The Power of Graph Convolutional Networks to Distinguish Random Graph Models

Abram Magner, Mayank Baranwal and Alfred O. Hero III

Abstract

Graph convolutional networks (GCNs) are a widely used method for graph representation learning. We investigate the power of GCNs, as a function of their number of layers, to distinguish between different random graph models 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 exhibit an infinite class of graphons that are well-separated in terms of cut distance and 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, and furthermore show that, for this application, ReLU activation functions and non-identity weight matrices with non-negative entries do not help in terms of distinguishing power. These results theoretically match empirical observations of several prior works. Finally, we show that for pairs of graphons satisfying a degree profile separation property, a very simple GCN architecture suffices for distinguishability. 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

Abram Magner (email: amagner@albany.edu) is with the Department of Computer Science, University at Albany, SUNY, Albany, NY, USA. Mayank Baranwal (email: mayankb@umich.edu) and Alfred Hero (email: hero@eecs.umich.edu) are with the Department of EECS, University of Michigan, Ann Arbor, MI, USA.

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.
of graph representation learning schemes have been proposed and used [5]. One particularly effective such method is the \textit{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 \textit{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. 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 \textit{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.

\section*{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 of graph limit theory, 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 connection between GCNs and random walks on graphs. We then give lower bounds 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 graphons, arising from stochastic block models, that are hard for any test to distinguish on the basis of these embeddings. We first establish our result for the case of identity weight matrices and identity initial embedding matrix using the random walk connection in Theorems [1] and [2]. We then generalize these results to a broad class of choices of weight matrices (essentially, all sequences of weight matrices whose product has bounded norm) in Theorem [3]. We further generalize beyond stochastic block models to a larger class of pairs of graphons that are hard to distinguish in Theorem [5]. In the course of the proofs of these results, we illustrate the limitations of a linear GCN architecture.
Finally, we show an achievability result in Theorem 4 that says, roughly, that provided that the number of layers is sufficiently large ($K = \Omega(\log n)$), there exists 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 nonlinearity is necessary.

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 the content of Theorems [1] [2] [3] and [5]. On the other hand, Theorem 4 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 do not arise frequently 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.

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 $\hat{A}$, and starting with an initial embedding matrix $\hat{M}^{(0)}$, the $\ell$th embedding matrix is defined as follows:

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

(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 $\hat{M}^{(\ell)}$:

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

(2)

Typical examples of activation functions in neural network and GCN contexts include

- Rectified linear unit (ReLU): $\sigma(x) = x \cdot I[x > 0]$.
- Sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$.
- Hyperbolic tangent: $\sigma(x) = \tanh(x)$.

Empirical work has given evidence that the performance of GCNs on certain classification tasks is unaffected by replacing nonlinear activation functions by the identity [29]. Frequently, $\hat{A}$ is replaced by either the normalized adjacency matrix $D^{-1/2}A D^{-1/2}$, 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 generally write this model as \( \text{SBM}(p_1, p_2, q) \), suppressing \( k_1 \).

An important metric on graphons is the cut distance [30]. It is induced by the cut norm, which is defined as follows: fix a graphon \( W \). Then

\[
\|W\|_{\text{cut}} = \sup_{S,T} \left| \int_{S \times T} W(x, y) \, d\mu(x) \, d\mu(y) \right|,
\]

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 \( d_{\text{cut}}(W_0, W_1) \) between graphons \( W_0, W_1 \) is then defined as

\[
d_{\text{cut}}(W_0, W_1) = \inf_{\phi} \|W_0 - W_1(\phi(\cdot), \phi(\cdot))\|_{\text{cut}},
\]

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 \( B \sim \text{Bernoulli}(1/2) \) 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 \([-\epsilon_{\text{res}}, \epsilon_{\text{res}}]\), for a parameter \( \epsilon_{\text{res}} > 0 \) 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 [31]. 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 \( d = n \).
III. MAIN RESULTS

Our first theorems establish a convergence result and a subsequent error probability lower bound in the simplest case where all weight matrices, as well as the initial embedding matrix $\hat{M}^{(0)}$, are equal to the $n \times n$ identity matrix, and the activation function $\sigma$ is simply the identity. Note that, for the given choice of initial embedding and weight matrices, this is equivalent to using ReLU activations. In this setting, the $K$th embedding matrix is given by $\hat{A}^K$, where $\hat{A}$ is the random walk matrix of the input graph, and is therefore exactly the $K$-step random walk matrix. We will then generalize this result to a larger class of weight matrices and to a much larger class of pairs of graphons.

Theorem 1 (Convergence of powers of random walk matrices). There exists an infinite family $\mathcal{F}$ of graphons with the following properties:

- They are well-separated in terms of cut distance, in the sense that there is some $\delta > 0$ such that infinitely many pairs $W_0, W_1 \in \mathcal{F}$ satisfy $d_{cut}(W_0, W_1) \geq \delta$.
- Let $K \geq D \log n$ for a sufficiently large constant $D > 0$. For any two graphons $W_0, W_1$ in $\mathcal{F}$, the scaled output embedding matrices $n\hat{M}^{(K,0)}, n\hat{M}^{(K,1)}$ of a $K$-layer GCN with identity weight matrices and initial embedding matrix, with input graphs sampled from $W_0$ and $W_1$, converge to the same point in the $L_\infty$ norm.

The above theorem may be interpreted in a negative light in relation to GCNs as representation learning methods for the hypothesis testing problem that we consider. In particular, a useful representation maps pairs of well-separated points in the input graph space to pairs of well-separated points in the latent representation space.

Theorem 2 (Probability of error lower bound for distinguishing graphons using identity weight matrices). Let $\mathcal{F}$ be the family of graphons in Theorem\[.] For any two graphons $W_0, W_1 \in \mathcal{F}$, let $K$ be as before. Then there exist two sequences $\{G^{(b)n}\}_{n=1}^{\infty}$, $\{G^{(1)n}\}_{n=1}^{\infty}$ of random graph models such that

- with probability 1, samples $G^{(b)n} \sim G^{(b)n}$ converge in cut distance to $W_b$,
- Let $F(G)$ denote the embedding vector output of the GCN with $K$ layers, all weight matrices and initial embedding matrix equal to the $n \times n$ identity matrix. Then the probability of error of any test in distinguishing between $W_0$ and $W_1$ based on $H^{(K)}$, the $\epsilon_{res}$ uniform
perturbation of $F(G_{B,n})$, is at least
\[ \exp \left( -\frac{C}{\epsilon_{\text{res}} n} \right) (1 + o(1)), \] (5)
for some fixed positive constant $C$. If $\epsilon_{\text{res}} = \Omega(1/n)$, this is $\Omega(1)$.

