The Role of Dimension in Graph Convolutional Networks

Abram Magner AMAGNER@ALBANY.EDU
Department of Computer Science
University at Albany, SUNY
Albany, NY, USA

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

Graph convolutional networks are a popular representation learning method for graphs, wherein an input graph is mapped to a d-dimensional *embedding vector*, yielding a latent representation. We continue the project of theoretically elucidating the roles of various aspects of GCN architectures by studying the power and limitations of GCNs in distinguishing random graph models based on embedding vectors of sample graphs. In the present work, we show how the embedding dimension affects the set of pairs of models that can be distinguished from one another. We additionally extend the theory to the setting of graphs with vertex colors that are potentially locally correlated to graph structure. We also consider the application of GCNs to multi-hypothesis testing and use channel capacity results to show a lower bound on how the embedding dimension must scale with respect to the number of hypotheses and the signal-to-noise ratio in order to guarantee a probability of error tending to 0.

1. Introduction

Many modern machine learning tasks involve supervised and semi-supervised learning on graphs. In order to leverage deep learning approaches, wherein inputs take the form of vectors in a d-dimensional Euclidean space, for these problems, representation learning methods for graphs have become important. A menagerie of methods (see Hamilton et al. (2017) for a survey) have been proposed and explored empirically. Theoretical examination, which we discuss in detail in Section 1.1, has been from a variety of different angles, from generalization bounds to stability results to characterizations in terms of isomorphism tests. However, important questions about the design of architectures remain. To study these questions, in this work, we take the perspective of studying the learned representations in terms of the information-theoretic limits of their use in downstream tasks, specifically classification. This provides a unifying mathematical objective by which different representation learning methods can be compared.

One popular representation learning architecture is the *graph convolutional network* (GCN) (Defferrard et al., 2016; Kipf and Welling, 2016), the main subject of the present work. At a high level, the graph convolutional network assigns an initial embedding vector with dimension d (a hyperparameter, the *embedding dimension* of the GCN) to each node of an input graph G. These initial embedding vectors are then passed through a sequence of layers, where each layer consists of a neighborhood averaging phase, followed by multiplication by a weight matrix (which is typically learned from a training set), and, finally, followed by component-wise application of a nonlinear activation function. The resulting matrix is then converted to a graph embedding vector via, e.g., averaging of its rows.

Given the wide application of GCNs to various application domains, it is desirable to have theoretical insight into the roles played by its hyperparameters in its performance limits on a task of interest.

Here we attempt to provide insight into the role played by the embedding dimension in the ability of a GCN to distinguish between different random graph models from their samples. The problem of distinguishing random graph models is simply a restatement of the problem of *graph classification*, wherein there are two (or more) different class labels, each corresponding to a different conditional probability distribution on graphs. Thus, it is a natural downstream task on which we can study the performance and fundamental/information-theoretic limits of representation learning methods. Since our focus is on information-theoretic and convergence results without the involvement of training samples, our results are about the representation capabilities of GCNs and are thus complementary to results on their generalization ability.

It is desirable to have a sufficiently rich class of graph models with a convenient parameterization. For this reason, we focus our attention on those models that are parameterized by *graphons* (Lovász and Szegedy, 2006). A graphon is a symmetric, Lebesgue-measurable function $W : [0,1]^2 \to [0,1]$ and can be interpreted from a variety of perspectives, as described in our discussion of prior work.

1.1. Prior work

This work builds on Magner et al. (2020); Magner et al. (2020), which provided the theoretical framework within which we study the performance limitations of GCNs, in particular posing the task of distinguishing graphons on the basis of GCN embedding outputs with input graphs sampled from the graphons. That work proved several results concerning GCNs in a setting where the embedding dimension is fixed to be n, the size of the sample graph. This leaves numerous open questions regarding the behavior of GCNs. We specifically go further by (i) studying the effect of varying the embedding dimension and (ii) extending the theory to graphs with vertex colors.

We now review some relevant prior work in the theory and application of GCNs. Several recent works focus on the extent to which GCNs, as functions from graphs to vectors, exhibit injectivity (Xu et al., 2018; Chen et al., 2019; Morris et al., 2019). They show that, in the sense of injectivity, GCNs can be made to distinguish between the same pairs of graphs as the classical Weisfeiler-Lehman invariant for graph isomorphism (Weisfeiler and Lehman, 1968). In particular, distinguishing between two graphs G_1, G_2 here means that the GCN maps G_1 and G_2 to distinct embedding vectors. There are a few primary limitations of such an approach for studying the capabilities of a representation learning method: (i) it ignores the fact that there is a *metric structure* on the input space (so they make no statements about the distance between the embedding vectors of graphs that are very dissimilar in some appropriate metric on graphs), and (ii) it does not directly imply anything about the capabilities of GCNs as representation learning methods employed for downstream supervised learning tasks. The framework introduced in Magner et al. (2020); Magner et al. (2020) addresses both of these limitations. This is the reason that we extend it.

Other works bound the generalization error of GCNs and consider questions of stability (Garg et al., 2020; Verma and Zhang, 2019). Our work is complementary to these sorts of results, as the generalization error bounds do not fully explain some of the phenomena observed in practice, such as the curious dependence of the performance of GCNs on depth or on embedding dimension.

Closer in spirit, in some respects, to our work is Keriven et al. (2020), which studies convergence and stability of graph convolutional networks on sparse random graphs. In particular, they define a notion of a continuous GCN, and they show convergence of discrete GCNs to continuous limits as the size of the input graphs tends to infinity. This allows them to study stability of GCNs to random parameter perturbations.

On the applied/empirical side, several recent works have studied the properties of GCN architectures and have provided evidence of potentially surprising phenomena. For example, in Wu et al. (2019), it is argued through empirical data that GCNs with linear activation functions perform as well as nonlinear ones on certain classification problems. Additionally, some work has observed that **untrained** GCNs have nontrivial performance (Kipf and Welling, 2016; Kawamoto et al., 2018). This highlights the need for more theory to elucidate the roles of the various design choices in the performance of GCNs. The author of the present work is not aware of any empirical papers hinting at general principles governing the choice of the embedding dimension of GCNs.

