Transferability Properties of Graph Neural Networks

Luana Ruiz, Member, IEEE, Luiz F. O. Chamon, Member, IEEE, and Alejandro Ribeiro, Member, IEEE

Abstract—Graph neural networks (GNNs) are composed of layers consisting of graph convolutions and pointwise nonlinearities. Due to their invariance and stability properties, GNNs are provably successful at learning representations from data supported on moderate-scale graphs. However, they are difficult to learn on large-scale graphs. In this article, we study the problem of training GNNs on graphs of moderate size and transferring them to large-scale graphs. We use graph limits called graphons to define limit objects for graph filters and GNNs—graph filters and graphon neural networks (WNNs)—which we interpret as generative models for graph filters and GNNs. We then show that graphon filters and WNNs can be approximated by graph filters and GNNs sampled from them on weighted and stochastic graphs. Because the error of these approximations can be upper bounded, by a triangle inequality argument we can further bound the error of transferring a graph filter or a GNN across graphs. Our results show that (i) the transference error decreases with the graph size, and (ii) that graph filters have a transferability-discriminability tradeoff that in GNNs is alleviated by the scattering behavior of the nonlinearity. These findings are demonstrated empirically in a recommendation problem and in a decentralized control task.

Index Terms—Graph neural networks, transferability, graph signal processing, graphons.

I. INTRODUCTION

GRAPH neural networks (GNNs) are deep learning architectures for network data popularized by their state-of-the-art performance in a number of learning tasks [1], [2], [3], [4]. Besides working well in practice, GNNs are verifiably invariant to node relabelings and stable to graph perturbations [5], [6], which are properties that help understand their practical success. Each layer of a GNN composes one or more graph convolutions and a pointwise nonlinearity; the permutation equivariance and the stability properties are inherited from the graph convolution. Because it is a local graph operator, the graph convolution is also to credit for the ability to implement GNNs in a distributed way [7].

An important consequence of locality is that, unlike in a fully connected neural network (FCNN), in a GNN the number of parameters does not depend on the number of nodes. In fact, the number of parameters of a GNN is typically much smaller than the graph size. In theory, this should mean that GNNs can scale to graphs with large number of nodes. In practice, learning GNNs for large-scale graphs is difficult. One reason for that is that graph convolutions need full knowledge of the graph structure, which may be hard to measure for large-scale graphs. Another is that they require calculating large matrix-vector multiplications, which, for an $n$-node graph, have in the order of $O(n^2)$ complexity if the graph is not sparse.

Although the independence between the GNN weights and the graph does not guarantee scalability, it allows for the same GNN to be applied to different graphs. Thus, we could train the GNN on a graph of moderate size and transfer the learned weights to a larger graph. For this to work, the graph on which the GNN is trained—the training graph—and the one on which it is executed—the target graph—have to be similar. In this article, we define similar graphs as graphs that are associated with the same graphon [8], [9], [10], [11]. Graphons can be used to describe families of similar graphs because they are both the limits of convergent graph sequences and generative models for weighted (Definitions 1 and 2) and stochastic graphs (Definition 3). They proved to be a useful model in large-scale graph problems in statistics [12], [13], [14], game theory [15], network science [16], [17] and controls [18].

To understand whether a GNN learned on the training graph would work well on the target graph, we need to estimate the difference between the outputs of the GNN on these graphs or its transference error. To do so, we define the graphon filter and the graphon neural network (WNN), which are the limit objects of the graph filter and the GNN on the graphon [19]. By interpreting the graphon filter and the WNN as generative models for graph filters and GNNs, we show that they can be approximated by graph filters (Thm. 1) and GNNs (Thm. 2) sampled from them. To illustrate these results with concrete examples, in Props. 1–3, 5 we particularize these theorems to template, weighted and stochastic graphs. Since the error of these approximations can be upper bounded, by a triangle inequality argument we can further bound the error of transferring graph filters (Prop. 4) and GNNs (Prop. 6) across graphs.

These results have two important implications. The first is that both graph filters and GNNs have an inherent transferability property, i.e., they can be transferred across similar graphs with performance guarantees. But although transferability increases with the size of both the training and the target graph, in the finite-sample regime (i.e., on graphs of finite size) some...
spectral components are not transferable. The second is that there is a tradeoff between the transferability and the spectral discriminability of the graph convolution. While it would be natural to expect this tradeoff to be inherited by GNNs, numerical experiments on a movie recommendation task (Section V) show that GNNs are more transferable than convolutional filters. A plausible explanation is that, in GNNs, the transferability-discriminability tradeoff is alleviated by the scattering behavior of the nonlinearities; see Section IV-C.

A. Related Work

This article follows from a series of works on graphon signal processing (WSP), including [20]—which introduces WSP and proves convergence of graph to graphon filters—and [19]—which defines WNNS and studies GNN transferability over a narrow class of weighted graphs, where nodes are sampled from a regular partition and edges have deterministic weights given by the graphon. Herein, we extend upon the results of [19] by considering graphs with both stochastic nodes and edges, and proving a tighter transferability bound (of order $O(n^{-1/2})$) than the $O(n^{-1})$ bound in [19].

Transferability of GNNs has also been studied in [21], [22], [23]. These papers focus on the same problem we address in this article, but use different methods and arrive at different results. More importantly, what sets our work apart is that it unveils a tradeoff between the transferability and the spectral behavior of the graph convolution (and thus of GNNs)—which is not brought up in any other works.

Article [21] considers graphs sampled from generic topological spaces as opposed to graphons. Graphs and graph signals are obtained by application of a generic sampling operator, and mapped back to the topological space via an analogous reconstruction map. The bound on the GNN transferability error depends on the quality of the sampling-reconstruction process, whose error is assumed bounded. Moreover, the transferability analysis in [21] is restricted to signals in Paley-Wiener spaces, i.e., signals which are bandlimited. In contrast, in our article we give three examples of how one might sample graphs and graph signals from graphons and graphon signals (Definitions 1–3), and do not make any bandlimitedness assumptions on the input signal. The latter is particularly important because it allows understanding the transferability behavior of the GNN when the data has high-frequency information.

Same as in our article, article [22] considers graphs sampled from graphons and proves transferability of GNNs by showing that graph convolutions are transferable. However, while in our article we do not impose any restrictions on the input signal and only require the filter $h$ to be Lipschitz, in [22] more restrictive assumptions are placed on either the signal—which has to be in $L^p$ for $p > 2$—or the filter $h$—where $h'$ must be Lipschitz and $h(0) = 0$. Our assumptions lead to a detailed analysis of the spectral implications of transferability, showing that graph convolutions have a transferability-discriminability tradeoff and suggesting a hypothesis for why this tradeoff is less pronounced in GNNs. These considerations, which are the main result from our analysis, are not made in [22].

Article [23] gives finite-sample bounds for the convergence and stability of GNNs on graphs sampled from a random graph models $G(n, p_n)$ where the probability $p_n$ of sampling an edge is given by a weighting of a graphon and a decreasing function of $n$. As a consequence, their analysis applies to graphs ranging from dense (i.e., sampled from graphons) to moderately sparse. The main difference of our article with respect to [23] is that their transferability bound depends on the number of coefficients of the graph convolution, whereas ours is a function of the filter’s Lipschitz constant. It is by focusing on the Lipschitz property of the filter—without imposing any restrictions on its number of coefficients—that we are able to visualize the relationship between the transferability and the spectral discriminability of the graph convolution. This relationship is not discussed in [23].

Somewhat related to our article, [24] analyzes transferability from a different angle by assuming that the graph structure is a latent function of the node features, which can be seen as a homophily assumption. This latent function then defines a graph family in which GNNs are shown to transfer well empirically. Finally, GNN stability is closely related to transferability and is studied in the seminal article by [6]—which follows from stability analyses of graph scattering transforms [25]—, as well as in [26], [27], [28]. Other related work includes the convergence analyses of graph filters in [29] and [30]. These differ from our article in that they only focus on graph filters, and only give asymptotic transferability results.

Notation. All norms $\| \cdot \|$ refer to the $L^2$ norm. We use $\| \cdot \|$ to denote the operator norm induced by the $L^2$ norm.

II. Preliminary Definitions

We consider data supported on undirected graphs $G = (\mathcal{V}, \mathcal{E}, W)$ where $\mathcal{V}$ is a set of $|\mathcal{V}| = n$ nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of edges, and $\mathcal{W} : \mathcal{E} \to \mathbb{R}$ is a symmetric function assigning real weights to the edges in $\mathcal{E}$. This data is represented in the form of vectors $x \in \mathbb{R}^n$ called graph signals where $|x|_i$ is the value of the signal at node $i$ [31], [32]. To parametrize signal processing operations that depend on the graph topology by the graph $G$, a generic graph matrix representation $S \in \mathbb{R}^{n \times n}$, called the graph shift operator (GSO), is defined. The GSO $S$ is such that $S|_{ij} = s_{ij} \neq 0$ if and only if $i = j$ or $(i, j) \in \mathcal{E}$. The graph adjacency $A|_{ij} = (A)_{ij} = \mathcal{W}(i,j)$ [33], the graph Laplacian $S = L = \text{diag}(A) - A$ [31], and the random walk Laplacian $S = \text{diag}(A) - 1 L$ [34] are all examples of matrices that satisfy this property. Unless otherwise specified, in this article we consider $S = A$ and replace $\mathcal{W}$ by $S$ in the triplet $G = (\mathcal{V}, \mathcal{E}, S)$ to simplify notation.

The GSO gets its name from the fact that it defines a notion of shift for signals $x$ on the graph $G$. This is because, at each node $i$, the signal $z = Sx$ resulting from an application of the GSO to $x$ can be written elementwise as

$$z_i = \sum_{j=1}^{n} |S|_{ij} |x|_j = \sum_{j : |S|_{ij} \neq 0} |S|_{ij} |x|_j = \sum_{j \in \mathcal{N}(i)} |S|_{ij} |x|_j$$

where $\mathcal{N}(i) = \{ j : |S|_{ij} \neq 0 \}$ is the neighborhood of $i$. In other words, the outcome of $Sx$ at node $i$ is a result of the neighboring
nodes, shifting their data values \(x_j\), each weighted by the proximity measure \(|S_{ij}|\), to the node \(i\).

The notion of a graph shift, in turn, allows defining linear shift-invariant (LSI) or convolutional filters on graphs. Let \(h = [h_0, \ldots, h_{K-1}]^T\) be a vector of \(K\) real coefficients. The graph convolution with coefficients \(h\) is defined as [35], [36]

\[
\mathbf{H}(\mathbf{S})\mathbf{x} = \sum_{k=0}^{K-1} h_k \mathbf{S}^k \mathbf{x}
\]

i.e., it is given by the application of an order \(K\) polynomial of \(\mathbf{S}\) to the signal \(\mathbf{x}\), similarly to how the convolution is defined as a shift-and-sum operation in Euclidean domains.

Because \(\mathbf{S}\) is symmetric, it can be diagonalized as \(\mathbf{S} = \mathbf{VAV}^H\). The matrix \(\mathbf{A}\) is a diagonal matrix whose diagonal elements are the graph eigenvalues which we assume to be ordered according to their sign and in decreasing order of absolute value, i.e., \(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{K} \geq \lambda_{K+1}\). These are interpreted as graph frequencies. We consider high magnitude eigenvalues to be low frequencies and low magnitude eigenvalues to be high frequencies because we relate the graph signal’s frequency with the absolute value of its normalized Rayleigh quotient. Explicitly, we define the frequency of a graph signal \(\mathbf{x}\) as 

\[
\nu = \frac{x^H \mathbf{A}^{-1} \mathbf{x}}{\sqrt{x^H \mathbf{A}^{-1} \mathbf{x} \mathbf{x}^H \mathbf{A}^{-1} \mathbf{x}}}
\]

The graph eigenvectors, given by the columns of \(\mathbf{V}\), can be seen as the graph’s oscillation modes [37]. Importantly, the eigenvectors in \(\mathbf{V}\) form an orthonormal basis of \(\mathbb{R}^{n \times n}\) which we call the graph spectral basis. The graph Fourier transform (GFT) of \(\mathbf{x}\) is defined as the projection of \(\mathbf{x}\) onto this basis,

\[
\hat{\mathbf{x}} = \mathbf{GFT}(\mathbf{x}) = \mathbf{V}^H \mathbf{x}.
\]

Similarly, the inverse GFT is defined as \(\mathbf{x} = \mathbf{V} \hat{\mathbf{x}}\). Substituting \(\mathbf{S} = \mathbf{VAV}^H\) into (1) and calculating the GFT of \(\mathbf{y} = \mathbf{H}(\mathbf{S})\mathbf{x}\), we get

\[
\hat{\mathbf{y}} = \mathbf{V}^H \mathbf{H}(\mathbf{S})\mathbf{x} = \sum_{k=0}^{K-1} h_k \mathbf{A}^k \mathbf{V}^H \mathbf{x} = \sum_{k=0}^{K-1} h_k \mathbf{A}^k \mathbf{x}
\]

from which we obtain the spectral representation of the graph convolution \(h(\lambda) = \sum_{k=0}^{K-1} h_k \lambda^k\). There are two important things to note about the function \(h\). First, while the GFTs of \(\mathbf{x}\) and \(\mathbf{y}\) depend on the eigenvectors of \(\mathbf{S}\), the spectral response of \(\mathbf{H}(\mathbf{S})\) is obtained by evaluating only at the eigenvalues of the graph. This is illustrated in Fig. 2, where the red curve represents \(h(\lambda)\) and the green lines represent the eigenvalues of the graph \(\mathbf{G}\). Second, on a given graph \(\mathbf{G}\) polynomial filters \(\mathbf{H}(\mathbf{S})\) of order \(K = N\) can be used to implement filters with spectral response given by any arbitrary function \(f\) at the eigenvalues \(\lambda\), i.e., \(h(\lambda_i) = f(\lambda_i)\), as a consequence of the Cayley-Hamilton theorem. In particular, if \(f\) is analytic, this function can be approximated by a polynomial filter \(h(\lambda)\) for all \(\lambda\). The error of the approximation will depend on the filter order \(K\).