The above results hold for a much larger class of parameters of the GCN, as stated in the next theorems, with some caveats. Recalling the empirical fact that nonlinear activation functions do not seem to improve the performance of GCNs [29], we consider below a GCN model in which they are replaced by the identity function. Whenever the weight matrices have only non-negative entries, this is equivalent to the case where the activation function is chosen to be the ReLU. We may alternatively view the following result as illustrating the limitations of linear GCNs and the necessity of adding nonlinearity.

**Theorem 3** (Convergence and probability of error lower bound for a broad class of weight matrices). Consider again the family $F$ in Theorems 1 and 2, and let $K$ be as before. Provided that there is a positive constant $C > 0$ for which the norm of the product of the weight matrices and the initial embedding matrix $\hat{M}^{(0)}$ satisfies
\[ \| \hat{M}^{(0)} K^{-1} \prod_{j=0}^{K-1} W^{(j)} \|_\infty \leq C, \] (6)
and provided that the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is the identity, the conclusions in Theorems 1 and 2 remain true.

The above results theoretically reflect a phenomenon that has been observed empirically in GCNs [6]: namely, the performance of a GCN for various classification tasks tends to degrade if the number of layers used is excessively large. In practice, this degradation of performance can occur for $K$ as small as 5. We note that for typical graphs encountered in GCN applications, the number of vertices $n$ is small enough that $\log n$ is not very large, so that $K = \Omega(\log n)$ restricts $K$ to a small range.

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

**Theorem 4** (Distinguishability result). Let $W_0, W_1$ denote two graphons with the following properties:

- There is some fixed $\ell \in (0, 1]$ such that $W_0(x, y), W_1(x, y) \geq \ell$ for every $x, y \in [0, 1]$. 

The degree profiles (equivalent to the random walk stationary distributions) of sample graphs from \( W_0 \) and \( W_1 \) are sufficiently different. More formally, define the degree function \( d_W(x) \) to be

\[
d_W(x) = \int_0^1 W(x, y) \, dy,
\]

and define the total degree function

\[
D(W) = \int_0^1 \int_0^1 W(x, y) \, dx \, dy.
\]

Then we stipulate that

\[
\int_0^1 \left| \frac{d_{W_0}(\phi(x))}{D(W_0)} - \frac{d_{W_1}(x)}{D(W_1)} \right| \, dx > \delta,
\]

for some fixed \( \delta > 0 \) and for all measure-preserving bijections \( \phi \).

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 \( \epsilon_{res} \)-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 \( \epsilon_{res} \leq \frac{\delta}{2n} \).

We may generalize Theorems 1, 2, and 3 in the following ways: we may consider a broader class of pairs of graphons, and we may consider sufficiently well-behaved nonlinear activation functions. The following theorem generalizes in both directions, essentially providing a converse to Theorem 4.

Before we state it, we define the following class of activation functions:

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

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

For simplicity, 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 [32]: \( \sigma(x) = \frac{x}{1+e^{-x}} \)
- SELU [33]:

\[
\sigma(x) = \begin{cases} 
  e^x - 1 & x \leq 0 \\
  x & x > 0 
\end{cases}
\] (10)

**Theorem 5** (Probability of error lower bound for a broader class of graphons and for nonlinear activations). Let \( W_0, W_1 \) denote two graphons with the following properties:

- There is some fixed \( \ell \in (0, 1] \) such that \( W_0(x, y), W_1(x, y) \geq \ell \) for every \( x, y \in [0, 1] \).
- The degree profiles (equivalent to the random walk stationary distributions) of sample graphs from \( W_0 \) and \( W_1 \) are sufficiently similar. More formally, define the degree function \( d_W(x) \) to be

\[
d_W(x) = \int_0^1 W(x, y) \, dy,
\] (11)

and define the total degree function

\[
D(W) = \int_0^1 \int_0^1 W(x, y) \, dx \, dy.
\] (12)

Then we stipulate that

\[
\int_0^1 \left| \frac{d_{W_0}(\phi(x))}{D(W_0)} - \frac{d_{W_1}(x)}{D(W_1)} \right| \, dx \leq \delta,
\] (13)

for some fixed \( \delta > 0 \) and for some measure-preserving bijection \( \phi \).

Suppose that \( \epsilon_{res} > \frac{\delta}{2n} \). Then there exist two sequences \( \{G_{0,n}\}_{n=1}^{\infty}, \{G_{1,n}\}_{n=1}^{\infty} \) of random graph models such that

- with probability 1, samples \( G_{b,n} \sim G_{b,n} \) converge in cut distance to \( W_b \),
- The conclusions of Theorem 2 and 3 hold with a GCN architecture that uses as an activation function some \( \sigma \in \mathcal{A} \).

We note that the class of stochastic block models parameterized by \( \mathcal{P} \) yields an infinite set of pairs of graphons satisfying the conditions in Theorem 5.
A. 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 information-theoretically 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.

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, captured by graphexes.

IV. PROOFS

A. Proof of Theorem 1

For our convergence result, we will consider the following family of graph models: the stochastic block models $\text{SBM}(k_1, k_2, p_1, p_2, q)$ with two blocks of size $k_1 n$ and $k_2 n$ and with three density parameters: $p_1, p_2, q$, where $p_j$ denotes the probability of an edge between two vertices in class $j$, $j \in \{1, 2\}$, and $q$ denotes the probability of an edge between two vertices in different classes. We will fix $k_1$ and $k_2$ throughout (say, $k_1 = k_2 = 1/2$) and, therefore, suppress
them in our notation. 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

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

(14)

where $\preceq$ in the lexicographic partial order.

We have chosen this parameter set because of the following fact.

**Lemma 1** (Equal expected degree sequence for every parameter tuple in $\mathcal{P}$). Let $P \in \mathcal{P}$. Then for any vertex $v$ in class 1 of a sample graph $G \sim \text{SBM}(P)$, we have

$$\mathbb{E}[\deg(v)] = k_1 n p_{*1} + k_2 n q_*, \quad (15)$$

and for any vertex $w$ in class 2 of $G$,

$$\mathbb{E}[\deg(w)] = k_2 n p_{*2} + k_1 n q_*.$$  

(16)

In other words, every $\text{SBM}(P)$ with $P \in \mathcal{P}$ has the same set of expected vertex degrees.