Graphons We next review relevant work on graphons. Extensive references to classical literature on graphons are available in Magner et al. (2020); Magner et al. (2020). For the purposes of the present work, the relevant notions are as follows: by the Aldous-Hoover theorem, the space of graphons parameterizes the space of infinite vertex-exchangeable random graph models – that is, probability distributions on graphs that are invariant under permutations of the vertices. In particular, to a given graphon W, we associate a probability distribution on n-vertex graphs as follows: we sample n *latent positions* $x_1, ..., x_n$ i.i.d. from the uniform distribution on $[0, 1]$. For each pair $i, j \in [n]$, we then put an edge between the vertices i and j with probability $W(x_i, x_j)$, independently of anything else.

Thus, graphons provide a convenient geometrization of a broad class of models, and other models of interest, such as Erdős-Rényi and stochastic block models (in the limit of infinitely many vertices), are contained within this class. There is a natural metric associated with graphons: the *cut distance* $d_{cut}(\cdot, \cdot)$. The details are spelled out in Magner et al. (2020); Magner et al. (2020). For our purposes, the important point is that our results concern classes of pairs of graphons that are well-separated in cut distance but are mapped to very proximal embedding vectors by (essentially) arbitrary GCNs. In what follows, we will not explicitly refer to the cut distance.

Markov chains and mixing times We refer to Levin et al. (2006) for the necessary background on Markov chains and mixing times.

1.2. Main problem: Distinguishing graphons from GCN embeddings of samples

We next introduce the main hypothesis testing problem for which we prove our results. Two graphons W_0, W_1 are fixed. A coin $B \sim \text{Bernoulli}(1/2)$ is flipped, and a sample graph $G_n \sim W_B$ on n vertices is generated. A GCN with fixed parameters and K layers embeds G_n to a vector $\hat{H}^{(B,K)}$. The hypothesis testing problem is to recover the value of B from $\hat{H}^{(B,K)}$. We will also consider a variant in which $\hat{H}^{(B,K)}$ is perturbed by independent noise in each coordinate, yielding $H^{(B,K)}$.

1.3. Summary of contributions

The main contributions of this work are as follows: we explicitly introduce the notion of a universal exceptional set of pairs of graphons, which was implicit in Magner et al. (2020); Magner et al. (2020). This is the set of all pairs of graphons that cannot be distinguished by any (simple, norm-constrained) GCN of moderate depth. We first give an operational characterization of the universal exceptional set for arbitrary embedding dimension $d \leq n$. We then study its size as a function of embedding dimension, and we find that the set remains fixed, regardless of the value of d. However, we show a probability of error lower bound in terms of the embedding dimension and number of hypothesis graphons in the multiple hypothesis testing setting.

We then consider the exceptional set of graphon pairs for an arbitrary fixed GCN with a given value of d. This is the set of graphon pairs that the particular GCN under consideration cannot distinguish. We define a natural notion of dimension associated with this set and lower bound it in terms of the embedding dimension.

Finally, we extend the theory to graphs with vertex colors in the setting where colors are potentially correlated with local graph structure. We give dimension results analogous to those described above in the colorless case. In particular, our bounds now involve the total variation distance between the two color distributions of the models to be distinguished.

In general, we find that while there is a dependence on embedding dimension, it is surprisingly weak in the two-hypothesis case.

2. Main results

2.1. Notation and preliminaries

We start with definitions relevant to graph convolutional networks (GCNs). A K-layer GCN with embedding dimension $d \in \mathbb{N}$ is a function mapping graphs to vectors over $\mathbb{R}$. It is parameterized by a sequence of K *weight matrices* $W^{(j)} \in \mathbb{R}^{d\times d}$, $j \in \{0, ..., K-1\}$. Each of these corresponds to a layer. From an input graph G with adjacency matrix A and random walk matrix $\hat{A}$ (i.e., $\hat{A}$ is A with every row normalized by the sum of its entries), and starting with an initial embedding matrix $\hat{M}^{(0)}$ that may be a function of the input graph and node/edge features, the ℓth embedding matrix is defined as follows:

$$\hat{M}^{(\ell)} = \sigma(\hat{A} \cdot \hat{M}^{(\ell-1)} \cdot W^{(\ell-1)}), \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a fixed nonlinear *activation function* and is applied element-wise to an input matrix. An *embedding vector* $\hat{H}^{(\ell)} \in \mathbb{R}^{1\times d}$ is then produced by averaging the rows of $\hat{M}^{(\ell)}$:

$$\hat{H}^{(\ell)} = \frac{1}{n} \cdot \mathbf{1}^T \hat{M}^{(\ell)}. \tag{2}$$

We note that many alternative ways of generating the final embedding vector from the final embedding matrix have been considered; e.g., it could be the output of a feedforward neural network. However, since we are interested specifically in the capabilities of GCNs (i.e., the idea of iterated application of a graph-derived convolution matrix, followed by a linear transformation of features, followed by a nonlinearity), it is natural to choose a more constrained output.

Typical examples of activation functions in neural network and GCN contexts include the ReLU, sigmoid, and hyperbolic tangent functions. However, some empirical work indicates that the performance of GCNs does not suffer from the use of linear activations (Wu et al., 2019).

We note that $\hat{A}$ in the architecture just described plays the role of a *convolution* matrix. It is frequently replaced by some related matrix, such as the normalized adjacency matrix or the graph Laplacian. We consider $\hat{A}$ for simplicity and comparability with the previous work of Magner et al. (2020); Magner et al. (2020).

Throughout the paper and proofs, we will refer to a few different norms. For a matrix or vector M, $\|M\|_\infty$ denotes the L_∞ norm (i.e., the maximum absolute value of any component of M). For a matrix M, we denote by $\|M\|_{op,\infty}$ the L_∞ operator norm of M. For two probability distributions μ_0, μ_1 on a discrete set $\mathcal{X}$, we denote by $d_{TV}(\mu_0, \mu_1)$ the total variation distance between μ_0, μ_1: $d_{TV}(\mu_0, \mu_1) = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu_0(x) - \mu_1(x)|$.We will frequently use alternative characterizations of the total variation distance between two random variables X and Y, for example, as the minimum possible probability that $X \neq Y$ under any coupling of the two.

The limiting behavior of GCNs for graphs G with size tending to infinity is related to the stationary distribution of the random walk on G. We denote this distribution by π_G and recall that the stationary probability of each vertex is its degree divided by the sum of the degrees of all vertices.

