The Power of Graph Convolutional Networks to Distinguish Random Graph Models

Abram Magner
University at Albany, SUNY
Albany, NY, USA
Email: amagner@albany.edu

Mayank Baranwal
University of Michigan
Ann Arbor, MI, USA
Email: mayankb@umich.edu

Alfred O. Hero III
University of Michigan
Ann Arbor, MI, USA
Email: hero@eecs.umich.edu

***Abstract*—Graph convolutional networks (GCNs) are a widely used method for graph representation learning. To elucidate the capabilities and limitations of GCNs, we investigate their power, as a function of their number of layers, to distinguish between different random graph models (corresponding to different class-conditional distributions in a classification problem) on the basis of the embeddings of their sample graphs. In particular, the graph models that we consider arise from graphons, which are the most general possible parameterizations of infinite exchangeable graph models and which are the central objects of study in the theory of dense graph limits. We give a precise characterization of the set of pairs of graphons that are indistinguishable by a GCN with nonlinear activation functions coming from a certain broad class if its depth is at least logarithmic in the size of the sample graph. This characterization is in terms of a degree profile closeness property. Outside this class, a very simple GCN architecture suffices for distinguishability. We then exhibit a concrete, infinite class of graphons arising from stochastic block models that are well-separated in terms of cut distance and are indistinguishable by a GCN. These results theoretically match empirical observations of several prior works on GCNs. To prove our results, we exploit a connection to random walks on graphs.**

I. Introduction

In applications ranging from drug discovery [1] and design to proteomics [2] to neuroscience [3] to social network analysis [4], inputs to machine learning methods take the form of graphs. In order to leverage the empirical success of deep learning and other methods that work on vectors in finite-dimensional Euclidean spaces for supervised learning tasks in this domain, a plethora of graph representation learning schemes have been proposed and used [5]. One particularly effective such method is the *graph convolutional network* (GCN) architecture [6], [7]. A graph convolutional network works by associating with each node of an input graph a vector of features and passing these node features through a sequence of *layers*, resulting in a final set of node vectors, called node embeddings. To generate a vector representing the entire graph, these final embeddings are sometimes averaged. Each layer of the network consists of a graph diffusion step, where a node's feature vector is averaged with those of its neighbors; a feature transformation step, where each node's vector is transformed by a weight matrix; and, finally, application of an elementwise nonlinearity such as the ReLU or sigmoid function. The weight matrices are trained from data, so that the metric structure of the resulting embeddings are (one hopes) tailored to a particular classification task.

While GCNs and other graph representation learning methods have been successful in practice, numerous theoretical questions about their capabilities and the roles of their hyperparameters remain unexplored. In this paper, we give results on the ability of GCNs to distinguish between samples from different random graph models. We focus on the roles that the number of layers and the presence or absence of nonlinearity play. The random graph models that we consider are those that are parameterized by *graphons* [8], which are functions from the unit square to the interval $[0, 1]$ that essentially encode edge density among a continuum of vertices. Graphons are the central objects of study in the theory of dense graph limits and, by the Aldous-Hoover theorem [9] exactly parameterize the class of infinite exchangeable random graph models – those models whose samples are invariant in distribution under permutation of vertices.

A. Prior Work

A survey of modern graph representation learning methods is provided in [5]. Graph convolutional networks were first introduced in [7], and since then, many variants have been proposed. For instance, the polynomial convolutional filters in the original work were replaced by linear convolutions [6]. Authors in [10] modified the original architecture to include gated recurrent units for working with dynamical graphs. These and other variants have been used in various applications, e.g., [11], [12], [13], [14].

Theoretical work on GCNs has been from a variety of perspectives. In [15], the authors investigated the generalization and stability properties of GCNs. Several works, including [16], [17], [18], have drawn connections between the representation capabilities of GCNs and the distinguishing ability of the *Weisfeiler-Lehman* (WL) algorithm for graph isomorphism testing [19]. These papers drawing comparisons to the WL algorithm implicitly study the injectivity properties of the mapping from graphs to vectors induced by GCNs. However, they do not address the metric/analytic properties, which are important in consideration of their performance as representation learning methods [20]. Finally, at least one work has considered the performance of untrained GCNs on community detection [21]. The authors of that paper provide

a heuristic calculation based on the mean-field approximation from statistical physics and demonstrate through numerical experiments the ability of untrained GCNs to detect the presence of clusters and to recover the ground truth community assignments of vertices in the stochastic block model. They empirically show that the regime of graph model parameters in which an untrained GCN is successful at this task agrees well with the analytically derived detection threshold. The authors also conjecture that training GCNs does not significantly affect their community detection performance.