### A. Graph Neural Networks

A GNN is a deep convolutional architecture where each layer consists of (i) a bank of convolutional filters like the one in (1) and (ii) a nonlinear activation function. The convolutional filterbank transforms the \(F_{\ell-1}\) features \(x_{\ell-1}^g\) from layer \(\ell - 1\) into \(F_\ell\) intermediate linear features given by

\[
u_\ell^f = \sum_{g=1}^{F_{\ell-1}} \mathbf{H}_{\ell}^g(S)x_{\ell-1}^g
\]

where \(1 \leq f \leq F_\ell\). Specifically, the filter \(\mathbf{H}_{\ell}^g(S)\) maps feature \(g\) of layer \(\ell - 1\) into feature \(f\) of layer \(\ell\). The collection of filters \(\mathbf{H}_{\ell}^g(S)\) for all \(f, g\) yields a filterbank \(\{\mathbf{H}_{\ell}^g(S)\}_{f,g}\) with a total of \(F_{\ell-1} \times F_\ell\) filters like the one in (1).

The activation function is usually a pointwise nonlinearity such as the ReLU or sigmoid, but localized implementations incorporating the graph structure have also been proposed [5]. Letting \(\sigma\) denote the activation function, the \(\ell\)th layer of the GNN can be written as

\[
x_{\ell}^f = \sigma(\nu_\ell^f).
\]

Because \(\sigma\) is pointwise, at each node \(i \in \mathcal{V}\) we have \([x_{\ell}^f]_i = \sigma([\nu_\ell^f]_i)\). If the total number of layers of the GNN is \(L\), (4)–(5) are repeated for layers \(\ell = 1, \ldots, L\), the output of each layer being the input to the next. At \(\ell = 1\), the input features \(x_0^g\) are the input data \(x^0\) for \(1 \leq g \leq F_0\). The GNN output is given by the features of the last layer, i.e., \(y^f = x_L^f\).

The goal of graph machine learning is to obtain meaningful representations \(y\) from input data \(x\) (e.g., through empirical risk minimization). This can be thought of as learning a map \(\Phi: \mathcal{X} \rightarrow \mathcal{Y}\) where \(\Phi\) is parametric on the graph. The function \(\Phi\) could be, for instance, parametrized by a graph filter \(\gamma = \mathbf{H}(\mathbf{S})\mathbf{x} = \sum_{k=0}^{K-1} h_k \mathbf{S}^k \mathbf{x}\) [cf. (1)]. In this case, we would write

\[
y = \Phi(\mathbf{x}; \mathbf{h}, \mathbf{S})
\]

where \(\mathbf{h} = [h_0, \ldots, h_{K-1}]\). Alternatively, \(\Phi\) could be parametrized by a \(L\)-layer GNN with inputs \(x = \{x^\ell\}_{\ell=0}^{F_0}\) and outputs \(y = \{y^f\}_{f=1}^{F_L}\) [cf. (4)–(5)].

In this case, we would write

\[
y = \Phi(\mathbf{x}; \mathcal{H}, \mathbf{S})
\]

where \(\mathcal{H} = \{h_{fg}\}_{\ell=1}^{F_L}\) is a tensor grouping the GNN parameters \(h_{fg}\) for all layers \(1 \leq \ell \leq L\) and features \(1 \leq f \leq F_\ell\) and \(1 \leq g \leq F_{\ell-1}\). This notation is particularly convenient for GNNs because it summarizes (4)–(5) for \(\ell = 1, \ldots, L\) into a single function \(\Phi\). It also highlights the fact that the graph filter parameters \(\mathbf{h}\) and the GNN parameters \(\mathcal{H}\) are independent of the GSO \(\mathbf{S}\) (and thus of the graph \(\mathbf{G}\)).

### B. Graphs and Graphons

Graphons are the limit objects of sequences of dense undirected graphs. They are defined as functions \(\mathbf{W}: [0, 1]^2 \rightarrow [0, 1]\) where \(\mathbf{W}\) is bounded, measurable and symmetric. A sequence of graphs converging to the constant or Erdős-Rényi graphon is shown in Fig. 1.

A sequence of graphs converges to a graphon in the sense that densities of homomorphisms from certain graph motifs onto these graphs converge. Given an unweighted graph \(\mathbf{F} = (\mathcal{V}, \mathcal{E})\) and a (possibly weighted) graph \(\mathbf{G} = (\mathcal{V}, \mathcal{E}, \mathbf{S})\), a homomorphism between \(\mathbf{F}\) and \(\mathbf{G}\) is defined as a map \(\beta: \mathcal{V} \rightarrow \mathcal{V}\) such as
that, if \((i, j) \in \mathcal{E}', (\beta(i), \beta(j)) \in \mathcal{E}\). Because there are multiple such maps between the nodes of \(F\) and \(G\), to count them we define the number of homomorphisms between \(F\) and \(G\) as

\[
\text{hom}(F, G) = \sum_{\beta} \prod_{(i,j) \in \mathcal{E}'} |S|_{\beta(i)\beta(j)}.
\]

For unweighted graphs \(G\), \(\text{hom}(F, G)\) is the total number of adjacency preserving maps between \(V'\) and \(V\). For weighted graphs, each map \(\beta\) is weighted proportionally to the product of the weights of the edges \((\beta(i), \beta(j)) \in \mathcal{E}\) for all \(i, j \in V\).

Since homomorphisms are only a fraction of the total number of maps between the nodes of \(F\) and the nodes of \(G\), we can define a density of homomorphisms from \(F\) to \(G\). This density is denoted \(t(F, G)\) and given by

\[
t(F, G) = \frac{\text{hom}(F, G)}{|V'|^{||V||}}.
\]

The density \(t(F, G)\) can be likened to the probability of sampling a homomorphism when picking a random map \(\beta : V' \rightarrow V\). This allows extending the notion of density of homomorphisms to graphons. Explicitly, we define the density of homomorphisms from the graph \(F\) to the graphon \(W\) as

\[
t(F, W) = \int_{[0,1]^{||V'||}} \prod_{(i,j) \in \mathcal{E}'} W(u_i, u_j) \prod_{i \in V'} du_i
\]

i.e., \(t(F, W)\) is the probability of sampling \(F\) from \(W\).

A sequence of graphs \(\{G_n\}\) converges to \(W\) if and only if the density of homomorphisms between any finite, undirected and unweighted graph \(F = (V', \mathcal{E}')\) and the \(G_n\) converges to the density of homomorphisms between \(F\) and \(W\). More formally, \(G_n \rightarrow W\) if and only if

\[
\lim_{n \to \infty} t(F, G_n) = t(F, W)
\]

for all finite, undirected and unweighted \(F\) [9]. We refer to graphs \(F\) as motifs to emphasize that a graph sequence only converges when the densities of these structural motifs converge.

C. Graphons as Graph Generative Models

The other important interpretation of graphons is as generative models for graphs. A graph \(G_n = (V_n, \mathcal{E}_n, S_n)\) can be obtained from the graphon \(W\) in two steps. First, \(n\) points \(u_i \in [0, 1]\) are chosen to be the labels of the nodes \(i \in V_n\). Then, the edges \(\mathcal{E}_n\) and the adjacency matrix \(S_n\) are determined either by assigning weight \(W(u_i, u_j)\) to \((i, j)\), or by sampling the edge \((i, j)\) with probability \(W(u_i, u_j)\). There are many different strategies for choosing the node labels \(u_i\) and, given the possibility of defining either weighted (deterministic) or stochastic (unweighted) edges \(\mathcal{E}_n\), there are many types of graphs that can be generated in this way. We focus on three: template graphs, weighted graphs, and stochastic graphs.

Definition 1: (Template graphs). Let \(\{u_i\}_{i=1}^n\) be the regular \(n\)-partition of \([0, 1]\),

\[
u_i = \frac{i-1}{n}
\]

for \(1 \leq i \leq n\). The \(n\)-node template graph \(G_n\), whose GSO we denote \(S_n\), is obtained from \(W\) as

\[
[S_n]_{ij} = W(u_i, u_j) \text{ for } 1 \leq i, j \leq n.
\]

Template graphs are the simplest type of graph that can be generated from a graphon. The node labels \(u_i\) are determined by partitioning \([0, 1]\) into \(n\) intervals of equal length, and the edge weights \([S_n]_{ij}\) are obtained by evaluating the graphon at \((u_i, u_j)\). Template graphs are always weighted and, if \(W\) is strictly positive, they are also complete (i.e., all nodes are connected to every node of the graph).

Definition 2: (Weighted graphs). Let \(\{u_i\}_{i=1}^n\) be \(n\) points sampled independently and uniformly at random from \([0, 1]\),

\[
u_i \sim \text{unif}(0, 1)
\]

for \(1 \leq i \leq n\). The \(n\)-node weighted graph \(G_n\), whose GSO we denote \(S_n\), is obtained from \(W\) as

\[
[S_n]_{ij} = W(u_i, u_j) \text{ for } 1 \leq i, j \leq n.
\]

Weighted graphs are obtained by sampling \(u_i\) uniformly at random from the unit interval and assigning weights \(W(u_i, u_j)\) to edges \((i, j)\). As the name implies, these graphs are weighted and, like template graphs, they can also be complete.

Definition 3: (Stochastic graphs). Let \(\{u_i\}_{i=1}^n\) be \(n\) points sampled independently and uniformly at random from \([0, 1]\),

\[
u_i \sim \text{unif}(0, 1)
\]
for $1 \le i \le n$. The $n$-node stochastic graph $G_n$, whose GSO we denote $S_n$, is obtained from $W$ as

$$[S_n]_{ij} = [S_n]_{ji} \sim \text{Bernoulli}(W(u_i, u_j)) \quad \text{for } 1 \le i, j \le n.$$ 

Stochastic graphs are stochastic in both the node labels $u_i$, which are sampled uniformly at random, and in the edges $(i, j)$, which are sampled with probabilities $W(u_i, u_j)$. Also called $W$-random graphs [8, Ch. 10.1], these graphs are always unweighted, i.e., $S_n \in \{0, 1\}^{n \times n}$. Note that stochastic graphs can also be obtained from weighted graphs as

$$[S_n]_{ij} = [S_n]_{ji} \sim \text{Bernoulli}([S_n]_{ij}). \quad (9)$$

Thus, weighted graphs may be used as random graph models for $n$-node unweighted graphs.

We conclude this section by introducing the notion of a graphon induced by a graph, which will be useful in later derivations. Let $G$ be a graph with GSO $S \in \mathbb{R}^{n \times n}$ and node labels $\{u_i\}_{i=1}^n$, $u_i \in [0, 1]$. If the node labels are unknown, simply define $u_i = (i-1)/n$ for $1 \le i \le n$. Further define the intervals $I_i = [u_i, u_{i+1})$ for $1 \le i \le n-1$ and $I_n = [u_n, 1) \cup [0, u_1)$. The graphon $W_G$ induced by $G$ is given by

$$W_G(u, v) = \sum_{i=1}^n \sum_{j=1}^n [S]_{ij}I(u \in I_i)I(v \in I_j) \quad (10)$$

where $I$ is the indicator function.