Classes of pairs of graphons

Definition 1 (ℓ-lower bounded graphon) *We say that a graphon W is ℓ-lower bounded, for an $\ell \in \mathbb{R}$ if for all $x, y \in [0,1]$, $W(x,y) \geq \ell$.*

Definition 2 *For a graphon W, we define the degree function $d_W : [0,1] \to \mathbb{R}$ to be $d_W(x) = \int_0^1 W(x,y)\,dy$.We define the total degree function to be $D(W) = \int_0^1\int_0^1 W(x,y)\,dx\,dy$.*

For a pair of graphons W_0, W_1, we define the function

$$d_{deg}(W_0, W_1) = \inf_{\phi} \int_0^1 \left| \frac{d_{W_0}(\phi(x))}{D(W_0)} - \frac{d_{W_1}(\phi(x))}{D(W_1)} \right| dx, \tag{3}$$

where the infimum ranges over all measure-preserving bijections ϕ.

We next introduce a few sets whose characterization is fundamental to our results.

Definition 3 (Operational universal δ-exceptional set) *The operational universal δ-exceptional set is the set of all pairs of graphons W_0, W_1 satisfying the following properties:*

- *Both W_0, W_1 are ℓ-lower bounded for some fixed $\ell > 0$.*
- *$d_{deg}(W_0, W_1) \leq \delta$.*

Definition 4 (Universal exceptional set) *The universal exceptional set (with respect to a class $\mathcal{C}$ of GCNs) of pairs of graphons is the set of pairs W_0, W_1 satisfying the following properties:*

- *Both W_0, W_1 are ℓ-lower bounded for some fixed $\ell > 0$.*
- *Consider a sequence $G_n^{(b)} \sim W_b$, for each $b \in \{0,1\}$, with an arbitrary coupling between the two sequences. For any graph convolutional network in $\mathcal{C}$ with output graph embedding function $\tilde{F}_n(\cdot)$, we have that with high probability as $n \to \infty$,*

$$\|\tilde{F}_n(G_n^{(0)}) - \tilde{F}_n(G_n^{(1)})\|_\infty = o(1/n). \tag{4}$$

The motivation for this definition is as follows: typically, the L_∞ norms of the embedding vectors are $\Theta(1/n)$. Thus, the condition (4) implies that the two embedding vectors are asymptotically equivalent as the size of the input graph tends to infinity.

We can also define an exceptional set for a **particular** GCN, which is simply the universal exceptional set with respect to a class $\mathcal{C}$ containing only that GCN.

Definitions relating to GCN architectures We next present definitions that specify the set of GCNs under consideration. These definitions are the same as in Magner et al. (2020); Magner et al. (2020).

Definition 5 (Nice activation functions) *We define $\mathcal{A}$ to be the class of activation functions $\sigma : \mathbb{R} \to \mathbb{R}$ satisfying the following conditions:*

- *$\sigma \in C^2$ (i.e., σ has a continuous second derivative at every point).*
- *$\sigma(0) = 0, \sigma'(0) = 1$, and $\sigma'(x) \leq 1$ for all x.*

We restrict our theorems and proofs to these activation functions; however, as stated in Magner et al. (2020); Magner et al. (2020), certain of the conditions can be relaxed in various ways in order to include the hyperbolic tangent, swish, and SeLU functions, among others.

Throughout, we will insist on some norm constraints on the GCN weight matrices. We consider that these are justified by the fact that GCNs are not very interesting if their distinguishing capabilities come primarily from having high-norm weight matrices and initial embeddings, and not from some more intricate structure that must be designed.

Definition 6 (Norm constraints on GCNs) *Fix two positive constants C and E. We say that a GCN is* norm-constrained *if the initial embedding matrix $\hat{M}^{(0)}$ and the weight matrices $\{W^{(j)}\}_{j=0}^{K}$ satisfy $\|\hat{M}^{(0)T}\|_{op,\infty} \cdot \prod_{j=0}^{K} \|W^{(j)T}\|_{op,\infty} \leq C$,and $\sum_{j=0}^{K} \|W^{(j)T}\|_{op,\infty} \leq E$.*

Finally, we will say that a GCN is of *moderate depth* (with respect to a collection $\mathcal{W}$ of graphons and a parameter $\epsilon > 0$, which will be clear from context) if it has K layers with K larger than the supremum of the ϵ-total variation mixing times of all graphons in W. In our setting, this will mean that $K \geq D \log n$, for some large enough D with respect to $\mathcal{W}$. Alternatively, it suffices to take $L \geq \omega(1) \log n$, where $\omega(1)$ is a function of n that grows arbitrarily slowly to infinity as $n \to \infty$.

In what follows, we will restrict our attention, not necessarily with explicit statement, to norm-constrained GCNs of moderate depth with activation functions coming from $\mathcal{A}$. We denote this class of GCNs by $\mathcal{C}$, and the subset of GCNs with embedding dimension d we denote by $\mathcal{C}_d$.

Noise model For reasons of efficiency, neural network architectures frequently operate using limited precision (Sakr et al., 2017; Gupta et al., 2015). In order to study the distinguishing power of GCNs in this setting, we adopt the following noise model, which was justified and used in Magner et al. (2020); Magner et al. (2020): we fix some $\epsilon_{res}(n) = \epsilon_{res} > 0$. Then the output embedding vector of the GCN is perturbed in every coordinate by an independent uniform random number taken from the interval $[-\epsilon_{res}, \epsilon_{res}]$. The *unperturbed* graph embedding vector is denoted by $\hat{H}^{(\ell)}$, as above, while the perturbed embedding vector is denoted by $H^{(\ell)}$.

2.2. Main results

2.2.1. EQUIVALENCE OF EXCEPTIONAL SETS

In previous work (Magner et al., 2020; Magner et al., 2020), it was established that for the set of norm-constrained GCNs with dimension $d = n$ and number of layers $K = D \log n$ for large enough constant D, the universal exceptional set is equal to the operational universal 0-exceptional set. Our first question, then, is whether or not this fact generalizes to lower embedding dimensions d. This is addressed in the following theorem.

Theorem 7 (Equivalence of exceptional sets) *For any $d \leq n$, the universal exceptional set with respect to the class $\mathcal{C}_d$ of GCNs is equal to the operational 0-exceptional set of graphon pairs (which we note is* not *dependent on d).*

Remark 8 *Theorem 7 has a noteworthy consequence: the embedding dimension has no effect on the universal exceptional set. However, it* does *affect the rate of convergence of the error probability in the ϵ_{res}-perturbed case.*

2.2.2. PROBABILITY OF ERROR IN THE MULTICLASS SETTING

Here, we consider the multi-hypothesis testing problem of distinguishing among k pairwise δ-exceptional graphons $W_0, W_1, ..., W_{k-1}$. In particular, we consider the setting where we choose an index B uniformly at random from the set $\{0, 1, ..., k-1\}$, a sample $G = G_n \sim W_B$ is drawn, and the task is to output an estimate $\hat{B}$ of B based on the ϵ_{res}-perturbed output embedding vector H of G. Here, H is as in the previous theorem.

