RESULTS FOR PERFECT CLASSIFICATION FOR GRAPH ATTENTION ON THE CONTEXTUAL STOCHASTIC BLOCK MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the ability of one layer Graph Attention Networks (GAT) to achieve perfect node classification for a simple synthetic data model called the contextual stochastic block model (CSBM). We determine a *positive* CSBM parameter regime such that GAT achieves perfect classification and a *negative* CSBM parameter regime such that GAT fails to achieve perfect classification. For the positive result we use a generalized attention mechanism of the original (Velickovic et al., 2018). For the negative result we consider a fixed attention mechanism which is determined using the labels of the nodes. We pose two questions. *Is the condition of GAT for achieving perfect classification better than that of a simple community detection method, i.e., thresholding the second principal eigenvector of the adjacency matrix (Abbe, 2018)?* The answer to this question is negative, and it depends on the parameter regime of the CSBM distribution. This happens because the graph information is coupled with the feature information using the operation of matrix multiplication. However, such matrix multiplication operation can be detrimental for perfect node classification. The second question is, *is the condition of GAT for achieving perfect classification better than that of simple graph convolution (GCN) (Kipf & Welling, 2017)?* We show that GAT is better than GCN if the attention mechanism of GAT is a Lipschitz function, while it is not better if it is not a Lipschitz function.

1 INTRODUCTION

Graph learning has received a lot of attention recently due to breakthrough learning models (Gori et al., 2005; Scarselli et al., 2009; Bruna et al., 2014; Duvenaud et al., 2015; Henaff et al., 2015; Atwood & Towsley, 2016; Defferrard et al., 2016; Hamilton et al., 2017; Kipf & Welling, 2017) that are able to exploit multi-modal data that consist of nodes and their edges as well as the features of the nodes. One of the most important problems in graph learning is the problem of node classification, where the goal is to classify the nodes of a graph given the node features and the topological graph information. One of the most popular models for node classification and graph learning in general is the Graph Attention Network (GAT). Graph attention (Velickovic et al., 2018) is usually defined as averaging the features of a node with the features of its neighbors by appropriately weighting the edges of a graph before spatially convolving the node features. Graph attention achieves this by using information from the given features for each node or given features for pairs of nodes.

In this work we explore the regimes where GAT can achieve perfect node classification for the contextual stochastic block model (CSBM) (Binkiewicz et al., 2017; Deshpande et al., 2018). The CSBM is a coupling of the stochastic block model (SBM) with a Gaussian mixture model, where the features of the nodes within a class are drawn from the same component of the mixture model. For a more precise definition, see Section 3. We focus on the case of two classes where the answer to the above question is sufficiently precise to understand the performance of graph attention and build useful intuition about it. Moreover, perfect classification is one of the three questions that has been traditionally asked for the SBM alone (Abbe, 2018). We see our work as an extension of previous work on perfect classification for SBM (Abbe, 2018) to CSBM using GAT. For more discussion see Section 2. One of the motivations for doing so is to understand the strengths and weaknesses of GAT on a controlled synthetic data model so that the graph learning community can build insights based on our findings.

We split our contributions in two parts. The first part compares one layer GAT to a simple community detection method, i.e., thresholding the second principal eigenvector of the adjacency matrix. The second part compares one layer GAT to one layer GCN.

1. *Regime where GAT's condition for perfect classification is not better than that of the simple community detection method.* In Theorem 1 we consider a fixed attention mechanism where the attention coefficients are determined using the labels of the nodes. Using ground-truth information to determine the attention coefficients is unrealistic, but it allows us to understand the CSBM parameter regime where GAT fails to achieve perfect classification in an ideal scenario. We show that if the distance between the means of the Gaussians is smaller than $\Theta(\sigma\sqrt{\log n/\Delta(p,q)})$, where σ is the standard deviation, n is the number of nodes and $\Delta(p,q)$ is the maximum expected intra- or inter-degree, then with high probability GAT fails to achieve perfect node classification for any parameter regime of the SBM. On the contrary, if the adjacency matrix has a large enough spectral gap, then the simple community detection method achieves perfect node classification.
2. *Regime where GAT's condition for perfect classification is better than that of the simple community detection method.* In Theorem 2 we consider a generalized attention mechanism of the original (Velickovic et al., 2018). We show two results. In the first result, if the distance of the means of the Gaussians is bigger than $\omega(\sigma\sqrt{\log n})$, then GAT achieves perfect node classification for any parameter setting of the SBM, while the simple community detection method requires a large enough spectral gap, otherwise it fails. While in the second result if the distance between the means is between $\Theta(\sigma\sqrt{\log n/\Delta(p,q)})$ and $\omega(\sigma\sqrt{\log n})$ and the adjacency matrix has a spectral gap then GAT and the community detection achieve perfect classification.
3. *Lipschitz GAT has up to a constant the same condition for perfect classification as GCN.* In Proposition 3 we show that if the attention function for GAT is a Lipschitz function then GCN and GAT have up to a constant the same separability threshold for the distance between the means to achieve perfect node classification.
4. *Non-Lipschitz GAT has a better condition for perfect classification than GCN.* In Section 6.2 we show that if the attention function of GAT is a non-Lipschitz function, then GAT has a better separability threshold for perfect node classification than GCN.