**Proof.** This is an easy expected degree calculation. The particular choice of the vector $(1/k_1, k_1/k_2^2, -1/k_2)$ in the definition of $\mathcal{P}$ leads to cancellation of perturbations. $\Box$

Intuitively, the fact that all models parameterized by $\mathcal{P}$ have the same set of expected degrees is important for the following reason: under the hypotheses of Theorem 2, a $K$-layer GCN outputs a $K$-step random walk matrix for any input graph $G$. Moreover, provided that the random walk Markov chain on $G$ is ergodic (which holds with high probability for samples from $\text{SBM}(P)$, $P \in \mathcal{P}$), it has a unique stationary distribution $\pi$ such that, for each vertex $v \in G$,

$$\pi_v = \frac{\deg(v)}{\sum_{w \in G} \deg(w)}. \quad (17)$$

Thus, we expect that large graphs with approximately equal degree sequences will result in embeddings that are close to one another, provided that $K$ is sufficiently large compared to the mixing time of the random walk on $G$ and provided that $\epsilon_{res}$ is bounded away from 0.

Additionally, the graphons corresponding to $\mathcal{P}$ have the following separation property.

**Lemma 2** (Cut distance property of $\mathcal{P}$). There is some $\delta > 0$ such that there exist infinitely many pairs $P_0, P_1 \in \mathcal{P}$ such that the corresponding graphons $W_0, W_1$, respectively, satisfy

$$d_{cut}(W_0, W_1) \geq \delta.$$  

(18)
Proof. This is a simple exercise in following the definitions and is omitted for brevity. \hfill \Box

The rest of the proof proceeds as follows: we show that for any pair of graphons $W_0, W_1$ corresponding to different parameterizations $P_0, P_1 \in \mathcal{P}$, with high probability, a sample graph from either one will have an embedding matrix that converges to the same point. We do this in a sequence of steps: we first establish the limit point. We then upper bound the distance to the limit point in terms of the mixing time of the random walk chain on $G$. Finally, we upper bound the mixing time of the random walk on $G$ in terms of the density parameters.

**Lemma 3** (Limit point of the embedding matrices). Consider the $n \times n$ matrix $M_\infty$ defined as follows: the rows of $M_\infty$ are all copies of the row vector $\pi$, where

$$
\pi_v = \frac{\mathbb{E}[\text{deg}(v)]}{\sum_w \mathbb{E}[\text{deg}(w)]}
$$

for each $v \in [n]$. More explicitly, if $v$ is in class 1, then

$$
\pi_v = \frac{k_1 p_1 + k_2 q}{n(k_1^2 p + 2k_1 k_2 q + k_2^2 p_2)};
$$

and if $v$ is in class 2, then

$$
\pi_v = \frac{k_2 p_2 + k_1 q}{n(k_1^2 p + 2k_1 k_2 q + k_2^2 p_2)}.
$$

We refer to these functions as $\rho_1(p_1, p_2, q)$ and $\rho_2(p_1, p_2, q)$, respectively.

We have that, for either $B = 0$ or 1,

$$
\mathbb{P}[\|\hat{A}_\infty^{(B)} - M_\infty\|_\infty > \epsilon(n)] \leq 2n e^{-\mu_* \delta^2/3},
$$

where $\mu_* \triangleq \min_{v \in [n]} \mathbb{E}[\text{deg}(v)]$, $\delta \in (0, 1)$ such that $\delta \gg 1/\sqrt{\mu_*}$ and $\delta = o(1)$, $\epsilon(n) = \Theta(\delta/n)$, and $\hat{A}_\infty^{(B)}$ is the stationary matrix of the random walk matrix $\hat{A}^{(B)}$.

**Proof.** Recall that the stationary probability of any vertex for the random walk on a given graph is given by

$$
\frac{\text{deg}(v)}{\sum_w \text{deg}(w)}.
$$

Thus, if we can show that the degrees of all vertices are concentrated around the same expected values for both values of $B$, then we can conclude that the sample embedding matrices converge to the same value with high probability.

We note that the degree of each vertex $v$ (say, without loss of generality, in class 1) is a sum of two independent binomial random variables, giving the number of connections to other vertices...
in the same class and to vertices in the other class. Thus, we can use the standard Chernoff bound to prove concentration of the degrees. Define $\mu_v = \mathbb{E}[\deg(v)]$. We have, for any vertex $v$, and for any $\delta \in (0, 1)$,

$$\mathbb{P}[|\deg(v) - \mathbb{E}[\deg(v)]| > \delta \mu_v] \leq 2e^{-\mu_v \delta^2/3}, \quad (24)$$

which tends to 0 exponentially fast as long as $\delta \gg 1/\sqrt{\mu_v}$.

Define $\mu^*$ to be $\min_{v \in [n]} \mu_v$. By union bounding over all vertices, we see that

$$\mathbb{P}[\bigcup_v |\deg(v) - \mathbb{E}[\deg(v)]| > \delta \mu_v] \leq 2n e^{-\mu^* \delta^2/3}. \quad (25)$$

Let $A$ denote the event that every vertex $v$ is within $\delta \mu_v$ of its expectation. When $A$ holds, we have

$$\sum_w \deg(w) \in \left[\sum_w \mu_w - \delta, \sum_w \mu_w + \delta \right]. \quad (26)$$

For brevity, we write $X = \sum_w \mu_w$. This implies the following, for every $v$:

$$\frac{\deg(v)}{\sum_w \deg(w)} = \frac{\deg(v)}{X(1 + \delta)} - \frac{\mathbb{E}[\deg(v)]}{X} = \frac{1}{X} \left(\frac{\deg(v)}{1 + \delta} - \mathbb{E}[\deg(v)]\right). \quad (27)$$

Upper bounding $\deg(v)$ by $(1 + \delta)\mu_v$, we have

$$\frac{1}{X} \left(\frac{\deg(v)}{1 + \delta} - \mathbb{E}[\deg(v)]\right) \leq \frac{1}{X} \left(\frac{1 + \delta}{1 - \delta} - \frac{\mu_v}{X} \left(\frac{1 + \delta}{1 - \delta} - 1\right) \right) \leq 3\delta \pi_v, \quad (28)$$

provided that $\delta = o(1)$.