Theorem 9 (Probability of error lower bound in terms of dimension) *In the setting of the multi-hypothesis testing problem with k hypotheses described above, the probability of error of any test based on the output embedding vector H is at least*

$$\Pr[\hat{B} \neq B] \geq 1 - \frac{d \cdot \log_2(1 + \frac{\delta}{n \cdot \epsilon_{res}}) + 1}{\log k}. \tag{5}$$

provided that the embedding dimension d is sufficiently large (i.e., there exists some fixed positive integer d_0 such that the above holds for every $d \geq d_0$).

As a consequence, if the signal to noise ratio $\delta/(n\epsilon_{res})$ is $\Omega(1)$, the embedding dimension must grow at least logarithmically as a function of the number of hypotheses in order for the probability of error to tend to 0. It may be seen by an appeal to bounds on the ϵ_{res}-packing number of L_∞ balls of radius δ/n that if d is superlogarithmic as a function of k, then there exist arrangements of limit points for which the optimal probability of error tends to 0. Thus, the order of (5) is correct.

The proof of this theorem relies on channel capacity results for the additive uniform noise channel.

2.2.3. FURTHER LINEAR ALGEBRAIC RESULTS

In the next results, we indicate another sense in which the value of d plays a role. In particular, while the exceptional set is typically of measure zero in the space of graphon pairs, we can still associate a notion of “size” of this set that is affected by the embedding dimension.

First, we consider the exceptional set for a GCN with *particular* (but arbitrary) choice of initial embedding matrix. We give a geometric description of this set in terms of the intersection of a hyperplane that is a function of the initial embedding matrix F, with dimension depending on d, and a set that is independent of F and d. We take the dimension of this hyperplane to be a measure of the “dimension” of the set of graphon pairs that this GCN cannot distinguish. The embedding dimension of the GCN impacts the minimum possible value of this dimension. To state our result, we need to formalize this notion. To guide our intuition, we will start by considering only *linear* GCNs with no weight matrices.

Consider two random walk matrices $\hat{A}^{(0)}$ and $\hat{A}^{(1)}$, which we think of as having been sampled (with an arbitrary coupling) from two graphons W_0, W_1. We note that $\hat{A}^{(0)\infty}$ and $\hat{A}^{(1)\infty}$ are $n \times n$ matrices with n repeated rows. We denote by π_0^T, π_1^T these row vectors, and we note that their elements respectively sum to 1. For an initial embedding matrix F, intuitively, W_0 and W_1 are indistinguishable by the corresponding GCN with sufficiently many layers if and only if $(\hat{A}^{(0)\infty} - \hat{A}^{(1)\infty}) \cdot F = 0$, which is the case if and only if $F^T(\pi_0 - \pi_1)^T = 0$. In other words, $(\pi_0 - \pi_1)^T \in \ker(F^T)$. We will thus be interested in the dimension of the kernel of a matrix related to F^T as a function of the embedding dimension. In particular, $\pi_0 - \pi_1$ cannot be an arbitrary vector, since each of π_0, π_1 is stochastic and, further, arises as the normalized degree sequence of a graph. This stochasticity implies the following:

$$\sum_{i=1}^{n}(\pi_{0,i} - \pi_{1,i}) = \sum_{i=1}^{n}\pi_{0,i} - \sum_{i=1}^{n}\pi_{1,i} = 1 - 1 = 0. \tag{6}$$

In other words, our interest is in the set of *balanced vectors* v such that $v \in \ker(F^T)$. With this in mind, we define $\hat{F}^T$ to be F^T with an additional row consisting of all 1s. Then our interest is in $\dim\ker(\hat{F}^T)$. We call this the *indistinguishability dimension* of the initial embedding matrix F. The following theorem establishes a connection between the embedding dimension of a GCN (now with nonlinearities and norm-constrained weight matrices and moderate depth) and the indistinguishability dimension of its initial embedding matrix.

Theorem 10 *For any $d \leq n$, consider a GCN F with embedding dimension $d \leq n$. For any $n \times d$ initial embedding matrix $\tilde{M}^{(0)}$, the indistinguishability dimension of $\tilde{M}^{(0)}$ is at least $n - d - 1$.*

Furthermore, there exist initial embedding matrices $\tilde{M}^{(0)}$ whose indistinguishability dimension is exactly $n - d - 1$.

In other words, as we decrease the embedding dimension, the lower bound on the indistinguishability dimension grows, reaching a maximum at $n - 2$, corresponding to $d = 1$.

Remark 11 *Frequently, the initial embedding matrix of a GCN is a function of the input graph. Theorem 10 does not consider this (more difficult) case. The reason for this is that without any constraints on this function, the distinguishing power of the resulting GCN can be unreasonably powerful. Thus, to prove results for such cases, one needs to restrict the class of functions that produce initial embedding matrices. One natural idea would be to consider initial embedding matrices such that the embedding vector for each vertex is a function only of the subgraph within radius r of that vertex. We do not explore such ideas further in the present work.*

The existence statement in the above theorem implies that there exist initial embedding matrices for which the resulting exceptional set has measure 0 in the set of all graphon pairs. In this sense, very little design or optimization is required to distinguish almost all graphon pairs.

We next indicate how to generalize results for linear GCNs to the more general GCNs that we consider in our theorem statements. This follows from Lemma 9 of Magner et al. (2020); Magner et al. (2020), stated below.