The theory of graphons as limits of dense graph sequences was initiated in [22] and developed by various authors [23], [24]. For a comprehensive treatment, see [8].

Several authors have investigated the problem of estimation of graphons from samples [25], [26], [27]. Our work is complementary to these, as our goal is to investigate the performance of a *particular* method on the problem of distinguishing graphons.

B. Our Contributions

We first establish a convergence result for GCN embedding vectors, which will give a lower bound on the probability of error of *any* test that attempts to distinguish between two graphons based on slightly perturbed K-layer GCN embedding matrices of sample graphs of size n, provided that $K = \Omega(\log n)$. In particular, we exhibit a family of pairs of graphons that are hard for any test to distinguish on the basis of these embeddings. This is the content of Theorems 1 and 2.

We then show a converse achievability result in Theorem 3 that says, roughly, that provided that the number of layers is sufficiently large ($K = \Omega(\log n)$), there exists a *linear* GCN architecture with a very simple sequence of weight matrices and a choice of initial embedding matrix such that pairs of graphons whose expected degree statistics differ by a sufficiently large amount are distinguishable from the noise-perturbed GCN embeddings of their sample graphs. In other words, this indicates that the family of difficult-to-distinguish graphons alluded to above is essentially the *only* sort of case in which a nonlinear GCN architecture could be necessary (though, as Theorem 2 shows, for several choices of activation functions, these graphons are still indistinguishable).

Our proofs rely on concentration of measure results and techniques from the theory of Markov chain mixing times and spectral graph theory [28].

1) Relations between probability of error lower and upper bounds: Our probability of error lower bounds give theoretical backing to a phenomenon that has been observed empirically in graph classification problems: adding arbitrarily many layers (more than $\Theta(\log n)$) to a GCN can substantially degrade classification performance. This is an implication of Theorem 2. On the other hand, Theorem 3 shows that this is *not* always the case, and that for *many* pairs of graphons, adding more layers improves classification performance. We suspect that the set of pairs of graphons for which adding arbitrarily many layers does not help forms a set of measure 0, though this does not imply that such examples never arise in practice.

The factor that determines whether or not adding layers will improve or degrade performance of a GCN in distinguishing between two graphons W_0 and W_1 is the distance between the stationary distributions of the random walks on the sample graphs from W_0 and W_1. This, in turn, is determined by the normalized degree profiles of the sample graphs.

An extended version of this paper is available on ArXiv [29].

II. Notation and Model

A. Graph Convolutional Networks

We start by defining the model and relevant notation. A K-layer graph convolutional network (GCN) is a function mapping graphs to vectors over $\mathbb{R}$. It is parameterized by a sequence of K *weight matrices* $W^{(j)} \in \mathbb{R}^{d\times d}$, $j \in \{0, ..., K-1\}$, where $d \in \mathbb{N}$ is the *embedding dimension*, a hyperparameter. 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 $\hat{M}^{(0)}$, 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 $\hat{M}^{(\ell)}$:

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

Typical examples of activation functions in neural network and GCN contexts include the ReLU, sigmoid, and hyperbolic tangent functions. Empirical work has given evidence that the performance of GCNs on certain classification tasks is unaffected by replacing nonlinear activation functions by the identity [30]. Our results lend theoretical credence to this.

Frequently, $\hat{A}$ is replaced by either the normalized adjacency matrix $D^{-1/2}AD^{-1/2}$, where D is a diagonal matrix with the degrees of the vertices of the graph on the diagonal, or some variant of the Laplacian matrix $D - A$. For simplicity, we will consider in this paper only the choice of $\hat{A}$.