This implies that if we set $\epsilon(n) = 3\delta \max_v \{\pi_v\} = \Theta(\delta/n)$, then

$$\mathbb{P}[\bigcup_v \left|\frac{\deg(v)}{\sum_w \deg(w)} - \pi_v\right| > \epsilon(n)] \leq 2ne^{-\mu^* \delta^2/3}. \quad (29)$$

This completes the proof. \qed

**Lemma 4** (Distance to limit of random walk matrix powers in terms of mixing times). Consider a Markov chain with transition matrix $P$ and stationary matrix $P_\infty$. Let $t_{\text{mix}}(P, \epsilon)$ denote the $\epsilon$-total variation mixing time of $P$. For any $t \geq t_{\text{mix}}(P, \epsilon)$, we have that

$$\|P^t - P_\infty\|_\infty \leq 2\epsilon. \quad (30)$$

**Proof.** By definition of $t_{\text{mix}}(P, \epsilon)$, whenever $t \geq t_{\text{mix}}(P, \epsilon)$, for any initial distribution $\mu_0$ over nodes (a row vector),

$$\frac{1}{2}\|\mu_0 P^t - \pi\|_1 \leq \epsilon. \quad (31)$$
We therefore choose \( \mu_0 = e_j^T \) (i.e., the transpose of the \( j \)th standard basis vector. We note that \( \mu_0 P^t \) is the \( j \)th row of \( P^t \), and the above implies that every element of the \( j \)th row of \( P^t \) is within \( 2\epsilon \) of the corresponding element of \( \pi \). This completes the proof.

\[ \square \]

**Lemma 5** (Mixing times in terms of model parameters). Fix a parameter tuple \( \Psi = (p_1, p_2, q) \in \mathcal{P} \). Let \( G \) be a sample graph from this model. Then there is some positive constant \( C \) such that the mixing time of the simple random walk on \( G \) satisfies

\[
\mathbb{P}[t_{mix}(G, \epsilon) \geq C \log \frac{n}{\epsilon}] \leq e^{-\Theta(n)}.
\]

(32)

**Proof.** We prove this by lower bounding the expansion (also called the bottleneck ratio) of \( G \). Through Cheeger’s inequality, this translates to a lower bound on the spectral gap \( \gamma_* \) of the random walk matrix \( P \) (equivalently, an upper bound on the relaxation time \( t_{rel} = 1/\gamma_* \)), which in turn directly upper bounds the mixing time in terms of the minimum stationary probability.

Recall that the bottleneck ratio of a Markov chain with transition matrix \( P \) with stationary distribution \( \pi \) is defined as follows:

\[
\Phi(P) = \min_{S : \pi(S) \leq 1/2} \frac{Q(S, S^c)}{\pi(S)},
\]

(33)

where \( S \) ranges over subsets of the state space, and \( Q(S, S^c) \) is the probability that, if we start from the stationary distribution and take a single step, we move from \( S \) to \( S^c \). This may be rewritten in the case of random walks on graphs as

\[
\Phi(P) = \min_{S : \pi(S) \leq 1/2} \frac{|\partial S|}{\sum_{x \in S} \deg(x)},
\]

(34)

where \( \partial S \) denotes the boundary of \( S \), which is the set of edges connecting nodes in \( S \) and \( S^c \). Our goal will be to show that with high probability, \( \Phi(P) > C \), for some positive constant \( C \) depending on \( \Psi \). This yields the following sequence of implications:

\[ C < \Phi(P) \implies C^2/2 < \gamma_* \implies t_{rel} \leq 2/C^2. \]

(35)

Here, the first implication is from Cheeger’s inequality. Next, this implies that

\[
t_{mix}(P, \epsilon) \leq 2/C^2 \log(\frac{1}{\epsilon \pi_{min}}) = \Theta(\log n),
\]

(36)

where we have used the fact that \( \pi_{min} \), the minimum stationary probability achieved by any vertex, satisfies \( \pi_{min} = O(1/n) \).
We now proceed to lower bound $\Phi(P)$. A simple upper bound on the denominator is as follows:

$$\sum_{x \in S} \deg(x) \leq |S|n. \quad (37)$$

Meanwhile, to lower bound the numerator, we reason as follows. Let $S$ be an arbitrary such subset of nodes with $s_1, s_2$ nodes in classes 1 and 2, respectively, and let $S_1, S_2, S'_1, S'_2$ denote the subsets of $S$ in classes 1 and 2 and the subsets of $S^c$ in classes 1 and 2, respectively. Then

$$\mathbb{E}[|\partial S|] = p_1s_1s'_1 + p_2s_2s'_2 + qs_1s'_2 + qs'_1s_2, \quad (38)$$

by linearity of expectation. Now, $|S^c|$ must be $\Theta(n)$, since $\pi(S) \leq 1/2$ and the minimum stationary probability of any vertex is, with high probability, $\Theta(1/n)$. This implies that at least one term in (38) is $\Theta(|S|n)$. An application of the Chernoff bound allows us to conclude that $|\partial S| = \Theta(|S|n)$ with probability at least $1 - e^{-\Theta(n)}$. Putting this together with (37) yields the desired lower bound on $\Phi(P)$. This completes the proof.

We are now ready to complete the proof of Theorem 1. Fix $\epsilon > 0$. By Lemma 5 provided that $K \geq D \log n$ with the constant $D$ large enough, the mixing time of $G \sim W_B$, with probability exponentially close to 1, satisfies

$$t_{mix}(\tilde{M}(0), \epsilon) \leq K. \quad (39)$$

By Lemmas 3 and 4, this implies that with probability $1 - e^{-\Theta(n)}$, the (unperturbed) embedding matrix $\tilde{M}^{(K)}$ satisfies

$$\|\tilde{M}^{(K)} - M_\infty\|_\infty \leq \epsilon_2(n) = o(n). \quad (40)$$

Note, in particular, that $M_\infty$ has the same value for both $W_0$ and $W_1$, by construction of $\mathcal{P}$.

This, in turn, implies that for any coupling of random graphs $G^{(0)} \sim W_0$ and $G^{(1)} \sim W_1$ with embedding matrices $\tilde{M}^{(K,0)}$ and $\tilde{M}^{(K,1)}$, respectively,

$$\|\tilde{M}^{(K,0)} - \tilde{M}^{(K,1)}\|_\infty \leq \|\tilde{M}^{(K,0)} - M_\infty\|_\infty + \|\tilde{M}^{(K,1)} - M_\infty\|_\infty \leq 2\epsilon_2(n) = o(n), \quad (41)$$

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


B. Proof of Theorem 2

In order to prove the probability of error lower bound, we will need to define our models $G_{n,0}, G_{n,1}$ in such a way that they may be coupled, with the property that almost all vertex degrees be approximately equal in samples from both models. We describe our construction as follows.