Theorem 12 (GCN generalization theorem Magner et al. (2020); Magner et al. (2020)) *Consider two random walk matrices $\hat{A}^{(0)}$ and $\hat{A}^{(1)}$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be in the class $\mathcal{A}$ of nice activation functions. Furthermore, let the sequence of weight matrices $W^{(0)}, ..., W^{(K)}$ satisfy the norm constraint conditions for some fixed positive constants C and D.*

Then, if $K \ll n^{1/2-\epsilon}$ for some arbitrarily small positive constant ϵ,

$$\|\hat{M}^{(0,K)} - \hat{M}^{(1,K)}\|_\infty \tag{7}$$

$$\leq \|\hat{A}^{(0)K}\hat{M}^{(0,0)}\prod_{j=0}^{K} W^{(j)} - \hat{A}^{(1)K}\hat{M}^{(1,0)}\prod_{j=0}^{K} W^{(j)}\|_\infty \tag{8}$$

$$\cdot (1 + o(1)). \tag{9}$$

In effect, this allows us to upper bound the L_∞ distance between two outputs of a given GCN (possibly with two different initial embedding matrices) with nonlinear activation functions by the distance between the outputs with the activation functions replaced by the identity. This allows us to directly generalize our results on linear GCNs to the case of nonlinear ones.

2.2.4. RESULTS ON COLORFUL GRAPHS

We next present results concerning the ability of GCNs to distinguish between different distributions on *colorful graphs*. In particular, we consider a very simple, but nontrivial, setting, where vertices of sample graphs are labeled with "colors" that are chosen from a fixed color distribution, independently of any graph structural information. We fix a finite alphabet A of colors, which may have cardinality growing with n. A colorful graph distribution in our simple setting is then parameterized by a pair (μ, W), where μ is a probability distribution on A and W is a graphon. Samples are constructed as follows:

1. Sample $G = G_n \sim W$.
2. For each vertex v of G, draw an independent sample $c = c(v) \sim \mu$. This is the color of vertex v in G. The result is a colorful graph.

As in the previous subsection, we seek to understand the limitations of GCNs in this setting, and so we consider first a simple case where we are to distinguish between (μ_0, W) and (μ_1, W), where μ_0 and μ_1 are two color distributions, but the distribution on graph structures is the same in both cases.

A few different questions are of interest:

- For a given pair of color distributions, do there exist graphons for which distinguishing these distributions is hard for "all" GCNs?
- What role does the embedding dimension of the GCN play in the distinguishability of pairs of color distributions?

Let us further constrain our GCNs: we insist on a particular manner of construction of the initial embedding matrix, which must necessarily depend on the vertex color information in order to distinguish between two color distributions on the same graph. In particular, to each color $c \in A$, we associate a row vector $v_c \in \mathbb{R}^d$, which we call a color embedding vector. With this mapping chosen, for a given colorful graph, we choose the initial embedding matrix whose row corresponding to vertex j in G is the vector $v_{c(j)}$.

To state our results, we need a few definitions.

Definition 13 (Confusing distributions) *Fix two color distributions μ_0, μ_1 on an alphabet A, along with a collection of color embeddings. A probability distribution $\pi \in \mathbb{R}^n$ is* confusing *for μ_0 and μ_1 if it satisfies $\pi \cdot (\hat{M}^{(0,0)} - \hat{M}^{(1,0)}) = 0$.In other words, $\pi^T \in \ker((\hat{M}^{(0,0)} - \hat{M}^{(1,0)})^T)$.This allows us to define a confusing graph: G is confusing for μ_0 and μ_1 if the stationary distribution of its random walk is a confusing distribution. A graphon W is confusing from μ_0, μ_1 if, with probability $1 - o(1)$ as $n \to \infty$, a graph $G_n \sim W$ is confusing for μ_0, μ_1.*

Definition 14 (Confusing dimension) *For two color distributions μ_0, μ_1 and any given n, we define the* confusing dimension *of μ_0, μ_1 to be the maximum dimension of the kernel of the matrix $(\hat{M}^{(0,0)} - \hat{M}^{(1,0)})^T$, optimized over all couplings of the two distributions. We denote the confusing dimension for μ_0, μ_1 and for a fixed collection of color embeddings, by $\mathcal{CD}(\mu_0, \mu_1)$.*

We have the following theorem regarding the set of degree distributions of graphs for which GCNs of moderate depth cannot distinguish between μ_0 and μ_1.

Theorem 15 (Confusing graphons for a color distribution pair) *Fix two color distributions μ_0, μ_1 on a finite alphabet A. The confusing dimension of μ_0, μ_1 for $n \to \infty$ is lower bounded as follows:*

$$\mathcal{CD}(\mu_0, \mu_1) \geq (1 - d_{TV}(\mu_0, \mu_1)) \cdot n \cdot (1 + O(1/\sqrt{n})), \tag{10}$$

with high probability.

In particular, this theorem says that the "dimension" of the set of confusing degree distributions is lower bounded by a decreasing linear function of the total variation distance between the color distributions.

Remark 16 *It is easy to see that this bound is tight, provided that d is large enough, that there are sufficiently many colors, and that their color embedding vectors are chosen to be linearly independent.*

In certain applications, the assumption of independence between vertex colors and graph structure does not hold, and, in fact, the color of a vertex is a deterministic function of its local graph structure. For example, in chemistry applications, one frequently considers graphs, representing chemical compounds, whose nodes are labeled by atom types (e.g., carbon, nitrogen, etc.). The type of an atom is usually only dependent on graph structure through the degree of the node (i.e., only a very local property). For instance, carbon atoms always have four bonds.

This motivates the following model for colorful graphs, in which vertex colors are dependent on very local properties of the graph. We again have a finite alphabet A of colors. We associate to each point in the unit interval a probability distribution on A (in particular, this may be encoded as a map $C : [0, 1] \to [0, 1]^{|A|}$, with the property that $\|C(x)\|_1 = 1$ for every $x \in [0, 1]$; we call such a map a *color distribution map*). A distribution on colorful graphs is then given by a pair (W, C), for a fixed graphon W. We sample a colorful graph as follows:

1. Sample a $G = G_n \sim W$ as before, with latent vertex positions $x_1, ..., x_n \in [0, 1]$.

2. For each vertex $j \in [n]$, sample $c(j) \sim C(x_j)$ independently of anything else.

We define the notions of confusing distributions, graphs, and graphons as we did previously, for a pair of color distribution maps C_0, C_1.

We have the following theorem, analogous to Theorem 15.

Theorem 17 (Confusing graphons for dependent colors) *Fix two color distribution maps C_0, C_1 on a finite alphabet A. Fix a collection of color embeddings. The confusing dimension of C_0, C_1 for $n \to \infty$ is lower bounded as follows:*

$$\mathcal{CD}(\mu_0, \mu_1) \geq (1 - \mathbb{E}[d_{TV}(C_0(X), C_1(X))]) \cdot n \cdot (1 + O(1/\sqrt{n})), \tag{11}$$

with high probability.