The defining equation (1) has the following interpretation: multiplication on the left by $\hat{A}$ has the effect of replacing each node's embedding vector with the average of those of its neighbors. Multiplication on the right by the weight matrix $W^{(\ell-1)}$ has the effect of replacing each coordinate (corresponding to a feature) of each given node embedding vector with a linear combination of values of the node's features in the previous layer.

B. Graphons

In order to probe the ability of GCNs to distinguish random graph models from samples, we consider the task of distinguishing random graph models induced by graphons.

A graphon W is a symmetric, Lebesgue-measurable function from $[0,1]^2 \to [0,1]$. To each graphon is associated a natural exchangeable random graph model as follows: to generate a graph on n vertices, one chooses n points $x_1, ..., x_n$ uniformly at random from $[0,1]$. An edge between vertices i, j is independent of all other edge events and is present with probability $W(x_i, x_j)$. We use the notation $G \sim W$ to denote that G is a random sample graph from the model induced by W. The number of vertices will be clear from context.

One commonly studied class of models that may be defined equivalently in terms of sampling from graphons is the class of stochastic block models. A stochastic block model on n vertices with two blocks is parameterized by four quantities: k_1, p_1, p_2, q. The two blocks of vertices have sizes $k_1 n$ and $k_2 n = (1-k_1)n$, respectively. Edges between two vertices v, w in block i, $i \in \{1,2\}$, appear with probability p_i, independently of all other edges. Edges between vertices v in block 1 and w in block 2 appear independently with probability q. We will write this model as $\mathrm{SBM}(p_1, p_2, q)$, suppressing k_1.

An important metric on graphons is the *cut distance* [31]. It is induced by the cut norm, which is defined as follows: fix a graphon W. Then

$$\|W\|_{cut} = \sup_{S,T} \left| \int_{S \times T} W(x,y)\ d\mu(x)\ d\mu(y) \right|, \tag{3}$$

where the supremum is taken over all measurable subsets of $[0,1]$, and the integral is taken with respect to the Lebesgue measure. For finite graphs, this translates to taking the pair of subsets S, T of vertices that has the maximum between-subset edge density. The cut *distance* $d_{cut}(W_0, W_1)$ between graphons W_0, W_1 is then defined as

$$d_{cut}(W_0, W_1) = \inf_{\phi} \|W_0 - W_1(\phi(\cdot), \phi(\cdot))\|_{cut}, \tag{4}$$

where the infimum is taken over all measure-preserving bijections of $[0,1]$. In the case of finite graphs, this intuitively translates to ignoring vertex labelings. The cut distance generates the same topology on the space of graphons as convergence of subgraph homomorphism densities (i.e., *left convergence*), and so it is an important part of the theory of graph limits.

C. Main Hypothesis Testing Problem

We may now state the hypothesis testing problem under consideration. Fix two graphons W_0, W_1. A coin $B \sim \mathrm{Bernoulli}(1/2)$ is flipped, and then a graph $G \sim W_B$ on n vertices is sampled. Next, G is passed through $K = K(n)$ layers of a GCN, resulting in a matrix $\hat{M}^{(K)} \in \mathbb{R}^{n \times d}$ whose rows are node embedding vectors. The graph embedding vector $\hat{H}^{(K)}$ is then defined to be $\frac{1}{n}\mathbf{1}^T \hat{M}^{(K)}$. As a final step, the embedding vector is perturbed in each entry by adding an independent, uniformly random number in the interval $[-\epsilon_{res}, \epsilon_{res}]$, for a parameter $\epsilon_{res} > 0$ that may depend on n, which we will typically consider to be $\Theta(1/n)$. This results in a vector $H^{(K)}$. We note that this perturbation step has precedent in the context of studies on the performance of neural networks in the presence of numerical imprecision [32]. For our purposes, it will allow us to translate convergence results to information theoretic lower bounds.