- Generate $G'_{n,0} \sim \text{SBM}(P_0)$, $G'_{n,1} \sim \text{SBM}(P_1)$ independently. Set $G_{n,1} = G'_{n,1}$.
- Fix some $C > 1$. In any graph $G$, we say that a vertex $v$ is $C$-small if its degree satisfies
  \[ \deg_G(v) < \deg_{G'_{n,1}}(v) - C. \]  
  Similarly, we say that it is $C$-large if
  \[ \deg_G(v) > \deg_{G'_{n,1}}(v) + C. \]  
All other vertices will be said to be $C$-just right. Now, we will repair $G'_{n,0}$ so that its degrees are very close to those of $G_{n,1}$. In $G'_{n,0}$, we sequentially add in an arbitrary order all possible edges connecting $C$-small vertices, until no more such edges may be added. Note that in the course of doing this, some vertices may cease to be $C$-small. We then analogously remove edges between $C$-large vertices. This results in a graph which we call $G_{n,0}$.
- Finally, we are left with a set of $C$-small and a set of $C$-large vertices in $G_{n,0}$, as well as a set of $C$-just right vertices. We will show in Lemma 7 below that only an asymptotically small number of vertices are $C$-small or $C$-large. It will additionally be important that at most $O(n^{1/2+\delta})$ edges per vertex have been modified in transforming $G'_{n,0}$ into $G_{n,0}$.

Regarding the above construction, we have the following facts.

**Lemma 6.** The graphs $G_{n,0}$ and $G_{n,1}$ converge in cut distance to $W_0$ and $W_1$, respectively, with probability 1.

**Proof.** This follows directly from the cut distance convergence of the stochastic block models to $W_0$ and $W_1$, along with the fact that we did not add or remove more than $O(n^{1/2+\delta})$ edges incident on any given vertex. In particular, this means that the density (normalized by $n^2$) of edges between any pair of subsets of vertices was not perturbed by more than $O(n^{-1/2-\delta})$. \[ \square \]

**Lemma 7** (Number of large, small, and just right vertices in $G_{n,0}$). With high probability, the numbers of large, small, and $C$-just right vertices in $G_{n,0}$ are $O(n^{1/2+\delta}), O(n^{1/2+\delta}),$ and
\(n(1 - O(n^{-1/2+\delta}))\), respectively, where \(\delta > 0\) is arbitrarily small. Moreover, all of the large and small vertices are \(O(n^{1/2+\delta})\)-just right.

**Proof.** We note that after the edge addition process stops, the set of small vertices in \(G_{n,0}\) forms a clique. If this were not so, then we could continue by adding at least one edge between small vertices. Furthermore, since, with probability exponentially close to 1 as a function of \(n\), the discrepancies between the degrees of corresponding vertices in \(G'_{n,0}\) and \(G'_{n,1}\) are all at most \(O(n^{1/2+\delta})\), for any fixed \(\delta > 0\), we know that we added at most \(O(n^{1/2+\delta})\) edges per small vertex in \(G'_{n,0}\) to get \(G_{n,0}\). In \(G'_{n,0}\), with high probability, the size of the largest subset of vertices such that adding \(O(n^{1/2+\delta})\) edges per vertex yields a clique is \(O(n^{1/2+\delta})\). This implies that the set of small vertices in \(G_{n,0}\) has cardinality \(O(n^{1/2+\delta})\). The same holds for the set of large vertices. Finally, this implies that the number of just right vertices is \(n(1 - O(n^{-1/2+\delta}))\).

Lemma 7 will be important in establishing our error probability lower bound.

Le Cam’s method is the tool of choice for lower bounding the error probability of any test that distinguishes between \(W_0\) and \(W_1\) from a sample graph \(G_{n,B}\). In particular, Le Cam’s method requires us to upper bound the following quantity:

\[
d_{TV}(H^{(K,0)}, H^{(K,1)}),
\]

the total variation distance between the laws of the perturbed output embedding vectors under the models corresponding to \(W_0\) and \(W_1\).

To upper bound the quantity in (44), we will need the following lemma, which gives an expression for the total variation distance between the random perturbations of two fixed matrices in terms of their \(L_\infty\) distance.

**Lemma 8** (Exact expression for \(d_{TV}\) of perturbed matrices). Suppose that \(M^{(0)}\) and \(M^{(1)}\) are \(m \times n\) matrices, for arbitrary positive integers \(m, n\). Suppose, further, that \(\tilde{M}^{(0)}\) and \(\tilde{M}^{(1)}\) are independent element-wise \(\epsilon_{res}\) perturbations of \(M^{(0)}\) and \(M^{(1)}\). Then

\[
d_{TV}(\tilde{M}^{(0)}, \tilde{M}^{(1)}) = \frac{\text{Vol}(\mathbb{H}(M^{(0)}, \epsilon_{res}) \triangle \mathbb{H}(M^{(1)}, \epsilon_{res}))}{2\text{Vol}(\mathbb{H}(M^{(1)}, \epsilon_{res}))} = \frac{\text{Vol}(\mathbb{H}(M^{(0)}, \epsilon_{res}) \triangle \mathbb{H}(M^{(1)}, \epsilon_{res}))}{2\text{Vol}(\mathbb{H}(M^{(0)}, \epsilon_{res}))},
\]

(45)
where $\triangle$ denotes the symmetric difference between two sets, and $\mathbb{H}(M, r)$ denotes the axis-aligned hypercube of radius $r$ centered at $M$. This can be simplified as follows:

$$d_{TV}(\tilde{M}^{(0)}, \tilde{M}^{(1)}) = 1 - \prod_{i,j \in [m] \times [n]} \left( \frac{2\epsilon_{res} - |M_{i,j}^{(0)} - M_{i,j}^{(1)}|}{(2\epsilon_{res})^{mn}} \right)^+.$$  

(46)

**Proof.** By definition of total variation distance,

$$d_{TV}(\tilde{M}^{(0)}, \tilde{M}^{(1)}) = \frac{1}{2} \int_{x \in \mathbb{H}(M^{(0)}, \epsilon_{res}) \cup \mathbb{H}(M^{(1)}, \epsilon_{res})} \left| \frac{\mathbb{I}[x \in \mathbb{H}(M^{(0)}, \epsilon_{res})]}{\text{Vol}(\mathbb{H}(M^{(0)}, \epsilon_{res}))} - \frac{\mathbb{I}[x \in \mathbb{H}(M^{(1)}, \epsilon_{res})]}{\text{Vol}(\mathbb{H}(M^{(1)}, \epsilon_{res}))} \right| dx.$$  

(47)

and $\tilde{M}^{(1)}$, and making this substitution completes the first part of the proof. Now, we can provide a closed-form formula as follows.