3. Conclusions

We considered in this paper the impact of the embedding dimension d of GCNs on their ability to distinguish pairs of graphons from their sample graphs. We found that the set of graphon pairs that are indistinguishable by "any" GCN does not depend on d. We showed a lower bound on the probability of error in terms of the embedding dimension, signal to noise ratio, and number of hypothesis graphons in the multiple hypothesis testing setting. We then extended the framework to colorful graphs (graphs whose nodes are labeled with colors from a finite alphabet), possibly with dependence between colors and local graph structure. We exhibited a connection between the distance between the color distributions and a notion of the "dimension" of the set of graphons on which a given pair of color distributions is indistinguishable.

We regard this as a step in the direction of a more complete understanding of the roles that different hyperparameters of GCNs play in their capabilities as representation learning methods.

4. Proofs

4.1. Proof of Theorem 7

In order to prove the claim, we need to show two things:

- that a 0-exceptional pair of graphons is in the universal exceptional set,
- and that a pair of graphons in the universal exceptional set is 0-exceptional.

Remark 18 *The proof below is for the case of identity activation functions. As stated earlier, this is sufficient to conclude the same result for the case of activation functions from the set $\mathcal{A}$, using Theorem 12.*

0-exceptional implies universal exceptional We will show the former by upper bounding the following quantity:

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty, \tag{12}$$

where we recall that $\hat{H}^{(b,K)}$ denotes the output graph embedding vector for the GCN with input $G_b \sim W_b$ (here, W_0, W_1 are a 0-exceptional pair). It will turn out that we can upper bound this by something that is independent of the dimension d.

We will first do this in the case where there are no weight matrices and where the activation function is the identity. We then give the details necessary for generalizing to the claimed setting.

Explicitly, by the triangle inequality,

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty \leq \|\hat{H}^{(0,K)} - \frac{1}{n}\mathbf{1}^T\hat{A}^{(0)\infty}\hat{M}^{(0)}\|_\infty \tag{13}$$

$$+ \|\frac{1}{n}\mathbf{1}^T(\hat{A}^{(0)\infty} - \hat{A}^{(1)\infty})\hat{M}^{(0)}\|_\infty \tag{14}$$

$$+ \|\hat{H}^{(1,K)} - \frac{1}{n}\mathbf{1}^T\hat{A}^{(1)\infty}\hat{M}^{(0)}\|_\infty \tag{15}$$

To proceed, we need a few lemmas.

Lemma 19 (Distance between limiting distributions (Magner et al., 2020; Magner et al., 2020)) *Suppose that W_0, W_1 satisfy $d_{deg}(W_0, W_1) = \delta$, and suppose further that $\delta > 0$. Furthermore, suppose that $G_0 \sim W_0$ and $G_1 \sim W_1$, both with n vertices, and with an arbitrary coupling. Then with probability $1 - o(1)$ as $n \to \infty$, we have $\|\pi_{G_0} - \pi_{G_1}\|_\infty = \frac{\delta}{n}(1 + O(1/\sqrt{n}))$.*

In the case where $\delta = 0$, we have that with probability $1 - o(1)$ as $n \to \infty$, $\|\pi_{G_0} - \pi_{G_1}\|_\infty = O(n^{-3/2+const})$.

Lemma 20 (Random walk matrix powers (Magner et al., 2020; Magner et al., 2020)) *Consider a Markov chain with transition matrix P and stationary matrix P_∞. Let $t_{mix}(P, \epsilon)$ denote the ϵ-total variation mixing time of P. For any $t \geq t_{mix}(P, \epsilon)$, we have that $\|P^t - P_\infty\|_\infty \leq 2\epsilon$.*

We can upper bound (13) and (15) using Lemma 20. The remaining term (14) can be upper bounded using Lemma 19. In particular, we have, using the submultiplicativity of operator norms, that

$$\|\hat{H}^{(b,K)} - \frac{1}{n}\mathbf{1}^T\hat{A}^{(b)\infty}\hat{M}^{(0)}\|_\infty = \|\frac{1}{n}\mathbf{1}^T(\hat{A}^{(b)K} - \hat{A}^{(b)\infty})\hat{M}^{(0)}\|_\infty \tag{16}$$

$$\leq \|\frac{1}{n}\mathbf{1}^T\|_{op,\infty}\|\hat{A}^{(b)K} - \hat{A}^{(b)\infty}\|_\infty \cdot \|\hat{M}^{(0)T}\|_{op,\infty} \tag{17}$$

$$= \|\hat{A}^{(b)K} - \hat{A}^{(b)\infty}\|_\infty \cdot \|\hat{M}^{(0)T}\|_{op,\infty}. \tag{18}$$

Provided that $K > t_{mix}(\hat{A}^{(b)}, \epsilon_0)$, for an arbitrary $\epsilon_0 > 0$, we thus have that (13) and (15) are both upper bounded by

$$(13), (15) \leq 2\epsilon_0\|\hat{M}^{(0)T}\|_{op,\infty}. \tag{19}$$

We can upper bound (14) by

$$\|\frac{1}{n}\mathbf{1}^T(\hat{A}^{(0)\infty} - \hat{A}^{(1)\infty})\hat{M}^{(0)}\|_\infty \leq \|\hat{A}^{(0)\infty} - \hat{A}^{(1)\infty}\|_\infty \cdot \|\hat{M}^{(0)T}\|_{op,\infty} \tag{20}$$

$$= \begin{cases} \frac{\delta}{n}(1 + O(1/\sqrt{n})) \cdot \|\hat{M}^{(0)T}\|_{op,\infty} & \delta > 0 \\ O(n^{-3/2+const}) \cdot \|\hat{M}^{(0)T}\|_{op,\infty} & \delta = 0 \end{cases} \tag{21}$$

Here, we have used Lemma 19.