Our goal is to study the effect of the number of layers K and presence or absence of nonlinearities on the representation properties of GCNs and probability of error of optimal tests $\Psi(H^{(K)})$ that are meant to estimate B. Throughout, we will consider the case where $d = n$. We will frequently use two particular norms: the ℓ_∞ norm for vectors and matrices, which is the maximum absolute entry; and the operator norm induced by ℓ_∞ for matrices: for a matrix M,

$$\|M\|_{op,\infty} = \sup_{v\ :\ \|v\|_\infty = 1} \|Mv\|_\infty. \tag{5}$$

III. Main Results

To state our results, we need a few definitions. For a graphon W, we define the degree function $d_W : [0,1] \to \mathbb{R}$ to be

$$d_W(x) = \int_0^1 W(x,y)\ dy, \tag{6}$$

and define the total degree function

$$D(W) = \int_0^1 \int_0^1 W(x,y)\ dx\ dy. \tag{7}$$

We will assume in what follows that all graphons W have the property that there is some $\ell > 0$ for which $W(x,y) \geq \ell$ for all $x, y \in [0,1]$.

For any $\delta \geq 0$, we say that two graphons W_0, W_1 are a *δ-exceptional* pair if

$$\int_0^1 \left| \frac{d_{W_0}(\phi(x))}{D(W_0)} - \frac{d_{W_1}(x)}{D(W_1)} \right| dx \leq \delta, \tag{8}$$

for some measure-preserving bijection $\phi : [0,1] \to [0,1]$. If a pair of graphons is not δ-exceptional, then we say that they are δ-separated.

We define the following class of activation functions:

Definition 1 (Nice activation functions)**.** *We define $\mathcal{A}$ to be the class of activation functions $\sigma : \mathbb{R} \to \mathbb{R}$ satisfying the following conditions:*

- $\sigma \in C^2$.
- $\sigma(0) = 0$, $\sigma'(0) = 1$ *and* $\sigma'(x) \leq 1$ *for all* x.

For simplicity, in Theorems 1 and 2 below, we will consider activations in the above class; however, some of the conditions may be relaxed without inducing changes to our results: in particular, we may remove the requirement that $\sigma'(0) = 1$, and we may relax $\sigma'(x) \leq 1$ for all x to only hold for x in some constant-length interval around 0. This expanded class includes activation functions such as $\sigma(x) = \tanh(x)$ and the *swish* and *SELU* functions:

- swish [33]: $\sigma(x) = \frac{x}{1+e^{-x}}$
- SELU [34]: $\sigma(x) = I[x \leq 0](e^x - 1) + I[x > 0]x$.

We also make the following stipulation about the parameters of the GCN: the initial embedding matrices $\hat{M}^{(b,0)}$ (with $b \in \{0,1\}$) and weight matrices $\{W^{(j)}\}_{j=0}^K$ satisfy

$$\left\| M^{(b,0)T} \right\|_{op,\infty} \cdot \prod_{j=0}^{K} \|W^{(j)T}\|_{op,\infty} \leq C, \tag{9}$$

and $\sum_{j=0}^{K} \|W^{(j)T}\|_{op,\infty} \leq E$, for some fixed positive constants C and E.

Theorem 1 (Convergence of embedding vectors for a large class of graphons and for a family of nonlinear activations)**.** *Let W_0, W_1 denote two δ-exceptional graphons, for some fixed $\delta \geq 0$.*

Let K satisfy $D \log n < K$, for some large enough constant $D > 0$ that is a function of W_0 and W_1. Consider the GCN with K layers and output embedding matrix $\hat{M}^{(K)}$, with the additional properties stated before the theorem.

Suppose that $\delta > 0$. Then there exists a coupling (alternatively, a relabeling and an arbitrary coupling) of the graphs $G^{(0)} \sim W_0, G^{(1)} \sim W_1$, as $n \to \infty$ such that the embedding vectors $\hat{H}^{(0,K)}$ and $\hat{H}^{(1,K)}$ satisfy

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty \leq \frac{\delta}{n}(1 + O(1/\sqrt{n})) \tag{10}$$

with high probability.

If $\delta = 0$, then we have

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty \leq O(n^{-3/2+const}), \tag{11}$$

and for a $1 - o(1)$-fraction of coordinates i, $|\hat{H}_i^{(0,K)} - \hat{H}_i^{(1,K)}| = O(1/n^2)$.