### III. Transferability of Graph Filters

In the following, we discuss the transferability properties of graph filters. We begin by defining the limits of graph filters—graphon filters—in Section III-A. Then, in Section III-B we interpret graphon filters as generative models for graph filters, and derive an upper bound to the transferability or approximation error between a graphon filter and a graph filter with the same coefficients on a generic graph sampled from the graphon (Theorem 1). To illustrate this result with concrete examples, in Propositions 1–3 we particularize Theorem 1 to template, weighted and stochastic graphs. Sticking with the stochastic graph example, in Proposition 4 we then show that the error of transferring a filter across graphs associated with the same graphon can also be upper bounded as an immediate consequence of the graph filter approximation result.

#### A. Graphon Filters

The natural limit of signals supported on graphs that are sampled from or converge to a graphon is the graphon signal. A graphon signal is a function $X \in L^2([0, 1])$ and, like the graphon, it is both the limit of a convergent sequence of graph signals and a generative model for graph signals on graphs instantiated from a graphon [20]. For instance, given a template graph $G_0$ generated from the graphon $W$ as in Definition 1, the graphon signal $X$ can be used to generate the graph signal

$$[x_n]_i = X\left(\frac{i-1}{n}\right) \quad (11)$$

for $1 \le i \le n$. Similarly, on the weighted graph $G_n$ and on the stochastic graph $G_n$ [cf. Definitions 2 and 3], $X$ can be used to generate equivalent signals $x_n$ and $\hat{x}_n$ defined as

$$[x_n]_i = [x_n]_i = X(u_i) \quad (12)$$

for $1 \le i \le n$, where the $u_i$ are sampled independently and uniformly at random from $[0, 1]$.

Graphon signals can take any form, but the most straightforward ones are those induced by graph signals. The graphon signal induced by a graph signal $x_n \in \mathbb{R}^n$ is a step function given by [38]

$$X_n(u) = \sum_{i=1}^n [x_n]_i I(u \in I_i) \quad (13)$$

where $I_i = [u_i, u_{i+1})$ for $1 \le i \le n-1$ and $I_n = [u_n, 1) \cup [0, u_1)$. When the node labels are not defined, we set $u_i = (i-1)/n, i/n$ for $1 \le i \le n$.

The diffusion operator for graphon signals, denoted $T_W$, is the integral operator with kernel $W$, i.e.,

$$T_W X(v) = \int_0^1 W(u, v) X(u) du. \quad (14)$$

In analogy with the GSO, we refer to this operator as the graphon shift operator (WSO) of $W$.

Because graphons are bounded and symmetric, the WSO is a self-adjoint Hilbert-Schmidt (HS) operator. This allows expressing $W$ in the operator’s spectral basis as $W(u, v) = \sum_{i \in \mathbb{Z} \setminus \{0\}} \lambda_i \phi_i(u) \phi_i(v)$ and rewriting $T_W$ as

$$T_W X(v) = \sum_{i \in \mathbb{Z} \setminus \{0\}} \lambda_i \phi_i(v) \int_0^1 \phi_i(u) X(u) du \quad (15)$$

where the eigenvalues $\lambda_i, i \in \mathbb{Z} \setminus \{0\}$, follow the ordering $1 \ge \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n \ge \lambda_{n+1} \ge -1$. Since $T_W$ is compact, the eigenvalues accumulate around 0 as $|i| \to \infty$ [39, Thm. 3, Chapter 28]. This is illustrated in Fig. 2.

Similarly to how the graph eigenvalues are interpreted as graph frequencies, we interpret the graphon eigenvalues as graphon frequencies and the graphon eigenfunctions as the graphon’s oscillation modes. High magnitude eigenvalues represent the low frequencies, and vice-versa. Since the eigenfunctions $\phi_i$ form an orthonormal basis of $L^2([0, 1])$, we can define the graphon Fourier transform (WFT) of a graphon signal $X$ as its projection onto this eigenbasis. Explicitly, the WFT of $X$ is denoted $\hat{X}$ and given by [38]

$$[\hat{X}]_i = \int_0^1 X(u) \phi_i(u) du. \quad (16)$$

Conversely, the inverse graphon Fourier transform or iWFT is $X = \sum_{i \in \mathbb{Z} \setminus \{0\}} [\hat{X}]_i \phi_i$.

A graph filter is an operator $T_H : L^2([0, 1]) \to L^2([0, 1])$ which maps graphon signals $X$ to $Y = T_H X$. In particular, the graphon shift operator allows defining LSI graphon filters given by

$$T^{(k)}_H X = \sum_{k=0}^{K-1} h_k T^{(k)}_W X \quad \text{with}$$

$$T^{(k)}_W X = \int_0^1 W(u, v) T^{(k-1)}_W X(u) du \quad (17)$$
where \( T_{W}^{(0)} = I \) is the identity operator and \( h = [h_0, \ldots, h_{K-1}] \) are the filter coefficients. This filter is shift-invariant because \( T_{H} \) and \( T_{W} \) commute. I.e., if we shift the input as \( X' = T_{W}X \), the output is equivalently shifted as \( Y' = T_{H}X' = T_{H}T_{W}X = T_{W}T_{H}X = T_{W}Y \). LSI graph filtering is also called graphon convolution [20].

Using the spectral decomposition of \( T_{W} \) in (15), \( T_{H} \) can be written as

\[
T_{H}X(u) = \sum_{i \in \mathcal{V}(0)} \sum_{k=0}^{K-1} h_k \lambda_i^k \phi_i(v) \int_0^1 \phi_i(u)X(u)du
\]

This expression highlights the filter’s spectral representation, which is given by \( h(\lambda) = \sum_{k=0}^{K-1} h_k \lambda^k \). Because the WFT of the filter is a polynomial on the graphon eigenvalues, note that as \( K \to \infty \) LSI graphon filters can be used to approximate any graphon filter with spectral representation \( h(\lambda) = f(\lambda) \) where \( f \) is analytic. This is a consequence of the Cayley-Hamilton theorem in the limit of matrices. Moreover, unlike the WFT of the signals \( X \) and \( Y \), the spectral representation of the LSI graphon filter does not depend on the graphon eigenvalues—only on the graphon eigenvalues and the coefficients \( h \). This is illustrated in Fig. 2 where the red curve represents \( h(\lambda) \) and the blue lines are the eigenvalues of \( W \).

The fact that the spectral response of the LSI graphon filter only depends on the eigenvalues of the graphon and on the filter coefficients is something that these filters have in common with LSI graph filters. This can be seen by comparing (3) and (18). Indeed, if the coefficients \( h_k \) in these equations are the same, the spectral responses of the LSI graph filter and of the LSI graphon filter are determined by the same function \( h(\lambda) \). Given a graph \( G_n \) (with GSO \( S_n \)) and a graphon \( W \), the only thing that changes is where this function is evaluated: at the graphon eigenvalues \( \{\lambda_i(S_n)\}_{i \in \mathcal{V}(0)} \), or at the graphon eigenvalues \( \{\lambda_i(T_{W})\}_{i \in \mathcal{V}(0)} \) [cf. Fig. 2]. The implication is that graphon filters can be used as generative models for graph filters as long as the coefficients \( h_k \) in (3) and (18) are the same.

For example, let \( Y = T_{H}X \) be the LSI graphon filter which we represent, for simplicity, as the map \( Y = \Phi(X; h, W) \) to emphasize the dependence on the coefficients \( h \) and the graphon \( W \). We can generate a graph filter \( y_n = \Phi(x_n; h, S_n) \) from this graph filter by instantiating a graph \( G_n \) from \( W \) as in Definitions 1, 2 or 3, and the corresponding graph signal \( x_n \) as in (11) or (12). Generating graph filters from graphon filters is useful because it allows designing filters for graphons and transferring them to graphs. Since graphons are limits of convergent graph sequences, it also justifies transferring filters across graphs obtained from the same graphon.

Before delving into the transferability analysis of graph filters, we introduce the definition of graphon filters induced by graph filters, which will allow comparing graph and graphon filters directly. The graphon filter induced by the graph filter \( \Phi(x_n; h, S_n/n) = \sum_{k=0}^{K-1} h_k S_n^k x_n/n^k \) is given by

\[
\Phi(X_n; h, W_n) = \sum_{k=0}^{K} h_k T_{W_n}^{(k)} X_n(v) \quad \text{with}
\]

\[
T_{W_n}^{(k)} X_n(v) = \int_0^1 W_n(u,v)T_{W_n}^{(k-1)} X_n(u)du \quad (19)
\]

where the graphon \( W_n = W_{G_n} \) is the graphon induced by \( G_n \) (10) and \( X_n \) is the graphon signal induced by the graph signal \( x_n \) (13). Note that in the graph filter \( \Phi(x_n; h, S_n/n) \), we consider the normalized shift operator \( S_n/n \). This is important to guarantee that the output of the graph filter and its corresponding induced graphon filter match, as shown by the following lemma. The proof is deferred to Appendix B in the supplementary material.

**Lemma 1**: Let \( y_n = \Phi(x_n; h, S_n/n) = \sum_{k=0}^{K-1} h_k S_n^k x_n/n^k \) be a graph filter, and let \( y_n = \Phi(X_n; h, W_n) \) be its corresponding induced graphon filter (19). Then, it holds that

\[
y_n(u) = \sum_{i=1}^{n} y_i I_i(u) \quad (20)
\]

where \( I_i = [u_i, u_{i+1}] \) for \( 1 \leq i \leq n-1 \) and \( I_n = [u_n, 1] \cup [0, u_1] \), and \( u_i \) are the node labels associated with \( G_n \). When the node labels are not defined, \( u_i = (i - 1)/n \) for \( 1 \leq i \leq n \).
B. Graph Filter Transferability

The graph filter (and GNN) transferability results require introducing two definitions: the $c$-band cardinality and the $c$-eigenvalue margin, which are both graphon spectral properties associated with a fixed eigenvalue threshold $c$.

Definition 4: ($c$-band cardinality of $W$). The $c$-band cardinality of a graphon $W$, denoted $B_{cW}$, is the number of eigenvalues $\lambda_i$ of $T_W$ with absolute value larger or equal to $c$, i.e.,

$$B_{cW} = \#\{\lambda_i : |\lambda_i| \geq c\}.$$  

Definition 5: ($c$-eigenvalue margin of $W$ and $W'$). For two graphons $W$ and $W'$, the $c$-eigenvalue margin, denoted $\delta_{WW'}$, is given by

$$\delta_{WW'} = \min_{i,j \neq i} \{\lambda_i(T_{W'}) - \lambda_j(T_W) : |\lambda_i(T_{W'})| \geq c\}$$

where $\lambda_i(T_{W'})$ and $\lambda_j(T_W)$ denote the eigenvalues of $T_{W'}$ and $T_W$ respectively.

We also introduce the following Lipschitz continuity assumptions on the filter spectral response.

$AS1$: The spectral response of the convolutional filter $h$ is $A_h$-Lipschitz in $[-1, -c] \cup [c, 1]$ and $A_h$-Lipschitz in $(-c, c)$, with $A_h < A$. Moreover, $|h(\lambda)| < 1$.

Under assumption $AS1$, we can prove the following theorem, which states that a graphon filter can be approximated by a graph filter on a large graph sampled from this graphon.

Theorem 1: (Graphon filter transferability to a generic graph). Let $Y = \Phi(X; h, W)$ be a graphon filter [cf. (18)] satisfying assumption $AS1$. Given a generic graph $G_n$ with GSO $S_n$ sampled from $W$, let $y_n = \Phi(x_n; h, S_n/n)$ be a graph filter instantiated from $\Phi(X; h, W)$ on this graph. For any $0 < c \leq 1$, it holds that

$$\|y_n - Y\| \leq (A_h + \frac{\pi B_{WW_n}}{\delta_{WW_n}}) \|W - W_n\| \|X\| + (A_h + 2)\|X - X_n\| + 2a_h c \|X\|$$

(21)

where $W_n := W_{S_n}$ is the graphon filter induced by $G_n$ and $Y_n = \Phi(x_n; h, W_{S_n})$ is the graphon filter induced by $y_n = \Phi(x_n; h, S_n/n)$ (19).

Proof: Refer to Appendix A.

Thm. 1 shows that the error incurred when transferring a graph filter to a graph sampled from the graphon is upper bounded by the sum of three terms. The first depends on the distance between the graph and the graphon, i.e., it is the graph sampling error. Since sequences of graphs sampled from a graphon converge to it, we can expect this distance to be small for large $n$, and to eventually vanish as $n \to \infty$. Disregarding the norm of the input signal $X$ (it is finite), this term also depends on what we call the filter’s transferability constant. By being proportional to $A_h$ and $B_{WW_n}$, and inversely proportional to $\delta_{WW_n}$, this constant is directly proportional to the filter variability, so the smaller it is, the more transferable the filter. Note that although the bound in Thm. 1 is somewhat implicit due to the dependence of $\delta_{WW_n}$ on $n$, for a given graph $G_n$ and graphon $W_n$, this quantity can always be calculated, and for large $n$ it approaches the smallest graphon eigengap in the interval $[-1, -c] \cup [c, 1]$.