The volume of the hypercube $\mathbb{H}(M^{(0)}, \epsilon_{res})$ is given by

$$\text{Vol}(\mathbb{H}(M^{(0)}, \epsilon_{res})) = (2\epsilon_{res})^{mn}. \quad (48)$$

To compute the volume of the symmetric difference between the two hypercubes, we first compute the volume of their intersection, then the volume of their union (which is the sum of their volumes, minus that of their intersection). Then the volume of the symmetric difference is the volume of the union minus that of the intersection.

The volume of the intersection of the two hypercubes can be computed by noting that it is an axis-aligned rectangle, where the length along the $(i, j)$ axis is given by

$$\ell_{i,j} = (2\epsilon_{res} - |M_{i,j}^{(0)} - M_{i,j}^{(1)}|)^+, \quad (49)$$

so that the volume of the intersection is

$$\text{Vol}(\mathbb{H}(M^{(0)}, \epsilon_{res}) \cap \mathbb{H}(M^{(1)}, \epsilon_{res})) = \prod_{i,j \in [m] \times [n]} \ell_{i,j}. \quad (50)$$

This implies

$$\text{Vol}(\mathbb{H}(M^{(0)}, \epsilon_{res}) \triangle \mathbb{H}(M^{(1)}, \epsilon_{res})) = 2 \cdot (2\epsilon_{res})^{mn} - 2 \prod_{i,j \in [m] \times [n]} \ell_{i,j}. \quad (51)$$

Finally, we get the following simplified expression for the total variation distance:

$$d_{TV}(\tilde{M}^{(0)}, \tilde{M}^{(1)}) = 1 - \prod_{i,j \in [m] \times [n]} \ell_{i,j} \left( \frac{(2\epsilon_{res})^{mn}}{(2\epsilon_{res})^{mn}} \right). \quad (52)$$

This completes the proof. \qed
Now, we use the expression in (46) to complete the proof of the theorem as follows. We set \( \tilde{M}^{(0)} \) and \( \tilde{M}^{(1)} \) in Lemma 8 to be \( \hat{H}_{j}^{(K,0)} \) and \( \hat{H}_{j}^{(K,1)} \), respectively. Thus, in the lemma, \( m = 1 \) and \( n \) is the number of vertices in the sample graphs. We need to upper bound the differences between \( \hat{H}_{j}^{(K,0)} \) and \( \hat{H}_{j}^{(K,1)} \). To do this, we note that the stationary distributions \( \hat{H}_{j}^{(\infty,0)} \) and \( \hat{H}_{j}^{(\infty,1)} \) of the random walks on \( G_{n,0} \) and \( G_{n,1} \), respectively, have the following structure, from Lemma 7: for \( O \left( \frac{n^{1/2+\delta}}{n^2} \right) \) indices \( j \in [n] \), which correspond to large and small vertices in \( G_{n,0} \), we have

\[
|\hat{H}_{j}^{(\infty,0)} - \hat{H}_{j}^{(\infty,1)}| = O\left( \frac{n^{1/2+\delta}}{n^2} \right) = O\left( \frac{n^{-3/2+\delta}}{n^2} \right).
\] (53)

The remaining \( n \left( 1 - O\left( n^{-1/2+\delta} \right) \right) \) indices \( j \) corresponding to \( C \)-just right vertices in \( G_{n,0} \) satisfy

\[
|\hat{H}_{j}^{(\infty,0)} - \hat{H}_{j}^{(\infty,1)}| = O\left( \frac{1}{n^2} \right).
\] (54)

Furthermore, from Lemma 4 with \( \epsilon = O\left( \frac{1}{n^2} \right) \), provided that we choose \( K > \max\{ t_{\text{mix}}(\hat{A}^{(0)}, \epsilon), t_{\text{mix}}(\hat{A}^{(1)}, \epsilon) \} \), we have that \( \hat{H}^{(K,0)} \) and \( \hat{H}^{(K,1)} \) have the same structure.

Finally, from Lemma 5 and noting that our construction of \( G_{n,0} \) did not substantially alter its mixing time, we see that \( K \) only needs to be \( D \log n \), for some large enough \( D \).

Applying Lemma 8, we see that the total variation distance is upper bounded by

\[
1 - \left( 1 - \frac{C}{2\epsilon_{\text{res}}n^2} \right)^n \cdot \left( 1 - O\left( \frac{n^{1/2+\delta}}{\epsilon_{\text{res}}n^2} \right) \right)^n = 1 - \left( 1 - \frac{C}{2\epsilon_{\text{res}}n^2} \right)^n \left( 1 - O\left( \frac{n^{-3/2+\delta}}{\epsilon_{\text{res}}} \right) \right)^n
\] (55)

\[
= 1 - \exp\left( -\frac{C}{2\epsilon_{\text{res}}n} \right) \left( 1 + o(1) \right),
\] (57)

which is strictly less than 1, as desired. We note that by a standard fact about the total variation distance, this translates to a lower bound of \( e^{-\frac{C}{2\epsilon_{\text{res}}n}} \left( 1 + o(1) \right) \) for the optimal error probability.

C. Proof of Theorem 3

This follows immediately by our previous results and the following: fixing \( W_0, W_1 \) as in the previous theorems and letting \( \hat{A}^{(b)} \) denote the random walk matrix arising from \( W_b \) and \( \tilde{M}^{(K,b)} \).
denote the unperturbed GCN output from \( \hat{A}^{(b)} \),

\[
\| \tilde{M}^{(K,0)} - \tilde{M}^{(K,1)} \|_{\infty} = \| (\hat{A}^{(0)} - \hat{A}^{(1)}) \cdot M^{(0)} \prod_{j=0}^{K-1} W^{(j)} \|_{\infty} \tag{58}
\]

\[
\leq \| \hat{A}^{(0)} - \hat{A}^{(1)} \|_{\infty} \cdot \| M^{(0)} \prod_{j=0}^{K-1} W^{(j)} \|_{\infty} \tag{59}
\]

\[
\leq \| \hat{A}^{(0)} - \hat{A}^{(1)} \|_{\infty} \cdot C. \tag{60}
\]

The previous theorems then show that the remaining norm in the above expression tends to 0.

D. Proof of Theorem 4

It suffices to show two things: that the stationary distributions of the random walks on the sample graphs from the two models are sufficiently well-separated, in the sense that their total variation distance under every reordering of vertices is asymptotically positive with high probability under every coupling, which will imply that their \( L_\infty \) distance is sufficiently large in comparison to the \( \epsilon_{res} \) perturbation; and that the \( \epsilon \)-mixing time of both random walks is at most \( O(\log(n/\epsilon)) \). If these two facts are granted, then we will have that the unperturbed output embedding vectors of samples from \( W_0 \) and \( W_1 \) will converge in probability to two points bounded away from each other by at least a constant in the \( L_\infty \) norm, which yields the claim.