Remark 1. *We stress that the above convergence bounds are for the* unperturbed *GCN embedding vectors.*

Remark 2. *The convergence bounds (10) and (11) should be interpreted in light of the fact that the embedding vectors have entries on the order of $\Theta(1/n)$.*

Theorem 1 can be translated, with some effort, to the following result.

Theorem 2 (Probability of error lower bound)**.** *Consider again the setting of Theorem 1. Furthermore, suppose that $\epsilon_{res} > \frac{\delta}{2n}$. Let K additionally satisfy $K \ll n^{1/2-\epsilon_0}$, for an arbitrarily small fixed $\epsilon_0 > 0$. Then there exist two sequences $\{\mathcal{G}_{0,n}\}_{n=1}^{\infty}, \{\mathcal{G}_{1,n}\}_{n=1}^{\infty}$ of random graph models such that*

- *with probability 1, samples $G_{b,n} \sim \mathcal{G}_{b,n}$ converge in cut distance to W_b,*
- *When $\delta > 0$, the probability of error of any test in distinguishing between W_0 and W_1 based on $H^{(b,K)}$, the ϵ_{res}-uniform perturbation of $\hat{H}^{(b,K)}$, is at least*

$$\left(1 - \frac{\delta}{2\epsilon_{res} n}\right)^n \tag{12}$$

When $\delta = 0$, the probability of error lower bound becomes

$$\exp\left(-\frac{const}{\epsilon_{res} \cdot n}\right). \tag{13}$$

Remark 3. *When $\epsilon_{res} = \Theta(1/n)$ and $\delta = \Omega(1)$, the error probability lower bound (12) is exponentially decaying to 0. On the other hand, when $\epsilon_{res} \gg 1/n$ and $\delta = \Omega(1)$, it becomes $\exp\left(-\frac{\delta}{2\epsilon_{res}}\right)(1 + o(1))$, which is $\Theta(1)$.*

When $\delta = 0$ and $\epsilon_{res} = \Omega(1/n)$, the probability of error lower bound in (13) is $\Omega(1)$.

We next turn to a positive result demonstrating the distinguishing capabilities of very simple, linear GCNs.

Theorem 3 (Distinguishability result)**.** *Let W_0, W_1 denote two δ-separated graphons. Then there exists a test that distinguishes with probability $1 - o(1)$ between samples $G \sim W_0$ and $G \sim W_1$ based on the ϵ_{res}-perturbed embedding vector from a GCN with K layers, identity initial and weight matrices, and ReLU activation functions, provided that $K > D \log n$ for a sufficiently large D and that $\epsilon_{res} \leq \frac{\delta}{2n}$.*

The convergence rate of the probability in the above theorem is available in the full version of this paper.

Finally, we exhibit a family of stochastic block models that are difficult to distinguish and are such that infinitely many pairs of them have large cut distance.

To define the family of models, we consider the following density parameter set: we pick a base point $P_* = (p_{*,1}, p_{*,2}, q_*)$ with all positive numbers and then define

$$\mathcal{P} = \left\{ P \ : \ (0,0,0) \prec P = P_* + \tau \cdot (\frac{1}{k_1}, \frac{k_1}{k_2^2}, \frac{-1}{k_2}) \preceq (1,1,1) \right\},$$

where $\preceq$ is the lexicographic partial order, and $\tau \in \mathbb{R}$. We have defined this parameter family because the corresponding SBMs all have equal expected degree sequences.

It may be checked that δ in Theorems 1 and 2 is 0 for pairs of graphons from $\mathcal{P}$. This gives the following result.

Theorem 4. *For any pair W_0, W_1 from the family of stochastic block models parameterized by $\mathcal{P}$, there exists a $K > D \log n$, for some large enough positive constant D, such that the following statements hold:*

a) Convergence of embedding vectors: There is a coupling (alternatively, there is a relabeling and an arbitrary coupling) of the graphs $G^{(0)} \sim W_0$ and $G^{(1)} \sim W_1$ such that, as $n \to \infty$, the embedding vectors $\hat{H}^{(0,K)}$ and $\hat{H}^{(1,K)}$ satisfy

$$\|\hat{H}^{(0,K)} - \hat{H}^{(1,K)}\|_\infty = O(n^{-3/2+const}) \tag{14}$$

with probability $1 - e^{-\Theta(n)}$.