The second term of the bound is related to the distance between the graphon signal and the graph signal. i.e., it stems from sampling the signal. Since $X$ is in $L^2$, $X_n$ converges up to node permutations. Hence, $\|X_n - X\|$ vanishes as $n \to \infty$. Note that this term also depends on $a_h c$, which upper bounds the maximum filter amplification in a part of the eigenvalue spectrum determined by a threshold $c$—the band $(-c, c)$, see Fig. 2. Since $a_h < A$, the contribution of this term to the approximation error is small.

The constant $c$ also appears in the third term of the bound, $2a_h c \|X\|$, which is related to the output signal’s non-transferable energy for eigenvalues smaller than $c$. For a given value of $c$, spectral components associated with $\lambda \in [-1, -c] \cup [c, 1]$ (i.e., low-frequency components) are completely transferable in the sense that, if the signal $X$ or the filter $\Phi(X; h, W)$ are $c$-bandlimited (i.e., spectral components associated with eigenvalues $|\lambda| \in [0, c]$ are zero), the transferability bound is only given by the first two terms of (22) and thus vanishes with the size of the graph. In the more general case in which the input signal is not bandlimited, regardless of how we fix $c \in (0, 1]$ there will be some residual non-transferable energy associated with the spectral components $|\lambda| \in [0, c]$. However, the effect of the non-transferable spectral components on the transferability error is attenuated by the filter since its Lipschitz constant on the $(-c, c)$ interval, $a_h$, is smaller than $A_h$, the Lipschitz constant on $[-1, -c] \cup [c, 1]$. In particular, if $a_h \ll A_h$ the filter resembles a filter with constant frequency response on the $(-c, c)$ band. See Section IV-C for more in-depth discussion about this non-transferable energy and filters with constant band.

Next, we show how Thm. 1 particularizes to the simplest type of graph that can be sampled from a graphon—the template graph in Definition 1. To do so, we need two additional assumptions on the Lipschitz continuity of the graphon and the graphon signal respectively.

$AS2$: The graphon $W$ is $A_w$-Lipschitz, i.e., $|W(u_2, v_2) - W(u_1, v_1)| \leq A_w(|u_2 - u_1| + |v_2 - v_1|)$.

$AS3$: The graphon signal $X$ is $A_x$-Lipschitz.

Proposition 1: (Graphon filter transferability to a template graph). Let $Y = \Phi(X; h, W)$ be a graphon filter [cf. (18)] satisfying assumptions $AS1$–$AS3$. Given a template graph $G_n$ with GSO $S_n$ [cf. Definition 1], let $y_n = \Phi(x_n; h, S_n/n)$ be a graph filter instantiated from $\Phi(X; h, W)$ on this graph. For any $0 < c \leq 1$, it holds that

$$\|y_n - Y\| \leq (A_h + \frac{\pi B_{WW_n}}{\delta_{WW_n}}) \frac{2A_w}{n} \|X\| + A_x (a_h + 2) \frac{\|X\|}{n} + 2a_h c \|X\|$$

(22)

where $W_n = W_{S_n}$ is the graphon filter induced by $G_n$ and $y_n = \Phi(x_n; h, S_n/n)$ is the graphon filter induced by $y_n = \Phi(x_n; h, S_n/n)$ (19).

Proof: Refer to Appendix C in the supplementary material.
From Prop. 1, we see that, for template graphs, the parts of the graph filter approximation bound associated with the graph and the graph signal sampling errors decrease as $n^{-1}$. However, the dependence on the Lipschitz constants $A_w$ and $A_x$ implies that graph filters are harder to approximate when the graphon and the graphon signal are not smooth.

Using concentration inequalities for the uniform distribution, Prop. 1 can be extended to weighted graphs. First, we need an additional definition—node stochasticity—which arises when the node labels are no longer given by a regular partition, but sampled uniformly at random.

**Definition 6:** (Node stochasticity). For a fixed probability $\chi \in [0, 1]$, the node stochasticity constant on $n$ nodes, denoted $\alpha(\chi, n)$, is defined as

$$\alpha(n, \chi) = \log ((n + 1)^{\frac{2}{\chi}} / \log (1 - \chi)^{-1}).$$

**Proposition 2:** (Graphon filter transferability to a weighted graph). Let $Y = \Phi(X; h, W)$ be a graphon filter [cf. (18)] satisfying assumptions AS1–AS3. Given a weighted graph $\tilde{G}_n$ with GSO $S_n$ [cf. Definition 2], let $y_n = \Phi(x_n; h, S_n/n)$ be a graph filter instantiated from $\Phi(X; h, W)$ on this graph. Let $\chi_1, \chi_2 \in (0, 0.3]$. For any $0 < c \leq 1$ and $n \geq 4/\chi_2$, with probability at least $[1 - 2\chi_1] \times [1 - \chi_2]$ it holds that

$$\|Y_n - Y\| \leq \left(A_h + \frac{\pi B_{\tilde{G}_n}}{\delta_{\tilde{G}_n}}\right)\frac{2A_w\alpha(n, \chi_1)}{n}\|X\| + \frac{A_x\alpha(n, \chi_1)(a_h c + 2)}{n} + 2a_h c\|X\| \quad (23)$$

where $\tilde{w}_n = W_{\tilde{G}_n}$ is the graphon induced by $\tilde{G}_n$ and $\tilde{Y}_n = \Phi(x_n; h, \tilde{S}_n/n)$ is the graph filter induced by $y_n = \Phi(x_n; h, S_n/n)$ (19).

**Proof:** Refer to Appendix C in the supplementary material.

Therefore, graphon filters are also transferable to weighted graphs. The main difference between Props. 1 and 2 is that the first and second terms of the transferability bound are now multiplied by the node stochasticity constant $\alpha(n, \chi_1)$ [cf. Definition 6]. This constant shows up because of the randomness associated with the node labels $u$, which are sampled uniformly at random. Note that $\alpha(n, \chi_1)$ depends on both the graph size and $\chi_1$. The value of $\chi_1$ determines the confidence of the transferability bound. This confidence, given by $[1 - 2\chi_1] \times [1 - \chi_2]$, also depends on the parameter $\chi_2$. Although $\chi_2$ does not appear in (23), there is an interplay between the confidence bound and the minimum graph size, since Prop. 2 holds with probability $[1 - 2\chi_1] \times [1 - \chi_2]$ for $n \geq 4/\chi_2$.

Next, we extend the graphon-graph filter approximation result to its more general form—graphon filter approximation by graph filters supported on stochastic graphs. To do so, we need to define edge stochasticity and make one additional assumption on the size of the graph.

**Definition 7:** (Edge stochasticity). For a fixed probability $\chi \in [0, 1]$, the edge stochasticity constant on $n$ nodes, denoted $\beta(\chi, n)$, is defined as

$$\beta(n, \chi) = \sqrt{n \log (2n/\chi)}.$$

Assumption 4 imposes a restriction on $n$ related to the graphon variability $A_w$ and its maximum degree $d_W$.

**AS4:** Given $\chi_3 \in (0, 1)$, $n$ is such that

$$n - \frac{\log 2n/\chi_3}{d_W} > 2A_w$$

where $d_W = \max_u \int_0^1 W(u, v) dv$.

We also need the following lemma, which upper bounds the transferability error of a graph filter across a weighted graph and a stochastic graph that is sampled from it (see (9)).

**Lemma 2:** (Graph filter transferability from weighted to stochastic graphs). Consider a graphon $W$ satisfying assumption AS2, and let $\tilde{G}_n$ be a weighted graph with GSO $S_n$ [cf. Definition 2] and $G_n$ a stochastic graph with GSO $S_n$ obtained from $S_n$ as in (9). Given a graph signal $X$ satisfying assumption AS3, let $y_n = \Phi(x_n; h, S_n/n)$ and $\tilde{y}_n = \Phi(x_n; h, S_n/n)$ be graph filters satisfying assumption AS1 and acting on the graph signals $x_n$ and $\tilde{x}_n$ instantiated from $X$ on $G_n$ and $\tilde{G}_n$ respectively (note that $x_n = x_{x,n}$). Let $\chi_3 \in (0, 1)$. For any $0 < c \leq 1$ and $n$ satisfying assumption AS4, with probability at least $1 - \chi_3$ it holds that

$$\|\tilde{y}_n - y_n\| \leq \left(A_h + \frac{\pi B_{\tilde{G}_n}}{\delta_{\tilde{G}_n}}\right)\frac{2\sqrt{\log (2n/\chi_3)}}{n}\|x_n\| + 2a_h c\|x_n\|$$

where $\tilde{w}_n = W_{\tilde{G}_n}$ and $W_{\tilde{n}} = W_{G_n}$ are the graphons induced by $G_n$ and $\tilde{G}_n$ respectively.

**Proof:** Refer to Appendix D in the supplementary material.

The graphon filter approximation result is obtained by combining Prop. 2 and Lemma 2 through the triangle inequality.

**Proposition 3:** (Graphon filter transferability to a stochastic graph). Let $Y = \Phi(X; h, W)$ be a graphon filter [cf. (18)] satisfying assumptions AS1–AS3. Given a stochastic graph $G_n$ with GSO $S_n$ [cf. Definition 3], let $y_n = \Phi(x_n; h, S_n/n)$ be a graph filter instantiated from $\Phi(X; h, W)$ on this graph. Let $\chi_1, \chi_2, \chi_3 \in (0, 0.3]$. For any $0 < c \leq 1$ and $n \geq 4/\chi_2$ satisfying assumption AS4, with probability at least $[1 - 2\chi_1] \times [1 - \chi_2] \times [1 - \chi_3]$ it holds that

$$\|Y_n - Y\| \leq \left(A_h + \frac{\pi B_{G_n}}{\delta_{G_n}}\right)\frac{2A_w\alpha(n, \chi_1) + \beta(n, \chi_3)}{n}\|X\| + \frac{A_x\alpha(n, \chi_1)(a_h c + 2)}{n} + 4a_h c\|x_n\|$$

where $W_n = W_{G_n}$ is the graphon induced by $G_n$ and $y_n = \Phi(x_n; h, W_n)$ is the graph filter induced by $y_n = \Phi(x_n; h, S_n/n)$ (19).

**Proof:** Prop. 3 follows directly from Lemma 2, Prop. 2, and the triangle inequality. Note that the probabilities $[1 - \chi_1] \times [1 - \chi_2]$ and $1 - \chi_3$ are multiplied because Lemma 2 holds for any weighted graph $G_n$, i.e., it is independent of the weighted graph.

This approximation bound is similar to the approximation bound derived for weighted graphs in Prop. 2, with two important differences. The first is that, in addition to depending on the node stochasticity constant $\alpha(n, \chi_1)$, the transferability constant also depends on the edge stochasticity $\beta(n, \chi_3)$ [cf. Definition 7] which accounts for the randomness of the edges of $G_n$. Besides modifying the value of the transferability constant for stochastic
graphs, note that \(\beta(n, \chi_3)\) also lowers the confidence of the bound. The second difference is that the part of the bound due to non-transferable spectral components is twice that of Prop. 2. This is due to the summation of the non-transferable energy between the graphon and the weighted graph (Prop. 2), and between the weighted graph and the stochastic graph (Lemma 2).

Since a graphon identifies a family of graphs, the graph filter approximation result in Thm. 1 and Propositions 1–3 can be readily extended to a graph filter transferability result. To stick with the same example and avoid repetition, herein we only show transferability of graph filters across stochastic graphs. Let \(G_{\tilde{n}_1}\) and \(G_{\tilde{n}_2}\), \(n_1 \neq n_2\), be two stochastic graphs with GSOs \(S_{\tilde{n}_1}\) and \(S_{\tilde{n}_2}\) respectively [cf. Definition 3], and let \(y_{\tilde{n}_1} = \Phi(x_{\tilde{n}_1}; h, S_{\tilde{n}_1}/n)\) and \(y_{\tilde{n}_2} = \Phi(x_{\tilde{n}_2}; h, S_{\tilde{n}_2}/n)\) be graph filters instantiated from \(\Phi(X; h, W)\) on these graphs. Let \(\chi_1, \chi_2, \chi_3 \in (0, 0.3]\). For any \(0 < c \leq 1\) and \(n_1, n_2 \geq 4/\chi_2\) satisfying assumption AS4, with probability at least \([1 - 2\chi_1]^2 \times [1 - \chi_2]^2 \times [1 - \chi_3]^2\) it holds that

\[
\|Y_{\tilde{n}_1} - Y_{\tilde{n}_2}\| \leq 4 \left( A_h + \frac{\pi B_{\text{max}}^c}{\delta_{\text{min}}^c} \right) \max_{i \in \{1, 2\}} \frac{A_n \alpha(n_i, \chi_1) + \beta(n_i, \chi_3)}{n_i} \|X\| \\
+ 2 A_x (a_h c + 2) \max_{i \in \{1, 2\}} \frac{\alpha(n_i, \chi_1)}{n_i} + 8 a_h c \|X\|
\]

where \(B_{\text{max}}^c = \max_{i \in \{1, 2\}} \max (B_{W_{\tilde{n}_1}}^c, B_{W_{\tilde{n}_2}}^c)\) and \(\delta_{\text{min}}^c = \min_{i \in \{1, 2\}} \min (\delta_{W_{\tilde{n}_1}}^c, \delta_{W_{\tilde{n}_2}}^c, \delta_{W_{\tilde{n}}}^c)\).