Now, we prove the two facts, starting with the separation of the stationary distributions. Letting \( \pi_{G_0} \) and \( \pi_{G_1} \) denote stationary distributions of samples \( G_0 \sim W_0, G_1 \sim W_1 \), and letting the vertices of \( G_1 \) be reordered so as to minimize the total variation distance between \( \pi_{G_0} \) and \( \pi_{G_1} \), we have

\[
2d_{TV}(\pi_{G_0}, \pi_{G_1}) = \sum_{v \in [n]} |\pi_{G_0}(v) - \pi_{G_1}(v)| = \sum_{v \in [n]} \left| \frac{\deg_{G_0}(v)}{\sum_{w} \deg_{G_0}(w)} - \frac{\deg_{G_1}(v)}{\sum_{w} \deg_{G_1}(w)} \right|. \tag{61}
\]

Now, by the concentration properties of degrees in samples from graphons, we have that with probability \( 1 - o(1) \) as \( n \to \infty \),

\[
\frac{\sum_{w} \deg_{G_b}(w)}{n^2} = D(W_b)(1 + O(1/n)). \tag{62}
\]

For each vertex \( v \), let \( x_{b,v} \in [0,1] \) denote the uniformly randomly chosen coordinate corresponding to \( v \) in the process of generating \( G_b \). Then, again by the concentration properties of degrees, and conditioning on the values of the \( x_{b,v} \) for \( b \in \{0,1\} \) and \( v \in [n] \), we have that with high probability,

\[
\deg_{G_b}(v) = d_{W_b}(x_{b,v}) n(1 + O(1/\sqrt{n})). \tag{63}
\]
By applying a measure-preserving bijection, we may assume that \( x_{0,v} = x_{1,v} = x_v \). We thus have that

\[
\sum_{v \in [n]} \left| \frac{\deg_{G_0}(v)}{\sum_w \deg_{G_0}(w)} - \frac{\deg_{G_1}(v)}{\sum_w \deg_{G_1}(w)} \right| = \sum_{v \in [n]} \int_{x_v=0}^{1} \left| \frac{dW_0(x_v)}{nD(W_0)} - \frac{dW_1(x_v)}{nD(W_1)} \right| (1 + O(1/\sqrt{n})) \, dx_v.
\]

By our assumption, we then have that the inner integral is at least \( \delta/n(1 + O(1/\sqrt{n})) \), so that

\[
\sum_{v \in [n]} \int_{x_v=0}^{1} \left| \frac{dW_0(x_v)}{nD(W_0)} - \frac{dW_1(x_v)}{nD(W_1)} \right| (1 + O(1/\sqrt{n})) \, dx_v \geq \delta(1 + O(1/\sqrt{n})).
\]

This establishes that the stationary distributions of the random walks of the sample graphs of the two models are bounded away from each other in the total variation distance. Thus, they are also bounded away from each other when viewed as vectors in the \( L_\infty \) distance: in particular, we have that with high probability,

\[
\|\pi_{G_0} - \pi_{G_1}\|_\infty \geq \frac{\delta}{n} \cdot (1 + O(1/\sqrt{n})�).
\]

We next establish that the mixing times of the random walks on \( G_0 \) and \( G_1 \) are both upper bounded by \( O(\log(n/\epsilon)) \). We can do this by lower bounding the bottleneck ratio by a constant for both graphs, as in the proof of Lemma 5. Since both graphons \( W_0(x, y) \) and \( W_1(x, y) \) are uniformly lower bounded over all \( (x, y) \) by \( \ell \), it is simple to show that with probability 1, for any set \( S \) of vertices, \( \sum_{x \in S} \deg(x) \) is upper bounded by \( |S|/n \). Furthermore, \( |\partial S| \) is of the same order, for the same reasons as in the proof of Lemma 5. This implies the desired lower bound on the bottleneck ratio of either graph, and, hence, the upper bound of \( O(\log(n/\epsilon)) \) on the \( \epsilon \)-mixing time of both graphs, with probability \( 1 - e^{-\Theta(n)} \).

Having established the two necessary facts, the proof is complete, since we may set \( \epsilon \) to be \( 1/n^2 \). If the number of layers \( K \) is at least \( D \log n \geq t_{\text{mix}}(\hat{A}^{(b)}, \epsilon) \) for a large enough \( D \), then, just as in the proof of Lemma 4, with high probability,

\[
\|\hat{A}^{(b)K} - \hat{A}^{(b)\infty}\|_1 \leq 1/n^2 \quad \Rightarrow \quad \|\hat{A}^{(b)K} - \hat{A}^{(b)\infty}\|_\infty \leq 1/n^2.
\]

Now, since, as we showed above,

\[
\|1^T \hat{A}^{(0)\infty}/n - 1^T \hat{A}^{(1)\infty}/n\|_\infty \geq \delta/n,
\]

we have that

\[
\|1^T \hat{A}^{(0)K}/n - 1^T \hat{A}^{(1)K}/n\|_\infty \geq \delta/n + O(1/n^2).
\]

This implies that if \( \epsilon_{\text{res}} < \frac{\delta}{2n} \), with probability \( 1 - o(1) \), the two output embedding vectors can be distinguished. This completes the proof.
V. PROOF OF THEOREM 5

We first explain how to prove the theorem in the case of identity activation functions. We then give a reduction from the case of nonlinear activations to identity ones. We construct the distributions $G_{b,n}$ and their sample graphs $G_{b,n}$ using the same coupling construction as we do in the proof of Theorem 2, except that we generate $G'_{b,n}$ according to $W_b$ instead of a stochastic block model. This results in a pair of graphs $G_{b,n}, b \in \{0, 1\}$ with the same guarantees on their degrees as in that construction. The remainder of the proof requires upper bounding the $L_\infty$ distance between the stationary distribution vectors $\pi_{G_0}, \pi_{G_1}$ and showing that the mixing times of the random walks on both graphs are $O(\log n)$. The upper bound on the distance between stationary distributions follows easily by modifying the derivation leading to (65). In particular, the assumption in the current theorem statement implies that

$$\|\pi_{G_0} - \pi_{G_1}\|_\infty \leq \frac{\delta}{n}(1 + O(1/\sqrt{n})).$$

(70)

Furthermore, the mixing time upper bounds, and the combination of these and (70) to yield the claimed error probability lower bound, follow for exactly the same reasons as in the proof of Theorem 2. We note that the stipulated lower bound on $\epsilon_{res}$ is required to ensure that the distributions of the $\epsilon_{res}$-perturbed embedding vectors for the two input graphs have overlapping support. The generalization to non-identity weight matrices is exactly as in the proof of Theorem 3.