Putting together (19) and (21), we have that

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty \leq \begin{cases} (4\epsilon_0 + \frac{\delta}{n}(1 + O(1/\sqrt{n}))) \cdot \|\hat{M}^{(0)T}\|_{op,\infty} & \delta > 0 \\ (4\epsilon_0 + O(n^{-3/2+const})) \cdot \|\hat{M}^{(0)T}\|_{op,\infty} & \delta = 0 \end{cases} \tag{22}$$

We choose $\epsilon_0 = 1/n^2$, which results in a mixing time of $\Theta(\log n)$. Thus, we finally get

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty \leq \begin{cases} \frac{\delta}{n}(1 + O(1/\sqrt{n})) \cdot \|\hat{M}^{(0)T}\|_{op,\infty} & \delta > 0 \\ O(n^{-3/2+const}) \cdot \|\hat{M}^{(0)T}\|_{op,\infty} & \delta = 0 \end{cases} \tag{23}$$

We note, crucially, that $\|\hat{M}^{(0)T}\|_{op,\infty}$ does not depend on d. That is, $\|\hat{M}^{(0)T}\|_{op,\infty}$ can take any non-negative real value, regardless of dimension. amAre we sure that this is all that we need? How do we generalize this to nonlinear activation functions and $d \times d$ weight matrices?

In particular, the implication is that closeness of degree profiles implies closeness of embedding vectors, *regardless of embedding dimension*. Furthermore, we note that the inequality in the $\delta = 0$ case implies the claimed result that the 0-exceptional set is a subset of the universal exceptional set.

Universal exceptional implies 0-exceptional In order to show the reverse inclusion, we will show the contrapositive: that a δ-separated pair, for some $\delta > 0$, is not in the universal exceptional set. To show this, we will exhibit an explicit construction of a GCN that can distinguish the pair. We assume that $d_{deg}(W_0, W_1) > \delta$. We will show that, for any $d \leq n$, there exists a choice of $\hat{M}^{(0)}$ (which is $n \times d$) such that the resulting embedding vectors $\hat{H}^{(b,K)}$ are well-separated in L_∞. We reason as follows: for the two stationary distribution row vectors π_0, π_1, from Lemma 19, we have $\|\pi_0 - \pi_1\|_\infty = \frac{\delta}{n}(1 + O(1/\sqrt{n}))$. We will use this to lower bound the quantity $\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty$. As weight matrices, we will choose $d \times d$ identity matrices. As an initial embedding matrix, we cannot choose an identity matrix. Instead, we need to choose a matrix $\hat{M}^{(0)}$ such that $(\pi_0 - \pi_1)\hat{M}^{(0)} = \Omega(1/n)$ as $n \to \infty$. We consider the initial embedding matrix all of whose columns are copies of $n \cdot (\pi_0 - \pi_1)^T$. With this definition in hand, we will show a lower bound on the L_∞ distance between the final embedding vectors of $\delta^2/n \cdot (1 + o(1))$. We have

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty = \|\frac{1}{n}\mathbf{1}^T(\hat{A}^{(0)K} - \hat{A}^{(1)K})\hat{M}^{(0)}\|_\infty \tag{24}$$

$$= \|\frac{1}{n}\mathbf{1}^T(\hat{A}^{(0)K} - \hat{A}^{(0)\infty} + \hat{A}^{(0)\infty} - \hat{A}^{(1)\infty} + \hat{A}^{(1)\infty} - \hat{A}^{(1)K})\hat{M}^{(0)}\|_\infty. \tag{25}$$

By the reverse triangle inequality, we can lower bound as follows:

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty \geq \|\frac{1}{n}\mathbf{1}^T(\hat{A}^{(0)\infty} - \hat{A}^{(1)\infty})\hat{M}^{(0)}\|_\infty \tag{26}$$

$$- \left| \|\frac{1}{n}\mathbf{1}^T(\hat{A}^{(0)K} - \hat{A}^{(0)\infty})\hat{M}^{(0)}\|_\infty + \|\frac{1}{n}\mathbf{1}^T(\hat{A}^{(1)K} - \hat{A}^{(1)\infty})\hat{M}^{(0)}\|_\infty \right|. \tag{27}$$

The first term is the dominant one, and we can handle it as follows:

$$\|\frac{1}{n}\mathbf{1}^T(\hat{A}^{(0)\infty} - \hat{A}^{(1)\infty})\hat{M}^{(0)}\|_\infty = \|(\pi_0 - \pi_1)\hat{M}^{(0)}\|_\infty. \tag{28}$$

Now, by our choice of $\hat{M}^{(0)}$, the vector in this norm consists of d copies of $n \cdot \|(\pi_0 - \pi_1)\|_2^2$, which is at least $\frac{\delta^2}{n} \cdot (1 + O(1/n))$, where we have used Lemma 19.

Now, the remaining terms (27) can be upper bounded in absolute value as follows: we apply Lemma 20, with $K > \max\{t_{mix}(\hat{A}^{(0)}, 1/n^2), t_{mix}(\hat{A}^{(1)}, 1/n^2)\} = \Theta(\log n)$ in order to conclude that the resulting norm is $O(1/n^2)$. This implies that $\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty = \Omega(\delta/n)$, as claimed. This, in turn, implies that a δ-separated pair of graphons does not lie in the universal exceptional set.

This concludes the proof of Theorem 7.

4.2. Proof of Theorem 9

To prove Theorem 9, we will need a few auxiliary results. We will identify each hypothesis graphon W_i with a resulting unperturbed embedding vector $\hat{H}^{(i)}$. The idea is to relate the probability of error to the capacity of the *uniform additive noise channel*. This capacity is bounded in an exercise from Cover and Thomas (2006) (see also Rioul and Magossi (2014)).

Lemma 21 (Capacity result for the uniform additive noise channel) *Consider the uniform additive noise channel with input X satisfying $|X| \leq A$ and output given by $Y = X + Z$, where Z is independent of X and is uniformly distributed in the interval $[-\Delta, \Delta]$. The capacity C' of this channel satisfies $C' \leq \log_2(1 + A/\Delta)$.*

Our proof requires a slight tweak of the proof of this lemma. In particular, we have

$$I(X^d; Y^d) \leq h(Y^d) - h(Y^d \mid X^d), \tag{29}$$

where $h(\cdot \mid \cdot)$ here is the conditional differential entropy.

The second term is dependent only on the distribution of Z^d, the i.i.d. uniform noise. The first term is maximized, subject to the constraint that $\|Y^d\|_\infty \leq A + \Delta$, by a d-dimensional vector consisting of i.i.d. uniform random variables on $[-(A + \Delta), A + \delta]$. Simple algebra completes the generalization that we need. Crucially, this derivation does not depend on the distribution of X^d, and, in particular, it is not necessary that it consist of i.i.d. random variables.