Proof: Prop. 4 follows from Prop. 3 and the triangle inequality.

Graph filters are thus transferable across graphs in the same graphon family. Since stochastic graphs have random edges, Prop. 4 can be applied to graphs of same size. If \(n_1 \neq n_2\), the transferability bound is slightly loose because it is a simplification of the sum of the graphon filter approximation bounds for \(G_{\tilde{n}_1}\) and \(G_{\tilde{n}_2}\). This simplification is however useful as it shows that the transferability error of graph filters is dominated by the graph with the largest ratio of node/edge stochasticity to graph size, typically the smallest graph.

Transferability of graph filters is an important result because it means that we can design a filter for one graph and transfer it to another. This is possible even if the graphs have different sizes, which considerably simplifies the linear processing of data supported on large-scale graphs. In the following, we will show that graph filter transferability also enables large-scale graph machine learning via GNNs. This is because the transferability properties of graph filters are not only inherited, but also augmented by GNNs.

Remark 1: (Graph filter transferability on template and weighted graphs). The graph filter transferability result for stochastic graphs in Prop. 4 also holds for weighted and template graphs. Simply set \(\beta(n, \chi_3) = 0\) for template and weighted graphs and \(\alpha(n, \chi_1) = 1\) for template graphs. Tighter bounds can be achieved by combining the triangle inequality with Propositions 1 and 2 respectively, in the same way that Prop. 3 was used to show Prop. 4.

IV. TRANSFERABILITY OF GRAPH NEURAL NETWORKS

In the following, we discuss the transferability properties of GNNs. We begin by defining WNNs in Section IV-A. Then, in Section IV-B we interpret WNNs as generative models for GNNs, and prove Thm. 2, which derives an upper bound for the transferability or approximation error between a WNN and a GNN sampled from it on a generic graph sampled from the graphon. The particularization of this theorem to stochastic graphs is stated in Prop. 5; finally, in Prop. 6, we show how Prop. 5 implies transferability of GNNs across stochastic graphs sampled from the same graphon.

A. Graphon Neural Networks

A graphon neural network (WNN) is a deep convolutional architecture consisting of layers where each layer implements a convolutional filterbank followed by a pointwise nonlinearity [19]. Consider layer \(\ell\), which maps the incoming \(F_{\ell-1}\) features from layer \(\ell-1\) into \(F_\ell\) features. The first step in this layer is to process the features \(X^\ell_{\ell-1}\), \(1 \leq \ell \leq F_\ell\), with a convolutional filterbank to generate the \(F_\ell\) intermediate linear features \(U^\ell_g\),

\[
U^\ell_g = \sum_{g=1}^{F_{\ell-1}} T_{H^g} X^\ell_{\ell-1}
\]

for \(1 \leq g \leq F_{\ell-1}\). Each intermediate feature \(U^\ell_g\) is obtained by aggregating the outputs of \(F_{\ell-1}\) filters like the one in (17) with coefficients \(h^g\). Since there are \(F_\ell\) such intermediate features, the filterbanks at each layer of the WNN contain a total of \(F_\ell \times F_{\ell-1}\) convolutional filters.

The next step is to process the intermediate features \(U^\ell_g\) with a pointwise nonlinearity, e.g., the ReLU. Denoting this nonlinearity \(\sigma\), the \(f\)th feature of the \(\ell\)th layer is given by

\[
X^\ell_f = \sigma \left( U^\ell_g \right)
\]

for \(1 \leq f \leq F_\ell\). Because the nonlinearity is pointwise, \(X^\ell_f(u) = \sigma(U^\ell_g(u))\) for all \(u \in [0, 1]\). If the WNN has \(L\) layers, (24)–(25) are repeated \(L\) times. The input features at the first layer, \(X^0_g\), are the input data \(X^0\) for \(1 \leq g \leq F_0\), and the WNN output is given by \(Y^f = X^\ell_f\) for \(1 \leq f \leq F_L\).

Similarly to the graphon filter, a WNN with inputs \(X = \{X^0\}_{f=1}^{F_0}\) and outputs \(Y = \{Y^f\}_{f=1}^{F_L}\) can be represented more compactly as the map \(Y = \Phi(X; H, W)\), where \(H\) is a tensor grouping the coefficients \(h^g\) for all features and all layers of the WNN, i.e., \(H = [h^g_{f\ell}]_{f\ell g}\) for \(1 \leq \ell \leq L, 1 \leq f \leq F_\ell\) and \(1 \leq g \leq F_{\ell-1}\). Comparing this map with the GNN map.
in (7), we see that, except for the fact that their supports—a graphon and a graph respectively—are different, if the tensors $\mathcal{H}$ are equal, these maps are the same. This allows interpreting WNNs as generative models for GNNs where the graph $G_n$ is instantiated from $W$ as in Definitions 1, 2 or 3, and the graph signal $x_n$, is instantiated from $X$ as in (11) or (12).

The interpretation of WNNs as generative models for GNNs is important for two reasons. First, it allows designing one WNN and instantiating as many GNNs as desired from it. I.e., it allows designing neural networks in the limit of very large graphs and transferring them to finite graphs without changes to the architecture. Second, it motivates analyzing the ability to transfer GNNs across graphs of same or different size, since a graphon sampled from the graphon $(10)$ and pre-interpreting $H$ in (7), we see that, except for the fact that their supports—a graphon and a graph respectively—are different, if the tensors $\mathcal{H}$ are equal, these maps are the same. This allows interpreting WNNs as generative models for GNNs where the graph $G_n$ is instantiated from $W$ as in Definitions 1, 2 or 3, and the graph signal $x_n$, is instantiated from $X$ as in (11) or (12).

To be able to compare WNNs with GNNs, or two GNNs supported on graphs of different sizes, we will need to define WNNs induced by GNNs. The WNN induced by a GNN $y_n = \Phi(x_n; \mathcal{H}, S_n/n)$ on the graph $G_n$ is given by

$$y_n = \Phi(X_n; \mathcal{H}, W_n)$$

(26)

where the graphon $W_n = W_{G_n}$ is the graphon induced by $G_n$ (10) and $x_n$ is the graphon signal induced by the graph signal $x_n$ (13). Note the normalization by $n$, which is necessary to have $y_n(u) = \sum_{i=1}^{n} |y_n|(u_i)$, see Lemma 1.

B. Graph Neural Network Transferability

Consider a WNN with $L$ layers, $F_0 = 1$ input feature, $F_L = 1$ output feature, and $F_i = F$, $1 \leq i \leq L - 1$ under a Lipschitz continuity assumption on the nonlinearity $\sigma$, this WNN can be approximated by a GNN on a graph $G_n$ sampled from the graphon $W$.

AS5: The activation functions are normalized Lipschitz, i.e., $|\sigma(x) - \sigma(y)| \leq |x - y|$, and $\sigma(0) = 0$.

Theorem 2: (WNN transferability to a generic graph). Let $Y = \Phi(X; \mathcal{H}, W)$ be a WNN with $L$ layers, $F_0 = 1$ input and output features and $F_L = F$, $1 \leq i \leq L$ [cf. (24)–(25)]. Assume that the convolutional filters that make up its layers all satisfy assumption AS1. Given a generic graph $G_n$ with GSO $S_n$ sampled from $W$, let $y_n = \Phi(x_n; \mathcal{H}, S_n/n)$ be a GNN instantiated from $\Phi(X; \mathcal{H}, W)$ on this graph. For any $0 < c < 1$, it holds that

$$\|Y_n - Y\| \leq LF_{L-1}(A_h + \frac{\pi B_p}{\delta_{WW_n}})\|W - W_n\|\|X\|$$

$$+ (A_h c + 2)\|X - X_n\| + LF_{L-2}a_h c\|X\|$$

(27)

where $W_n := W_{G_n}$ is the graphon induced by $G_n$ and $y_n = \Phi(x_n; \mathcal{H}, W_n)$ is the WNN induced by $y_n = \Phi(x_n; \mathcal{H}, S_n/n)$ (26).

Proof: Refer to Appendix E in the supplementary material.

This theorem follows recursively from Thm. 1 because, provided that Assumption AS5 is met, $|\sigma(U_t + \Delta U_t) - \sigma(U_t)|$ can be upper bounded by $|\Delta U_t|$. Most common activation functions, e.g., the hyperbolic tangent and the ReLU, satisfy this assumption. Thus, we can bound the error incurred of approximating WNN with a GNN by the transferability error of a cascade of $L$ graphon filterbanks.

Akin to the graphon filter approximation bound, the approximation bound in Thm. 2 has three terms: the first, which is controlled by the transferability constant and the distance between the graph and the graphon; the second, which is controlled by the distance between the graph signal and the graphon signal; and the third, which corresponds to the non-transferable energy of the input signal. Because it arises from the sampling error of the input signal, the second error term is exactly the same as in Thm. 1, while the others differ by a scaling factor of $LF_{L-1}$. Hence, the deeper and the wider the WNN, the more difficult it is to approximate it with a GNN.

In Prop. 5, we give a concrete example of how a WNN may be approximated by a GNN on a graph sampled from the graphon. For brevity, we only particularize the approximation result of Thm. 2 to the more general case of the stochastic graphs in Definition 3, but it is easy to write the same result for template and weighted graphs (see Remark 2).

Proposition 5: (WNN transferability to a stochastic graph). Let $Y = \Phi(X; \mathcal{H}, W)$ be a WNN with $L$ layers, $F_0 = F_L = 1$ input and output features and $F_i = F$, $1 \leq i \leq L$ [cf. (24)–(25)]. Assume that this WNN satisfies assumptions AS2, AS3, and AS5, and that the convolutional filters that make up its layers all satisfy assumption AS1. Given a stochastic graph $G_n$ with GSO $S_n$ [cf. Definition 3], let $y_n = \Phi(x_n; \mathcal{H}, S_n/n)$ be a GNN instantiated from $\Phi(X; \mathcal{H}, W)$ on this graph. Let $\chi_1, \chi_2, \chi_3 \in (0, 0.3)$. For any $0 < c < 1$ and $n \geq \frac{4}{\chi_2}$ satisfying assumption AS4, with probability at least $[1 - 2\chi_1] \times [1 - \chi_2] \times [1 - \chi_3]$ it holds that

$$\|Y_n - Y\| \leq LF_{L-1}(A_h + \frac{\pi B_p}{\delta_{WW_n}})\|W - W_n\|\|X\|$$

$$+ (A_h c + 2)\|X - X_n\| + LF_{L-2}a_h c\|X\|$$

(28)

where $W_n = W_{G_n}$ is the graphon induced by $G_n$ and $y_n = \Phi(x_n; \mathcal{H}, W_n)$ is the WNN induced by $y_n = \Phi(x_n; \mathcal{H}, S_n/n)$ (26).

Proof: The proof follows the same steps of the proof of Thm. 2 in Appendix E of the supplementary material, using Prop. 3 instead of Thm. 1 in (67).

From Prop. 5, it is ready to show that GNNs are transferable across stochastic graphs.