b) Probability of error lower bound: Let K additionally satisfy $K \ll n^{1/2-\epsilon_0}$, for an arbitrary small fixed $\epsilon_0 > 0$. Then there exist two sequences $\{\mathcal{G}_{0,n}\}_{n=1}^{\infty}$, $\{\mathcal{G}_{1,n}\}_{n=1}^{\infty}$ of random graph models such that

- *with probability 1, samples $G_{b,n} \sim \mathcal{G}_{b,n}$ converge in cut distance to W_b,*
- *the probability of error of any test in distinguishing between W_0 and W_1 based on $H^{(b,K)}$, the ϵ_{res}-uniform perturbation of $\hat{H}^{(b,K)}$, is lower bounded by $\exp\left(-\frac{C}{\epsilon_{res} n}\right)$.*

IV. Conclusions and future work

We have shown conditions under which GCNs are information-theoretically capable/incapable of distinguishing between sufficiently well-separated graphons.

It is worthwhile to discuss what lies ahead for the theory of graph representation learning in relation to the problem of

distinguishing distributions on graphs. As the present paper is a first step, we have left several directions for future exploration. Most immediately, although we have proven impossibility results for GCNs with nonlinear activation functions, we lack a complete understanding of the benefits of more general ways of incorporating nonlinearity. We have shown that architectures with too many layers cannot be used to distinguish between graphons coming from a certain exceptional class. It would be of interest to determine if more general ways of incorporating nonlinearity are able to generically distinguish between any sufficiently well-separated pair of graphons, whether or not they come from the exceptional class. To this end, we are exploring results indicating that replacing the random walk matrix $\hat{A}$ in the GCN architecture with the transition matrix of a related Markov chain with the same graph structure as the input graph G results in a linear GCN that is capable of distinguishing graphons generically.

Furthermore, a clear understanding of the role played by the embedding dimension would be of interest. In particular, we suspect that decreasing the embedding dimension results in worse graphon discrimination performance. Moreover, a more precise understanding of how performance parameters scale with the embedding dimension would be valuable in GCN design.

Additionally, different noise models may be important in practice: for instance, one may consider perturbation of the sample graphs or adversarial (but bounded) perturbation of the output embedding vectors.

Finally, we note that in many application domains, graphs are typically sparse. Thus, we intend to generalize our theory to the sparse graph setting by replacing graphons, which inherently generate dense graphs, with suitable nonparametric sparse graph models, e.g., *graphexes*.

V. Acknowledgments

This research was partially supported by grants from ARO W911NF-19-1026, ARO W911NF-15-1-0479, and ARO W911NF-14-1-0359 and the Blue Sky Initiative from the College of Engineering at the University of Michigan.

References

[1] M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, and F. Wang, "Graph convolutional networks for computational drug development and discovery," *Briefings in bioinformatics*, 2019.

[2] M. Randić and S. C. Basak, "A comparative study of proteomics maps using graph theoretical biodescriptors," *Journal of chemical information and computer sciences*, vol. 42, no. 5, pp. 983–992, 2002.

[3] O. Sporns, "Graph theory methods for the analysis of neural connectivity patterns," in *Neuroscience databases*. Springer, 2003, pp. 171–185.

[4] J. A. Barnes and F. Harary, "Graph theory in network analysis," 1983.

[5] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *IEEE Data Eng. Bull.*, vol. 40, pp. 52–74, 2017.

[6] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[7] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. USA: Curran Associates Inc., 2016, pp. 3844–3852. [Online]. Available: http://dl.acm.org/citation.cfm?id=3157382.3157527

[8] L. Lovász, *Large Networks and Graph Limits.*, ser. Colloquium Publications. American Mathematical Society, 2012, vol. 60.

[9] D. J. Aldous, "Representations for partially exchangeable arrays of random variables," *Journal of Multivariate Analysis*, vol. 11, no. 4, pp. 581 – 598, 1981. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0047259X81900993

[10] L. Ruiz, F. Gama, and A. Ribeiro, "Gated graph convolutional recurrent neural networks," *arXiv preprint arXiv:1903.01888*, 2019.