2 Previous work

Recently the concept of attention for neural networks (Bahdanau et al., 2015; Vaswani et al., 2017) was transferred to graph neural networks (Li et al., 2016; Bresson & Laurent, 2018; Velickovic et al., 2018; Lee et al., 2019; Puny et al., 2020). A few papers have attempted to understand the mechanism in (Velickovic et al., 2018). One work relevant to ours is (Brody et al., 2022). In this paper the authors show that a node may fail to assign large edge weight to its most important neighbors due to a global ranking of nodes that is generated by the attention mechanism in (Velickovic et al., 2018). Another related work is (Knyazev et al., 2019), which presents an empirical study of the ability of graph attention to generalize on larger, complex, and noisy graphs. In addition, in (Hou et al., 2019) the authors propose a different metric to generate the attention coefficients and show empirically that it has an advantage over the original GAT architecture.

Our work on GAT and the CSBM originates on previous work on understanding the performance of community detection methods for the SBM (Abbe, 2018). In previous work, researchers divided the parameter regimes of the SBM in three regimes. The first is the exact recovery regime, where the the community of all nodes is predicted correctly with high probability. In this paper we are also interested in predicting the class of all nodes correctly, i.e., perfect classification. However we do this for the graph learning model GAT and for the CSBM. We are not only interested in fitting the training data, our results hold for perfect classification for test data too.

The works by (Binkiewicz et al., 2017; Deshpande et al., 2018) are focused on the fundamental limits of learning for the CSBM. In this paper we are not interested in information theoretic limits, but we are interested in the parameter regime where GAT can achieve perfect classification. Of particular relevance is the work by (Baranwal et al., 2021), which studies the performance of graph convolution (Kipf & Welling, 2017) on CSBM as a semi-supervised learning problem. In this paper we extend this analysis to GAT.

The most relevant work is that of (Fountoulakis et al., 2022). In (Fountoulakis et al., 2022) it was shown that when the distance between the means of the Gaussians is $\omega(\sigma\sqrt{\log n})$ GAT is able to achieve perfect node separability (with high probability). In addition, it was shown that when the distance between the means is small, any attention architecture fails to distinguish inter from intra edges, which may hurdle the node classification performance.

Note that in (Fountoulakis et al., 2022) there is no condition provided for failing perfect classification. Thus it is not possible to compare GAT with community detection methods and GCN. In our paper we provide the regime where GAT fails to achieve perfect node classification when we provide ground-truth information to the attention mechanism

of GAT to distinguish edges. Finally, our results hold regardless if the intra-edge probability is bigger than the inter-edge probability for the SBM.

Finally, there are a few related theoretical works on understanding the performance and the universality of graph neural networks (Chen et al., 2019; Chien et al., 2021; Zhu et al., 2020; Xu et al., 2019; Garg et al., 2020; Loukas, 2020a;b; Jegelka, 2022). We provide theoretical results that characterize the precise performance of GAT compared to graph convolution and no convolution for CSBM with the goal of answering the particular questions that we imposed above.

3 PRELIMINARIES

Let $d, n \in \mathbb{N}$ be such that, and $\epsilon_1, \ldots, \epsilon_n \sim \text{Ber}(1/2)$, and define two classes as $C_k = \{j \in [n] \mid \epsilon_j = k\}$ for $k \in \{0, 1\}$. For each index $i \in [n]$, we draw feature vector $\mathbf{X}_i \in \mathbb{R}^d$ as $\mathbf{X}_i \sim N((2\epsilon_i - 1)\boldsymbol{\mu}, \mathbf{I} \cdot \sigma^2)$, where $\boldsymbol{\mu} \in \mathbb{R}^d$, $\sigma \in \mathbb{R}$ and $\mathbf{I} \in \{0, 1\}^{d \times d}$ is the identity matrix. For a given pair $p, q \in [0, 1]$ we consider the stochastic adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$ defined as follows. For $i, j \in [n]$ in the same class, we set $a_{ij} \sim \text{Ber}(p)$, and if i, j are in different classes, we set $a_{ij} \sim \text{Ber}(q)$[1]. We let $\mathbf{D} \in \mathbb{R}^{n \times n}$ be a diagonal matrix containing the degrees of the vertices. We denote by $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ a sample obtained according to the above random process. We consider node classification, and say that an estimator $\phi : \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times d} \times \mathbb{R}^{n \times n} \times [n] \to \mathbb{R}$ *perfectly classifies the nodes* if $\phi(\mathbf{X}, \mathbf{A}, i) < 0$ iff $i \in C_0$. We will use the following assumption on the parameters of the CSBM.