Proposition 6: (GNN transferability across stochastic graphs). Let $Y = \Phi(X; \mathcal{H}, W)$ be a WNN [cf. (7)] satisfying assumptions AS2, AS3, AS5, and such that the convolutional filters at all layers satisfy assumption AS1. Let $G_{\alpha_1}$ and $G_{\alpha_2}$, $n_1 \neq n_2$, be two stochastic graphs with GSOs $S_{\alpha_1}$ and $S_{\alpha_2}$ respectively [cf. Definition 3], and let $y_{\alpha_1} = \Phi(x_{\alpha_1}; \mathcal{H}, S_{\alpha_1}/n)$ and $y_{\alpha_2} = \Phi(x_{\alpha_2}; \mathcal{H}, S_{\alpha_2}/n)$ be GNNs instantiated from $\Phi(X; \mathcal{H}, W)$ on these graphs. Let $\chi_1, \chi_2, \chi_3 \in (0, 0.3)$. For any $0 < c < 1$ and $n_1, n_2 \geq \frac{4}{\chi_2}$...
satisfying assumption AS4, with probability at least \([1 - 2\chi_1^2] \times [1 - \chi_2^2] \times [1 - \chi_3^2]\) it holds that
\[
\|Y_{n_1} - Y_{n_2}\| \leq 4L^E\|A_h + \pi\delta_c^\text{max}\| \times \max_{i \in \{1, 2\}} \frac{\delta_c^i}{1 - \alpha(n_i, \chi_1) + \beta(n_i, \chi_3)} \|X\|
+ 2A_j(a_i c + 2) \max_{i \in \{1, 2\}} \frac{\alpha(n_i, \chi_1)}{n_i}
+ 8L^E\|a_i c\| \|X\|
\]
where \(B_c^\text{max} = \max_{i \in \{1, 2\}} \max(B_{c_W h_i}^\text{max}, B_{c_N h_i}^\text{max})\) and \(\delta_c^\text{min} = \min_{i \in \{1, 2\}} \min(\delta_{c_W h_i}, \delta_{c_N h_i}).\)

**Proof:** Prop. 6 follows directly from Prop. 5 and the triangle inequality. 

When \(n_1 = n_2\), the GNN transferability bound in Prop. 6 is approximately twice the WNN approximation bound in Prop. 5. When \(n_1 \neq n_2\), this bound can be improved by explicitly writing the sum of (28) for \(G_{n_1}\) and \(G_{n_2}\); we take the maximum to highlight that the error is dominated by the inverse of the size of smallest graph. Although Prop. 6 is explicitly written for stochastic graphs, it also holds for weighted and template graphs associated with the same graphon; see Remark 2.

The main implication of Prop. 6 is that a GNN can be trained on one graph to be transferred to another graph, which is especially helpful when the graph on which we want to execute the GNN is large and training on it is costly. In these cases, we can use the bound in Prop. 6 to determine the minimum size of the graph on which the GNN should be trained to meet a given error allowance. The transferability property of the GNN thus makes it a suitable model for machine learning on large-scale graphs.

**Remark 2:** The GNN transferability result derived for stochastic graphs in Prop. 6 holds for weighted graphs by setting \(\beta(n, \chi_3) = 0\) for weighted and template graphs and \(\alpha(n, \chi_1) = 1\) for the latter. However, the resulting bounds are slightly loose. For tighter bounds, combine the triangle inequality with Props. 1 and 2 respectively to obtain the graph filter transferability bounds for template and stochastic graphs. Then, follow the same proof steps used to show Thm. 2 from Thm. 1 in Appendix E of the supplementary material. The graph filter transferability result derived for stochastic graphs in Prop. 4 also holds for weighted and template graphs. Simply set \(\beta(n, \chi_3) = 0\) for both types of graphs and \(\alpha(n, \chi_1) = 1\) for template graphs. Tighter bounds can be achieved by combining the triangle inequality with Props. 1 and 2 respectively, in the same way that Prop. 3 was used to show Prop. 4.

### C. Discussion

Non-transferable energy, filters with constant band, and asymptotics. In order to be transferable, filters and GNNs have to be able to “match” the eigenvalues of the source graph with those of the target graph so that the amplifications of the output signal’s spectral components match on both graphs. This is possible because, as illustrated in Fig. 2, as \(n \to \infty\) the eigenvalues of a graph \(G_n\) sampled from a graphon \(W\) converge to the graphon eigenvalues [8, Chapter 11.6]. Hence, for large enough \(G_{n_1}\) and \(G_{n_2}\), their eigenvalues are close. However, because the graphon eigenvalues accumulate near zero, for small eigenvalues (i.e., for \(\lambda_j\) such that \(|j| \to \infty\)), this matching becomes very hard. This is due to the fact that the distance between the eigenvalues \(\lambda_j(G_{n_1})\) and \(\lambda_j(G_{n_2})\) might be larger than the distance between consecutive eigenvalues in this range. To mitigate this problem, as shown in Fig. 2 we restrict the variability of the filters to \(a_i < A_j\) below a certain threshold \(c\) [cf. Assumption AS4]. This ensures that the amplifications of the spectral components below \(c\) won’t be too different, and the fact that they cannot be discriminated less problematic. Still, because \(a_i\) is nonzero, the transference of these spectral components will incur in an error—the third term of the transferability bound in Thms. 1 and 2. We refer to the energy of this error as the non-transferable energy, since it stems from spectral components which cannot be transferred from graph to graph.

Due to \(a_i\) being small, the non-transferable energy does not contribute much to the transferability error and is dominated by the part of the bound that depends on the transferability constant. However, this is largely dependent on the value of \(c\). Reducing the value of \(c\) reduces the contribution of the non-transferable energy to the transferability bound. But decreasing \(c\) also has the effect of increasing the transferability constant both through \(B_{c_{\text{max}}}^\text{max}\), because a lower value of \(c\) results in a larger number of eigenvalues in \([-1, -c)\cup[c, 1]\]; and through \(\delta_c\), because as \(c\) approaches zero so does the margin between consecutive eigenvalues, which is the limit of \(\delta_c\) as \(n \to \infty\).

The non-transferable energy is so called because, for fixed \(c\), it is a constant term in the transferability bound which does not decrease with the size of the graphs. To avoid non-transferable spectral components, the graph filter (or, respectively, the graph filters of the GNN) would need to have a constant frequency response for \(|\lambda| < c\) or, equivalently, \(a_i = 0\). However, such filters are undesirable in practice because since they are not analytic they cannot be written in convolutional form (1). Nonetheless, in the limit it is possible to show that the difference between the outputs of Lipschitz continuous graph convolutions supported on \(G_{n_1}\) and \(G_{n_2}\) converging to the same graphon (and therefore of GNNs constructed with such convolutions) vanishes as \(n_1, n_2 \to \infty\). This convergence result, which can be seen as an asymptotic transferability result, is proved in [20, Thm. 4].

**Transferability-discriminability tradeoff and the effect of nonlinearities.** In both the graph filter transferability result (Thm. 1) and its GNN counterpart (Thm. 2), the transferability constant (i.e., the first term of the bound) depends on the parameters \(B_{c_{\text{max}}}^\text{max}\) and \(\delta_c\) of the graph convolutional filters, which in turn depend on the value of \(c\). The parameter \(B_{c_{\text{max}}}^\text{max}\) is the maximum \(c\)-band cardinality of a graphon [cf. Definition 4], which counts the number of graphon eigenvalues larger than \(c\). The parameter \(\delta_c\) is the minimum \(c\)-eigenvalue margin between two graphons [cf. Definition 5], which measures the minimum distance between eigenvalues of these two graphons with consecutive indices where one is smaller and the other is larger than \(c\). Since the eigenvalues of the limit graphon accumulate near zero, if \(c\) is large (i.e., close to one), \(B_{c_{\text{max}}}^\text{max}\) is small—because there are less eigenvalues in the \([-1, -c)\cup[c, 1]\) interval—and \(\delta_c\) is
large—because the further away eigenvalues are from zero, the larger the distance between two consecutive eigenvalues. This leads to a smaller transferability constant, decreasing the transferability error bound. On the other hand, larger values of \( c \) also reduce the model’s discriminative power (or discriminability) because they decrease the length of the interval \([−1, −c] \cup [c, 1]\) where the filter has full variability (i.e., \( A_b \)). Hence, the bounds in Thms. 1 and 2 exhibit a tradeoff between transferability and discriminability for both graph filters and GNNs. The value of the Lipschitz constant \( A_b \) also plays a role in this trade-off, as higher values of \( A_b \) lead to more discriminability but increase the transferability bound. Note that while for finite graphs this tradeoff arises from an analysis of the error bounds in Thms. 1 and 2 (and not of the transferability error per se, whose specific form we do not know), this tradeoff is not an artifact of the specific form of the bound, as it also arises in the asymptotic convergence result from [20].

In the case of GNNs, we conjecture that the transferability-discriminability tradeoff is better than in graph filters because of the addition of nonlinearities. This is because nonlinearities have an effect akin to rectifiers, effectively scattering some of the spectral components associated with small eigenvalues to the middle range of the spectrum where they can then be discriminated by the filters in the next layer. Indeed, we will observe this empirically in the experiments of Section V: for the same level of discriminability (i.e., comparable performances on the graphs on which they are trained), GNNs exhibit a lower transferability error than graph filters. It is hypothesized that nonlinearities play a similar role in the stability of GNNs, but in that case, it is the scattering of the components associated with large eigenvalues that improves stability [6].

**Graph as a design parameter.** While to prove transferability we focus on the limit object interpretation of WNNs, their interpretation as generative models is also valuable because it allows looking at the graph as a tunable parameter of the GNN. I.e., instead of considering the graph to be a fixed hyperparameter, we could think of it as a learnable parameter of the architecture like the weights \( \mathcal{H} \). The interpretation of the graph as a design parameter motivates a number of interesting research directions, e.g., building more general GNN architectures (i.e., with larger degrees of freedom); designing adversarial graph perturbations for GNNs; and drawing deeper connections between GNNs and transformers [40], which can be thought of as GNN architectures where the graph is learned.

V. NUMERICAL EXPERIMENTS

We illustrate the transferability properties of graph filters and GNNs in two applications: movie recommendation on a movie similarity network, and flocking via decentralized robot control. All architectures are trained using ADAM with learning rate \( 5 \times 10^{-4} \) and forgetting factors 0.9 and 0.999.

**A. Movie Recommendation**

We consider the MovieLens-1M dataset [41], which consists of one million ratings given by 6000 users to 4000 movies. Each rating is a score between 1 and 5 and higher scores indicate higher preference. Specifically, we aim to predict the ratings given by different users to the movie “Star Wars: Episode IV - A New Hope”, which has a total of 2991 ratings. To do so, we define a movie similarity network, where each node is a movie and each edge is the similarity between two movies. We restrict attention to movies with at least 5 ratings, which brings the number of nodes of the full movie similarity network down to 3416. We compute this network by defining a training set with 90% of the users, and by defining the similarity between two movies as the correlation between the ratings given by users in the training set to these movies. Each user corresponds to a graph signal. At each node, the value of a signal is the rating given by the user to the corresponding movie, or zero if a rating is not available. To train graph filters and GNNs in a supervised manner, we constructed input-output pairs where, for each user, the output is the rating to “Star Wars” and the input is the user’s graph signal with the rating to this movie zeroed out.

The architectures we consider are a 1-layer graph filter and a 1-layer GNN. Both have \( F_0 = 1 \) input features, \( F_1 = 64 \) features in the first layer, and \( K = 5 \) filter taps. The nonlinearity of the GNN is the ReLU and both the graph filter and the GNN are followed by a readout layer which outputs 1 feature, the rating. These architectures are trained by minimizing the MSE at the node corresponding to the movie “Star Wars”. We set aside 10% of the training samples for validation and train over 30 epochs with batch size 16. Performance is measured by recording the root mean squared error (RMSE) achieved by each architecture on the test set. To analyze transferability, we train the graph filter and the GNN on subnetworks with \( n = 500, 750, 1000, 1250, 1500, 1750, 2000 \) movies, and plot the difference between the RMSE they achieve on the subnetwork of size \( n \) and on the full 3416-movie network, relative to the former, in Fig. 3(a).

In Fig. 3(a), we observe that the RMSE difference decreases with the size of the movie subnetwork for both the graph filter and the GNN as predicted by Thms. 1 and 2. Although the GNN (orange) achieves a lower transferability error than the graph filter (blue) for most values of \( n \), the difference is very small. We hypothesize that this is because the graph convolutions being learned in the GNN have high variability. To test this hypothesis, we retrain the GNN by minimizing the penalized MSE loss with penalization factor given by the maximum Lipschitz constant of the filters of the GNN scaled by a penalty multiplier. The transferability error of this GNN is shown in green. The Lipschitz GNN is more transferable than the graph filter and the GNN trained without regularization, corroborating that GNNs are more transferable than graph filters, and illustrating the transferability-discriminability tradeoff of Thm. 2.