Thus, we have that for any sequence of joint distributions $X^d \in \mathbb{R}^d$, $d \to \infty$, with $\|X^d\|_\infty \leq A$, and vectors $Y^d = (X_1 + Z_1, X_2 + Z_2, ...X_d + Z_d)$, $\limsup_{d\to\infty} \frac{I(X^d;Y^d)}{d} \leq \log_2(1 + A/\Delta)$. Here, $I(\cdot;\cdot)$ is the Shannon mutual information. In other words, provided that d is large enough,

$$I(X^d; Y^d) \leq d \cdot \log_2(1 + A/\Delta). \tag{30}$$

We will apply the mutual information form of Fano's inequality to upper bound $\Pr[\hat{B} = B]$. We note that both random variables are supported on a set of size k. Then we have $\Pr[B = \hat{B}] \leq \frac{I(B;H)+1}{\log k}$.Now, $I(B; H) \leq I(\hat{H}; H)$, by the Markov property relating H, $\hat{H}$, and B, and by the data processing inequality. Finally, we can upper bound $I(\hat{H}; H)$ using (30), provided that d is large enough, which gives us $\Pr[B = \hat{B}] \leq \frac{d\cdot\log_2(1+\frac{\delta}{n\cdot\epsilon_{res}})+1}{\log k}$.This completes the proof.

4.3. Proof of Theorem 10

We note that the initial embedding matrix is $n \times d$, and its transpose is $d \times n$. Adding the row of 1s results in a $(d + 1) \times n$ matrix. If $d \leq n - 1$, the rank of this matrix is at most $d + 1$, so that the dimension of its kernel, by the rank-nullity theorem, is at least $n - (d + 1) = n - d - 1$, as desired.

4.4. Proof of Theorem 15

We will consider a coupling of μ_0 and μ_1 and upper bound the rank of the matrix $(\hat{M}^{(0,0)} - \hat{M}^{(1,0)})^T$ in probability. In particular, we can choose the maximal coupling. Then, for any graphon W, two colorful graphs (or, equivalently, a single graph on which every vertex is labeled with two colors) can be chosen as follows: we generate a graph $G \sim W$, and every vertex is assigned two colors according to the coupling.

Under this distribution, the number of zero rows in $\hat{M}^{(0,0)} - \hat{M}^{(1,0)}$ is lower bounded by a binomially distributed random variable with parameter $p = 1 - d_{TV}(\mu_0, \mu_1)$. This immediately implies that the number of zero rows is at least $(1 - d_{TV}(\mu_0, \mu_1)) \cdot n \cdot (1 + O(1/\sqrt{n}))$ with probability $1 - e^{-\Theta(n)}$, which, by the rank-nullity theorem, immediately implies the desired result.

4.5. Proof of Theorem 17

Given two distributions (W, C_0) and (W, C_1) on colorful graphs, we consider the following coupling: a single (non-colorful at this point) graph G is generated according to W, with latent positions $x_1, ..., x_n \in [0, 1]$. What remains is to choose two color maps for the vertices of G. For each x_j, we choose the maximal coupling of the distributions $C_0(x_j)$ and $C_1(x_j)$ and generate the colors according to this coupling. Note that the colors of different vertices are generated independently, conditioned on the latent positions. The result is that, for each $j \in [n]$, conditioned on the x_is, we have that the probability that the two colors of vertex j are not equal is given by $d_{TV}(C_0(x_j), C_1(x_j))$.

This, in turn, implies that the *unconditional* probability that the two colors of any given vertex are different is given by $\mathbb{E}[d_{TV}(C_0(X), C_1(X))]$, where the expectation is with respect to the uniformly random choice of X from the unit interval.

With this result in hand, we find via the Chernoff bound that the number of zero rows in $\hat{M}^{(0,0)} - \hat{M}^{(1,0)}$ is, with probability at least $1 - e^{-\Theta(n)}$, lower bounded by $(1 - \mathbb{E}[d_{TV}(C_0(X), C_1(X))]) \cdot n$,and this immediately gives us the lower bound on the dimension of the kernel of $\hat{M}^{(0,0)} - \hat{M}^{(1,0)}$, as claimed.

References

Zhengdao Chen, Soledad Villar, Lei Chen, and Joan Bruna. On the equivalence between graph isomorphism testing and function approximation with gnns. *arXiv preprint arXiv:1905.12560*, 2019.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006. ISBN 0471241954.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3844–3852, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9. URL http://dl.acm.org/citation.cfm?id=3157382.3157527.

Vikas K. Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks, 2020.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1737–1746. JMLR.org, 2015.

William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Eng. Bull.*, 40:52–74, 2017.

Tatsuro Kawamoto, Masashi Tsubaki, and Tomoyuki Obuchi. Mean-field theory of graph neural networks in graph partitioning. In *Advances in Neural Information Processing Systems*, pages 4361–4371, 2018.

Nicolas Keriven, Alberto Bietti, and Samuel Vaiter. Convergence and stability of graph convolutional networks on large random graphs, 2020.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.

László Lovász and Balázs Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933 – 957, 2006. ISSN 0095-8956. doi: https://doi.org/10.1016/j.jctb.2006.05.002. URL http://www.sciencedirect.com/science/article/pii/S0095895606000517.

Abram Magner, Mayank Baranwal, and Alfred O. Hero III. The power of graph convolutional networks to distinguish random graph models. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2664–2669, 2020.

Abram Magner, Mayank Baranwal, and Alfred O. Hero III. Fundamental limits of deep graph convolutional networks. *arXiv preprint arXiv:1910.12954*, 2020.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and lehman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4602–4609, 2019.

Olivier Rioul and José Magossi. On shannon's formula and hartley's rule: Beyond the mathematical coincidence. *Entropy*, 16(9):4892–4910, Sep 2014. ISSN 1099-4300. doi: 10.3390/e16094892. URL http://dx.doi.org/10.3390/e16094892.

Charbel Sakr, Yongjune Kim, and Naresh Shanbhag. Analytical guarantees on numerical precision of deep neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3007–3016, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/sakr17a.html.

Saurabh Verma and Zhi-Li Zhang. Stability and generalization of graph convolutional neural networks. *arXiv preprint arXiv:1905.01004*, 2019.

B. Yu. Weisfeiler and A. A. Lehman. Reduction of a graph to a canonical form and an algebra arising during this reduction (in Russian). *Nauchno-Technicheskaya Informatsia, Seriya*, 2(9):12–16, 1968.

Felix Wu, Amauri H. Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *ICML*, 2019.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.