Finally, the following lemma reduces the case of nonlinear activations $\sigma \in A$ to $\sigma(x) = x$.

Lemma 9. Consider two random walk matrices $\hat{A}^{(0)}$ and $\hat{A}^{(1)}$.

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be in $A$.

Furthermore, let the sequence of weight matrices $W^{(0)}, ..., W^{(K)}$ satisfy the following:

$$\|\hat{M}^{(b,0)}\|_{op,\infty} \prod_{j=0}^{K} \|W^{(j)}\|_{op,\infty} \leq C,$$

(71)

for some fixed positive $C$ and for $b \in \{0, 1\}$, and

$$\sum_{j=0}^{K} \|W^{(j)}\|_{op,\infty} \leq D,$$

(72)

for some fixed positive $D$.

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

$$\|\hat{M}^{(0,K)} - \hat{M}^{(1,K)}\|_\infty \leq \left\|\hat{A}^{(0)K} \hat{M}^{(0,0)} \prod_{j=0}^{K} W^{(j)} - \hat{A}^{(1)K} \hat{M}^{(1,0)} \prod_{j=0}^{K} W^{(j)}\right\|_\infty (1 + o(1)).$$

(73)
\textbf{Proof.} We start by noting that we can approximate $$\sigma(x)$$ by its first-order Taylor expansion:
\begin{equation}
\sigma(x) = x + \sigma''(\xi)x^2/2 = x(1 + \sigma''(\xi)x/2),
\end{equation}
where $$\xi$$ is some real number between 0 and $$x$$.

Furthermore, note that $$\sigma''(\xi) = O(x)$$. So we have $$\sigma(x) = x(1 + O(x^2))$$.

From this, we have that for each layer $$\ell \in [K]$$, and for each $$b \in \{0, 1\}
\begin{align*}
\hat{M}^{(b, \ell)} &= \sigma(\hat{A}^{(b)}M^{(b,\ell-1)}W^{(\ell-1)}) = \hat{A}^{(b)}\hat{M}^{(b,\ell-1)}W^{(\ell-1)} \cdot \left(1 + O\left(\|\hat{M}^{(b,\ell-1)T}\|_{op, \infty}^2\|W^{(\ell-1)T}\|_{op, \infty}^2\right)\right),
\end{align*}
where the relative error expression comes from the fact that
\begin{equation}
\|\hat{A}^{(b)}\hat{M}^{(b,\ell-1)}W^{(\ell-1)}\|_{\infty} \leq \|\hat{A}^{(b)}\|_{\infty}\|\hat{M}^{(b,\ell-1)}\|_{op, \infty}\|W^{(\ell-1)}\|_{op, \infty} = O\left(\|\hat{M}^{(b,\ell-1)T}\|_{op, \infty}^2\|W^{(\ell-1)T}\|_{op, \infty}^2/n^2\right)
\end{equation}
Iterating the recurrence (75) gives us
\begin{equation}
\hat{M}^{(b, \ell)} = \hat{A}^{(b)}\ell \hat{M}^{(b,0)} \prod_{j=0}^{\ell-1} W^{(j)} \cdot \prod_{j=0}^{\ell-1} \left(1 + O\left(\|\hat{M}^{(b,j)T}\|_{op, \infty}^2\|W^{(j)}\|_{op, \infty}^2\right)\right).
\end{equation}

Now, we will show an upper bound on $$\|\hat{M}^{(b,j)T}\|_{op, \infty}$$. We will in particular show that
\begin{equation}
\|\hat{M}^{(b,j)T}\|_{op, \infty} \leq \prod_{i=0}^{j} \|W^{(i)T}\|_{op, \infty} \cdot (1 + O(n^{-1/2+\delta}))^j,
\end{equation}
for arbitrarily small $$\delta > 0$$, which implies, by our initial assumption, that $$\|\hat{M}^{(b,j)T}\|_{op, \infty} = O(1)$$ as $$n \to \infty$$. To show this, we apply the fact that $$|\sigma(x)| \leq |x|$$ for all $$x$$. This implies that
\begin{equation}
\|\hat{M}^{(b,j)T}\|_{op, \infty} \leq \|W^{(j-1)T}\|_{op, \infty} \cdot \|\hat{M}^{(b,j-1)T}\|_{op, \infty} \cdot \|\hat{A}^{(b)}T\|_{op, \infty}
\end{equation}
\begin{equation}
\leq (1 + O(n^{-1/2+\delta})) \|W^{(j-1)T}\|_{op, \infty} \cdot \|\hat{M}^{(b,j-1)T}\|_{op, \infty},
\end{equation}
with probability $$1 - e^{-\Theta(n)}$$. Iterating this recurrence, we get
\begin{equation}
\|\hat{M}^{(b,j)T}\|_{op, \infty} \leq (1 + O(n^{-1/2+\delta}))^j \prod_{i=0}^{j-1} \|W^{(i)T}\|_{op, \infty} \cdot \|\hat{M}^{(b,0)T}\|_{op, \infty},
\end{equation}
as claimed. The inequality (78) implies that as long as $$\ell \ll n^{1/2-\delta},
\hat{M}^{(b, \ell)} = \hat{A}^{(b)}\ell \hat{M}^{(b,0)} \prod_{j=0}^{\ell-1} W^{(j)} \cdot \prod_{j=0}^{\ell-1} \left(1 + O(1/n^2)\right).
\end{equation}
This implies that
\[
\hat{M}^{(b,\ell)} = \hat{A}^{(b,\ell)} \hat{M}^{(b,0)} \prod_{j=0}^{\ell-1} W^{(j)} \cdot (1 + O(n^{-3/2+\delta})).
\] (83)

Finally, this implies
\[
\|\hat{M}^{(0,K)} - \hat{M}^{(1,K)}\|_\infty \leq \left\| \hat{A}^{(0,K)} \hat{M}^{(0,0)} \prod_{j=0}^{K-1} W^{(j)} - \hat{A}^{(1,K)} \hat{M}^{(1,0)} \prod_{j=0}^{K-1} W^{(j)} \right\|_\infty (1 + O(n^{-3/2+\delta})),
\] (84)

so that we have reduced the problem to the linear case, as desired.

REFERENCES