**B. Flocking via Decentralized Robot Control**

Decentralized control problems consist of a team of \( n \) agents which must accomplish a shared goal. Each agent has access to local states \( x_i \) and generates local control actions \( u_i \). In order to learn which actions to perform, agents exchange information across pairwise communication links determined by their geographical proximity, which defines an agent proximity network \( G_n \). Because communication incurs in delays, if agents \( i \) and
Fig. 3. (a) Difference between the RMSEs achieved on the training network and on the full 3416-movie network for the graph filter, the GNN, and the GNN trained by penalizing convolutions with high Lipschitz constant. The errors achieved by training on the full movie network are $0.81 \pm 0.04$ for the graph filter and $0.81 \pm 0.04$ for the GNN. (b) Difference between the control cost achieved on the training network and on the full 100-agent network, relative to the cost on the training network, for a graph filter and a GNN. (c) Difference between the control cost achieved on the training network and on the full 100-agent network, relative to the cost on the original network, for a GNN and a GNN trained by penalizing convolutions with high Lipschitz constant. The costs achieved by training on the 100 agent network are $19.31 \pm 20.66$ for the graph filter and $1.46 \pm 0.01$ for the GNN. The error bars in (a)–(c) are scaled by 0.5.

$j$ are $k$ hops away from one another in $G_n$, at time $t_i$ only has access to the delayed state $x_j(t - k)$. We define the information history of agent $i$ as [42]

$$X_i(t) = \bigcup_{k=0}^{K-1} \{ x_j(t-k) : j \in N^k_i(t) \}.$$  

(29)

This emphasizes that at time $t$ agent $i$ only knows its current state $x_i(t)$ and the states of its $k$-hop neighbors $j \in N^k_i$ at time $t - k$.

The information history in (29) allows defining a decentralized control scheme in which the actions $u_i(t)$ can be calculated as functions of the history $X_i(t)$. We can then use graph filters and GNNs to parametrize these functions by incorporating the delayed information structure into (1). Explicitly, to account for communication delays we rewrite the terms $S^k x_i$ in (1) as $\prod_{k=1}^{K} S(t-k)x(t-k)$. The convolutional filters in the GNN (4)-(5) are modified in the same way.

In the flocking problem [43], the shared goal is for all the agents to move with the same velocity while avoiding collisions. The states $x_i(t) \in \mathbb{R}^6$ are given by [42]

$$x_i^T(t) = \left[ \sum_{j \in N_i} v_{ij}(t) ; \sum_{j \in N_i} \frac{r_{ij}(t)}{\| r_{ij}(t) \|^2} ; \sum_{j \in N_i} \frac{r_{ij}(t)}{\| r_{ij}(t) \|^2} \right]$$

(30)

where $v_{ij}(t)$ are the positions and $v_{ij}(t)$ the velocities of agent $j$ measured relative to the positions and velocities of agent $i$, respectively. The neighborhood $N_i$ consists of nodes $j$ such that $\| r_{ij} \| \leq R$, i.e., which are within a communication and sensing radius of length $R$ from $i$. We consider $R = 2$. At $t = 0$, the agents’ positions and velocities are initialized at random. The actions $u_i(t)$ to predict are the agents’ accelerations.

While a centralized solution to the flocking problem is straightforward—it suffices to direct all the agents to move in the same direction—the optimal decentralized controller is unknown. Hence, we train decentralized graph filters and GNNs following the information structure in (29) to “imitate” the centralized controller. We consider 2-layer graph filters and a GNNs with $F_0 = 6$ input features, $F_1 = 64$ features in the first layer, and $F_2 = 32$ features in the second layer. At all layers, the convolutional filters have $K = 3$ filter taps. The readout layer outputs 2 features per node corresponding to the agents’ accelerations in the $x$ and $y$ directions. The nonlinearity used in the GNN is the hyperbolic tangent.

Both architectures are trained by minimizing the mean squared error (MSE) over 400 training trajectories of duration equal to 100 steps. We train for 30 epochs using minibatches of size 20, and validate the model over 20 validation trajectories every 5 training steps. Performance is measured by recording the cost of the decentralized controller, given by the sum of the deviations of each agent’s velocity from the mean, relative to the cost of the centralized controller. The test set consists of 20 trajectories with the same duration.

In decentralized control problems, transferability is important because, since we are training to imitate a centralized controller, the learning architecture has to be trained offline. Hence, the networks observed during training are different than those observed during execution, and typically smaller, because the cost of training graph filters and GNNs can be prohibitive for large graphs. To assess whether the policies learned with graph filters and GNNs are transferable in the flocking problem, we perform the following experiment. We train the models on networks of size $n = 25, 37, 50, 75, 87$. Then, we test them on both the original network and on a network with $n = 100$ agents, and record the difference between the cost achieved on the original network and on the 100-agent network relative to the cost on the original network. The results of this experiment for 10 random realizations of the dataset are shown in Fig. 3(b).

In Fig. 3(b), we observe that the difference between the outputs of the graph filter and the GNN on these networks both decrease as the number of nodes increases. This is consistent with the asymptotic behavior of the transferability bounds in Thms. 1 and 2, which decrease with $n$. We also observe that, for fixed $n$, the relative cost difference of the GNN is smaller than the relative cost difference of the graph filter. This evidences...
that GNNs are more transferable than graph filters as discussed in Section IV.

We further validate our results by comparing the 2-layer GNN with a GNN with same architecture but trained by penalizing filters with large Lipschitz constant. In this case, we minimize the sum of the MSE and of a penalization term given by the largest Lipschitz constant across all filters multiplied by a penalization factor. The results are shown in Fig. 3(c). Note that the models were only trained in networks with up to $n=62$ agents due to memory constraints. Although as expected the Lipschitz GNN is more transferable than the GNN, the difference is very small. This can be interpreted to mean that, in flocking, the optimal GNN filters have low variability regardless, which makes sense considering that flocking is a type of consensus task.

C. Discussion

The framework developed in this article assumes the existence of an underlying random graph model in the form of a graphon. Two questions that arise are then whether the graphon is a suitable model in a given application and, if so, which graphon. In the movie recommendation problem, the movie similarity graph is clearly sampled from a graphon. We do not know the closed-form expression of this graphon, but we do know that it is the limit of a sequence of correlation graphs. In practice, however, we do not need to know the graphon. The training graphs generated by sampling nodes uniformly at random from the large target graph are valid samples from this graphon.

In the decentralized robot control problem, the communication graphs are clearly not sampled from a graphon, as their degree does not necessarily scale with the graph size. In this case, a better continuous model is perhaps the manifold, as we do know that it is the limit of a sequence of correlation graphs. In practice, however, we do not need to know the graphon. The training graphs generated by sampling nodes uniformly at random from the large target graph are valid samples from this graphon.

VI. CONCLUSIONS

In this article, we defined graphon convolutions and WNNs, and showed that graph filters and GNNs sampled from them on stochastic and weighted graphs can be used to approximate graphon convolutions and WNNs. Building upon this result, we then showed that graph filters and GNNs are transferable across weighted and stochastic graphs. The transferability error decreases with the size of the graphs, however, some spectral components are not transferable even when the graphs are large, which is a consequence of the fact that the graphon eigenvalues accumulate near zero. Furthermore, our transferability results reveal a tradeoff between the transferability and the spectral discriminability of graph convolutions, which is inherited by GNNs. In practice, however, in GNNs this tradeoff is alleviated by the pointwise nonlinearities. These findings were corroborated empirically in the problems of movie recommendation and decentralized robot control.

A limitation of our analysis is that graphons are only good models for limits of dense graphs. Therefore, a future research direction is to extend the transferability analysis of this article to sparser graph limits, such as manifolds [44] and $L^p$ graphs [45]. Another limitation of our work is that it does not address scenarios where the graph on which the GNN should be trained to meet a specific error allowance is still too large. To address this problem, we will leverage the transferability results in this article to design training algorithms yielding more transferable GNNs [46]. A third relevant research direction is the empirical evaluation of the tightness of the graph filter and GNN transferability error bounds. This is important for practical purposes, to inform the choice of the training graph size, as well as for theoretical purposes, to motivate attempts at tighter results.

APPENDIX A

PROOF OF THM. 1

To prove Thm. 1, we need the following propositions.

Proposition 7: Let $W : [0, 1]^2 \rightarrow [0, 1]$ and $W' : [0, 1]^2 \rightarrow [0, 1]$ be two graphons with eigenvalues given by $\{\lambda_i(TW)\}_{i \in \mathbb{Z} \setminus \{0\}}$ and $\{\lambda_i(TW')\}_{i \in \mathbb{Z} \setminus \{0\}}$, ordered according to their sign and in decreasing order of absolute value. Then, for all $i \in \mathbb{Z} \setminus \{0\}$, the following inequalities hold

$$|\lambda_i(TW') - \lambda_i(TW)| \leq \|TW' - W\| \leq \|W' - W\|.$$  

Proof: See [19, Prop. 4].

Proposition 8: Let $T$ and $T'$ be two self-adjoint operators on a separable Hilbert space $\mathcal{H}$ whose spectra are partitioned as $\gamma \cup \Gamma$ and $\omega \cup \Omega$ respectively, with $\gamma \cap \Gamma = \emptyset$ and $\omega \cap \Omega = \emptyset$. If there exists $d > 0$ such that $\min_{x \in \gamma, y \in \Omega} |x - y| \geq 2d$, then the spectral projections $E_T(\gamma)$ and $E_{T'}(\omega)$ satisfy

$$\|E_T(\gamma) - E_{T'}(\omega)\| \leq \frac{\pi}{2} \frac{\|T - T'\|}{d}.$$  

Proof: See [47].

To prove Thm. 1, we fix a constant $0 < c < 1$ and decompose the filter $h$ into filters $h^{\geq c}$ and $h^{< c}$ defined as

\[
\begin{align*}
    h^{\geq c}(\lambda) & = \begin{cases} 
    0 & \text{if } |\lambda| < c \\
    h(\lambda) - h(c) & \text{if } \lambda \geq c \\
    h(\lambda) - h(-c) & \text{if } \lambda \leq -c
    \end{cases} \\
    h^{< c}(\lambda) & = \begin{cases} 
    h(c) & \text{if } \lambda \geq c \\
    h(-c) & \text{if } \lambda \leq -c
    \end{cases}
\end{align*}
\]  

so that $h = h^{\geq c} + h^{< c}$. Note that both functions have the same Lipschitz constants as $h$. We start by analyzing the transferability of $h^{\geq c}$.

Let $T^{\geq c}$ and $T^{< c}$ be graphon filters with filter function $h^{\geq c}$ on the graphons $W$ and $W_\lambda$, respectively. Using the triangle
inequality, we can write the norm difference \(|T_H^{c,c}X - T_{H_n}^{c,c}X_n|\) as
\[
\left| T_H^{c,c}X - T_{H_n}^{c,c}X_n \right| = \left| T_H^{c,c}X + T_{H_n}^{c,c}X - T_{H_n}^{c,c}X - T_{H_n}^{c,c}X_n \right|
\leq \left| T_H^{c,c}X - T_{H_n}^{c,c}X \right| + \left| T_{H_n}^{c,c}(X - X_n) \right|
\]
where the LHS is split between terms (1) and (2).

Writing the inner products \(\int_0^1 X(u)\varphi_i(u)du\) and \(\int_0^1 X(u)\varphi_i^n(u)du\) as \(\hat{X}(\lambda_i)\) and \(\hat{X}(\lambda_i^n)\) for simplicity, we can then express (1) as
\[
\left| T_H^{c,c}X - T_{H_n}^{c,c}X \right| = \left| \sum_i h^{\geq c}(\lambda_i)\hat{X}(\lambda_i)\varphi_i - \sum_i h^{\geq c}(\lambda_i^n)\hat{X}(\lambda_i^n)\varphi_i^n \right|
\leq \left| \sum_i \left( h^{\geq c}(\lambda_i) - h^{\geq c}(\lambda_i^n) \right) \hat{X}(\lambda_i)\varphi_i \right| + \left| \sum_i h^{\geq c}(\lambda_i^n) \left( \hat{X}(\lambda_i)\varphi_i - \hat{X}(\lambda_i^n)\varphi_i^n \right) \right|.
\]

Using the triangle inequality, this becomes
\[
\left| T_H^{c,c}X - T_{H_n}^{c,c}X \right| \leq \sum_i \left( h^{\geq c}(\lambda_i) - h^{\geq c}(\lambda_i^n) \right) \parallel \hat{X}(\lambda_i)\parallel \parallel \varphi_i \parallel + \sum_i h^{\geq c}(\lambda_i^n) \parallel \hat{X}(\lambda_i)\parallel \parallel \varphi_i - \varphi_i^n \parallel \]
where we have now split (1) between (1.1) and (1.2).

