussion in Appendix I.7). This avoids an interruption of the message passing scheme as in (14) and instead allows information to frrely flow *between* the cliques.

Additional details on training and models: All experiments were performed on a single NVIDIA Quadro RTX 8000 graphics card. We closely follow the experimental setup of (Gasteiger et al., 2019a) on which our codebase builds: All models are trained for a fixed maximum (and unreachably high) number of $n = 10000$ epochs. Early stopping is performed when the validation performance has not improved for 100 epochs. Test-results for the parameter set achieving the highest validation-accuracy are then reported. Ties are broken by selecting the lowest loss (c.f. (Velickovic et al., 2018)). Confidence intervals are calculated over multiple splits and random seeds at the 95% confidence level via bootstrapping.

We train all models on a fixed learning rate of $\text{lr} = 0.1$. Global dropout probability p of all models is optimized individually over $p \in \{0.3, 0.35, 0.4, 0.45, 0.5\}$. We use ℓ^2 weight decay and optimize the weight decay parameter λ for all models over $\lambda \in \{0.0001, 0.0005\}$. Where applicable (e.g. not for (He et al., 2021)) we choose a two-layer deep convolutional architecture with the dimensions of hidden features optimized over

$$K_\ell \in \{32, 64, 128\}. \tag{15}$$

In addition to the hyperparemeters specified above, some baselines have additional hyperparameters, which we detail here: BernNet uses an additional in-layer dropout rate of $\text{dp\_rate} = 0.5$ and for its filters a polynomial order of $K = 10$ as suggested in (He et al., 2021). Hyperparameters depth T and number of stacks K of the ARMA convolutional layer (Bianchi et al., 2019) are set to $T = 1$ and $K = 2$. ChebNet also uses $K = 2$ to avoid the known over-fitting issue (Kipf & Welling, 2017) for higher polynomial orders. The graph attention network (Velickovic et al., 2018) uses 8 attention heads, as suggested in (Velickovic et al., 2018).

For the ResolvNet model, we choose a depth of $L = 1$ with hidden feature dimension optimized over the values in (15) as for baselines. We empirically observed in the setting of *unweighted* graphs, that rescaling the Laplacian as

$$\Delta_{nf} := \frac{1}{c_{nf}} \Delta$$

with a normalizing factor c_{nf} before calculating the resolvent

$$R_z(\Delta_{nf}) := (\Delta_{nf} - z \cdot Id)^{-1} \tag{16}$$

on which we base our ResolvNet architectures improved performance.

For our ResolvNet architecture, we express this normalizing factor in terms of the largest singular value $\|\Delta\|$ of the (non-normalized) graph Laplacian. It is then selected among

$$c_{nf}/\|\Delta\| \in \{0.001, 0.01, 0.1, 2\}.$$

The value z in (16) is selected among

$$(-z) \in \{0.14, 0.15, 0.2, 0.25\}.$$