[11] T. S. Jepsen, C. S. Jensen, and T. D. Nielsen, "Graph convolutional networks for road networks," *arXiv preprint arXiv:1908.11567*, 2019.

[12] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen, "A graph-convolutional neural network model for the prediction of chemical reactivity," *Chemical science*, vol. 10, no. 2, pp. 370–377, 2019.

[13] W. Yao, A. S. Bandeira, and S. Villar, "Experimental performance of graph neural networks on random instances of max-cut," in *Wavelets and Sparsity XVIII*, vol. 11138. International Society for Optics and Photonics, 2019, p. 111380S.

[14] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, 2015, pp. 2224–2232.

[15] S. Verma and Z.-L. Zhang, "Stability and generalization of graph convolutional neural networks," *arXiv preprint arXiv:1905.01004*, 2019.

[16] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, "Weisfeiler and lehman go neural: Higher-order graph neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4602–4609.

[17] Z. Chen, S. Villar, L. Chen, and J. Bruna, "On the equivalence between graph isomorphism testing and function approximation with gnns," *arXiv preprint arXiv:1905.12560*, 2019.

[18] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint arXiv:1810.00826*, 2018.

[19] B. Y. 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*, vol. 2, no. 9, pp. 12–16, 1968.

[20] S. Arora and A. Risteski, "Provable benefits of representation learning," *CoRR*, vol. abs/1706.04601, 2017.

[21] T. Kawamoto, M. Tsubaki, and T. Obuchi, "Mean-field theory of graph neural networks in graph partitioning," in *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. USA: Curran Associates Inc., 2018, pp. 4366–4376. [Online]. Available: http://dl.acm.org/citation.cfm?id=3327345.3327349

[22] L. Lovász and B. Szegedy, "Limits of dense graph sequences," *Journal of Combinatorial Theory, Series B*, vol. 96, no. 6, pp. 933 – 957, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0095895606000517

[23] C. Borgs, J. Chayes, L. Lovász, V. Sós, and K. Vesztergombi, "Convergent sequences of dense graphs i: Subgraph frequencies, metric properties and testing," *Advances in Mathematics*, vol. 219, no. 6, pp. 1801 – 1851, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001870808002053

[24] C. Borgs, J. Chayes, L. Lovász, V. Sós, and K. Vesztergombi, "Convergent sequences of dense graphs. ii. multiway cuts and statistical physics," *Annals of Mathematics. Second Series*, vol. 1, 07 2012.

[25] S. H. Chan and E. M. Airoldi, "A consistent histogram estimator for exchangeable graph models," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. I–208–I–216. [Online]. Available: http://dl.acm.org/citation.cfm?id=3044805.3044830

[26] C. Gao, Y. Lu, and H. H. Zhou, "Rate-optimal graphon estimation," *Ann. Statist.*, vol. 43, no. 6, pp. 2624–2652, 12 2015. [Online]. Available: https://doi.org/10.1214/15-AOS1354

[27] O. Klopp and N. Verzelen, "Optimal graphon estimation in cut distance," *Probability Theory and Related Fields*, vol. 174, no. 3, pp. 1033–1090, Aug 2019. [Online]. Available: https://doi.org/10.1007/s00440-018-0878-1

[28] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*. American Mathematical Society, 2006.

[29] A. Magner, M. Baranwal, and A. O. Hero III, "Fundamental limits of deep graph convolutional networks," *arXiv preprint arXiv:1910.12954*, 2020.

[30] F. Wu, A. H. Souza, T. Zhang, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," in *ICML*, 2019.
[31] S. Janson, "Graphons, cut norm and distance, couplings, and rearrangements," *New York Journal of Mathematics*, vol. 4, pp. 1–76, 2013.
[32] C. Sakr, Y. Kim, and N. Shanbhag, "Analytical guarantees on numerical precision of deep neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 3007–3016. [Online]. Available: http://proceedings.mlr.press/v70/sakr17a.html
[33] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," *ArXiv*, vol. abs/1606.08415, 2017.
[34] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in neural information processing systems*, 2017, pp. 971–980.