Focusing on (1.1), note that the filter’s Lipschitz property allows writing \(|h(\lambda_i) - h^{\geq c}(\lambda_i^n)| \leq A_h |\lambda_i - \lambda_i^n|\). Hence, using Prop. 7 together with the Cauchy-Schwarz inequality, we get
\[
\left| \sum_i \left( h^{\geq c}(\lambda_i) - h^{\geq c}(\lambda_i^n) \right) \hat{X}(\lambda_i)\varphi_i \right| \leq A_h \parallel W - W_n \parallel \sum_i \parallel \hat{X}(\lambda_i)\parallel \parallel \varphi_i \parallel = A_h \parallel W - W_n \parallel \parallel X \parallel.
\]

For (1.2), we use the triangle and Cauchy-Schwarz inequalities to write
\[
\left| \sum_i h^{\geq c}(\lambda_i^n) \left( \hat{X}(\lambda_i)\varphi_i - \hat{X}(\lambda_i^n)\varphi_i^n \right) \right| = \left| \sum_i h^{\geq c}(\lambda_i^n) \left( \hat{X}(\lambda_i)\varphi_i + \hat{X}(\lambda_i)\varphi_i^n \right. \right.
\]
\[
\left. \left. - \hat{X}(\lambda_i^n)\varphi_i^n - \hat{X}(\lambda_i^n)\varphi_i^n \right) \right| \leq \sum_i h^{\geq c}(\lambda_i^n) \hat{X}(\lambda_i)\varphi_i^n
\]
\[
+ \sum_i h^{\geq c}(\lambda_i^n) \varphi_i^n \parallel \hat{X}(\lambda_i^n)\varphi_i^n \parallel \leq 2 \sum_i \parallel h^{\geq c}(\lambda_i^n)\parallel \parallel X \parallel \parallel \varphi_i - \varphi_i^n \parallel.
\]

Using Prop. 8 with \(\gamma = \lambda_i\) and \(\omega = \lambda_i^n\), we then get
\[
\left| \sum_i h^{\geq c}(\lambda_i^n) \left( \hat{X}(\lambda_i)\varphi_i - \hat{X}(\lambda_i^n)\varphi_i^n \right) \right| \leq \parallel X \parallel \sum_i \parallel h^{\geq c}(\lambda_i^n)\parallel \frac{\pi ||W - W_n||}{d_i} \]
where \(d_i \) is the minimum between \(\min(|\lambda_i - \lambda_i^{n+1}|, |\lambda_i - \lambda_i^{n-1}|)\) and \(\min(|\lambda_i^n - \lambda_i^{n+1}|, |\lambda_i^n - \lambda_i^{n-1}|)\) for each \(i\). Since \(\delta_{W,W_n} \leq d_i\) for all \(i\) and \(\parallel W - W_n \parallel \leq ||W - W_n||\) (i.e., the Hilbert-Schmidt norm dominates the operator norm), this becomes
\[
\left| \sum_i h^{\geq c}(\lambda_i^n) \left( \hat{X}(\lambda_i)\varphi_i - \hat{X}(\lambda_i^n)\varphi_i^n \right) \right| \leq \pi \frac{||W - W_n||}{\delta_{W,W_n}} \parallel X \parallel \parallel B_{W_n} \parallel.
\]

The final bound for (1.2) is obtained by noting that \(|h^{\geq c}(\lambda)| < 1\) and \(h^{\geq c}(\lambda) = 0\) for \(|\lambda| < c\). Since there are a total of \(B_{W_n}\) eigenvalues \(\lambda_i^n\) for which \(|\lambda_i^n| \geq c\), we get
\[
\left| \sum_i h^{\geq c}(\lambda_i^n) \left( \hat{X}(\lambda_i)\varphi_i - \hat{X}(\lambda_i^n)\varphi_i^n \right) \right| \leq \pi \frac{||W - W_n||}{\delta_{W,W_n}} \parallel X \parallel \parallel B_{W_n} \parallel.
\]

The bound for (2) follows immediately from the Cauchy-Schwarz inequality. Since \(|h(\lambda)| < 1\), the norm of the operator \(T_{H_n}\) is bounded by 1. Hence,
\[
\parallel T_{H_n}^{c,c}(X - X_n) \parallel \leq \parallel X - X_n \parallel
\]
which completes the bound on \(\parallel T_H^{c,c}X - T_{H_n}^{c,c}X_n \parallel\).

Next we analyze the transferability of the filter \(h^{<c}\). Note that, because this filter only varies in the interval \((-c,c)\) and has Lipschitz constant \(a_h\), we can bound \(\parallel T_H^{c,c}X - T_{H_n}^{c,c}X_n \parallel\) as
\[
\parallel T_H^{c,c}X - T_{H_n}^{c,c}X_n \parallel \leq \parallel (h(0) + a_h e)X - (h(0) - a_h e)X_n \parallel \leq |h(0)| \parallel X - X_n \parallel + a_h e \parallel X + X_n \parallel
\]
where the second inequality follows from the triangle and Cauchy-Schwarz inequalities. Using the fact that \(|h(\lambda)| < 1\), and adding and subtracting \(X\) to \(\parallel X + X_n \parallel\) to apply the triangle inequality once again, his bound becomes
\[
\parallel T_H^{c,c}X - T_{H_n}^{c,c}X_n \parallel \leq \parallel X - X_n \parallel + a_h e \parallel 2X + X_n - X \parallel \leq (1 + a_h e) \parallel X - X_n \parallel + 2a_h e \parallel X \parallel.
\]
Finally, to obtain the transferability bound for $h$ observe that,
since $h = h^{<c} + h^{<c}$, we can write
\[ ||T_H X - T_{H_n} X_n|| = \left|\left| T_H^{<c} X + T_{H^{<c}} X - T_{H^{<c}} X_n - T_{H^{<c}} X_n \right|\right| \]
\[ \leq ||T_H^{<c} X - T_{H^{<c}} X_n|| + ||T_{H^{<c}} X - T_{H^{<c}} X_n|| \]
i.e., the error made when transferring $h$ can be bounded by the sum of the transferability bounds of $h^{<c}$ and $h^{<c}$. Putting together equations (33), (34), (35) and (36) thus concludes the proof.

REFERENCES

[1] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in Proc. 5th Int. Conf. Learn. Representations. Toulon, France: ACM, Apr. 24–26, 2017, pp. 1–14.

[2] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in Proc. Neural Inf. Process. Syst. Barcelona, Spain: NIPS Foundation, Dec. 5–10, 2016, vol. 4, pp. 1–9.

[3] F. Gama, A. G. Marques, G. Leus, and A. Ribeiro, “Convolutional neural network architectures for signals supported on graphs,” IEEE Trans. Signal Process., vol. 67, no. 4, pp. 1034–1049, Feb. 2019.

[4] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, “Geometric deep learning: Going beyond Euclidean data,” 2017. [Online]. Available: http://arxiv.org/abs/1702.03281.

[5] L. Ruiz, F. Gama, G. Marques, and A. Ribeiro, “Invariant-preserving localized activation functions for graph neural networks,” 2019. [Online]. Available: https://arxiv.org/abs/1903.12575

[6] F. Gama, J. Bruna, and A. Ribeiro, “Stability properties of graph neural networks,” 2019. arXiv:1905.04497.

[7] F. Gama, E. Isufi, G. Leus, and A. Ribeiro, “Graphs, convolutions, and neural networks: From graph filters to graph neural networks,” IEEE Signal Process. Mag., vol. 37, no. 6, pp. 128–138, Nov. 2020.

[8] L. Lovász, Large Networks and Graph Limits, vol. 60. Providence, RI, USA: AMS, 2012.

[9] L. Lovász and B. Szegedy, “Limits of dense graph sequences,” J. Combinatorial Theory, Ser. B, vol. 96, no. 6, pp. 933–957, 2006.

[10] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi, “Convergent sequences of dense graphs. I: Subgraph frequencies, metric properties and testing,” Adv. Math., vol. 219, no. 6, pp. 1801–1851, 2008.

[11] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi, “Convergent sequences of dense graphs. II. Multiway cuts and statistical physics,” Ann. of Math., vol. 176, no. 1, pp. 151–219, Dec. 2012.

[12] P. J. Wolfe and S. C. Olshe, “Nonparametric graphon estimation,” 2013. arXiv:1309.5936.

[13] J. Xu, “Rates of convergence of spectral methods for graphon estimation,” 2017. arXiv:1709.03183.

[14] C. Gao, Y. Lu, and H. H. Zhou, “Rate-optimal graphon estimation,” Ann. Statist., vol. 43, no. 6, pp. 2624–2652, 2015.

[15] F. Parise and A. Ozdaglar, “Graphon games,” in Proc. Int. Conf. Mach. Learn. Cambridge, MA, USA: PMLR, 2021, pp. 5388–5397.

[16] H. Kenlay, D. Thanou, and X. Dong, “Interpretable stability bounds for spectral graph filters,” in Proc. Int. Conf. Mach. Learn. Cambridge, USA: PMLR, 2021, pp. 5388–5397.

[17] H. Kenlay, D. Thanou, and X. Dong, “On the stability of graph convolutional neural networks under edge rewiring,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. Piscataway, NJ, USA: IEEE, 2021, pp. 8513–8517.

[18] F. Gama, A. Ribeiro, and J. Bruna, “Graphon filters: Signal processing in very large graphs,” in Proc. 28th Eur. Signal Process. Conf. (EUSIPCO). Piscataway, NJ, USA: IEEE, 2021, pp. 1050–1054.

[19] M. W. Morency and G. Leus, “Graphon filters: Graph signal processing in the limit,” IEEE Trans. Signal Process., vol. 69, pp. 1740–1754, 2021.

[20] D. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains,” IEEE Signal Process. Mag., vol. 30, no. 3, pp. 83–98, May 2013.

[21] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, “Graph signal processing: Overview, challenges, and applications,” Proc. IEEE, vol. 106, no. 5, pp. 808–828, May 2018.

[22] A. Sandryhaila and J. M. F. Moura, “Discrete signal processing on graphs,” IEEE Trans. Signal Process., vol. 61, no. 7, pp. 1644–1656, Apr. 2013.

[23] R. Boukrab et al., “Random-walk Laplacian for frequency analysis in periodic graphs,” Sensors, vol. 21, no. 4, 2021, Art. no. 1275.

[24] J. Du, J. Shi, S. Kar, and J. M. F. Moura, “On graph convolution for graph CNNs,” in Proc. IEEE Data Sci. Workshop. Lausanne, Switzerland: IEEE, Jun. 4–6, 2018, pp. 239–243.

[25] S. Segarra, A. G. Marques, and A. Ribeiro, “Optimal graph-filter design and applications to distributed linear network operators,” IEEE Trans. Signal Process., vol. 65, no. 15, pp. 4117–4131, Aug. 2017.

[26] A. Sandryhaila and J. M. F. Moura, “Discrete signal processing on graphs: Frequency analysis,” IEEE Trans. Signal Process., vol. 62, no. 12, pp. 3042–3054, Jun. 2014.

[27] L. Ruiz, L. F. O. Chamon, and A. Ribeiro, “The graphon Fourier transform,” 2019. [Online]. Available: http://arxiv.org/abs/1910.10195.

[28] P. D. Lax, Functional Analysis. Hoboken, NJ, USA: Wiley, 2002.

[29] A. Venkata et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2021, pp. 5998–6008.

[30] F. M. Harper and J. A. Konstan, “The MovieLens datasets: History and context,” ACM Trans. Interact. Intell. Syst. (TiiS), vol. 5, pp. 1–19, Jan. 2016.

[31] F. Gama, E. Tolstaya, and A. Ribeiro, “Graph neural networks for decentralized controllers,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP). Piscataway, NJ, USA: IEEE, 2021, pp. 5260–5264.

[32] E. Tolstaya, F. Gama, J. Paulos, G. Pappas, V. Kurnar, and A. Ribeiro, “Learning decentralized controllers for robot swarms with graph neural networks,” in Proc. Conf. Robot Learn., 2020, pp. 671–682.

[33] Z. Wang, L. Ruiz, and A. Ribeiro, “Stability of manifold neural networks to perturbations,” 2021. [Online]. Available: http://arxiv.org/abs/2106.10372.

[34] C. Borgs, J. Chayes, H. Cohn, and Y. Zhao, “Anlp theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions,” Trans. Amer. Math. Soc., vol. 372, no. 5, pp. 3019–3062, 2019.

[35] J. Cervino, L. Ruiz, and A. Ribeiro, “Increase and conquer: Training graph neural networks on growing graphs.” 2021. [Online]. Available: http://arxiv.org/abs/2106.03693.

[36] A. Seelmann, “Notes on the sin 2 Θ theorem,” Integral Equ. Operator Theory, vol. 79, no. 4, pp. 579–597, 2014.

[37] F. Chung and M. Radrifche, “On the spectra of general random graphs,” Electron. J. Combinatorics, vol. 18, no. 1, pp. P215–P215, 2011.