Assumption 1. *$|p - q| > 0$, $p, q = \Omega(\log^2 n/n)$ and σ is constant.*

The $|p - q| > 0$ part of the assumption implies that there are either more intra- or -inter-edges on expectation. The $p, q = \Omega(\log^2 n/n)$ part of the assumption implies degree concentration. See the definition of the event $\boldsymbol{\mathcal{E}}$ below. The latter part of the assumption is without loss of generality since all that really matters is the ratio of the distance between the means $\|\boldsymbol{\mu}\|$ with σ. Let N_i denote the set of neighbors of node $i \in [n]$ and define the following event.

Definition 1. *Event $\boldsymbol{\mathcal{E}}$ is the intersection of the following events over the randomness of $\mathbf{A}$ and $\{\epsilon_i\}_i$.*

1. *$\boldsymbol{\mathcal{E}}_1$ is the event that $|C_0| = \frac{n}{2} \pm O(\sqrt{n \log n})$ and $|C_1| = \frac{n}{2} \pm O(\sqrt{n \log n})$.*
2. *$\boldsymbol{\mathcal{E}}_2$ is the event that for each $i \in [n]$, $\mathbf{D}_{ii} = \frac{n(p+q)}{2}\left(1 \pm \frac{10}{\sqrt{\log n}}\right)$.*
3. *$\boldsymbol{\mathcal{E}}_3$ is the event that for each $i \in [n]$, $|C_0 \cap N_i| = \mathbf{D}_{ii} \cdot \frac{(1-\epsilon_i)p + \epsilon_i q}{p+q}\left(1 \pm \frac{10}{\sqrt{\log n}}\right)$ and $|C_1 \cap N_i| = \mathbf{D}_{ii} \cdot \frac{(1-\epsilon_i)q + \epsilon_i p}{p+q}\left(1 \pm \frac{10}{\sqrt{\log n}}\right)$.*

Using standard Chernoff bound and a union bound one can obtain the following.

Lemma 1. *Event $\boldsymbol{\mathcal{E}}$ holds with probability at least $1 - o_n(1)$.*

4 GAT VS COMMUNITY DETECTION FOR PERFECT CLASSIFICATION

Given two node representations $\mathbf{X}_i, \mathbf{X}_j$ for nodes $i, j \in [n]$ and edge feature $\boldsymbol{e}_{ij}$, let $\Psi(\mathbf{X}_i, \mathbf{X}_j, \boldsymbol{e}_{ij}) \stackrel{\text{def}}{=} \alpha(\boldsymbol{w}^T \mathbf{X}_i, \boldsymbol{w}^T \mathbf{X}_j, \boldsymbol{e}_{ij})$, where function α and parameters $\boldsymbol{w}$ are learnable. We refer to Ψ as the attention architecture with attention coefficients

$$\gamma_{ij} \stackrel{\text{def}}{=} \frac{\exp(\Psi(\mathbf{X}_i, \mathbf{X}_j, \boldsymbol{e}_{ij}))}{\sum_{k \in N_i} \exp(\Psi(\mathbf{X}_i, \mathbf{X}_k, \boldsymbol{e}_{ik}))} \quad \forall i, j \in E \text{ and } \gamma_{ij} = 0 \text{ otherwise.}$$

Let $\boldsymbol{\Gamma} = \{\gamma_{ij}\}_{ij} \in \mathbb{R}^{n \times n}$, the attention layer is defined as

$$\widehat{\boldsymbol{x}} = (\boldsymbol{\Gamma} \odot \mathbf{A})\mathbf{X}\boldsymbol{w},$$

where $\odot$ denotes the elementwise multiplication. Note that we consider only one layer since we work with a simple two block CSBM for binary classification. The graph convolution model (Kipf & Welling, 2017) (GCN) is obtained by setting $\boldsymbol{\Gamma} \stackrel{\text{def}}{=} \mathbf{D}^{-1}$ [2]. Observe that the graph information is coupled with the feature information using the operation of matrix multiplication. The intuition is that combining node features with topological information will favour

[1] Matrix $\mathbf{A}$ can contain self-loops as in $\mathbf{A} + \mathbf{I}$. In any case our analysis and conclusions are the same.

[2] Other versions of GCN are captured by setting $\boldsymbol{\Gamma}$ differently. For example $\gamma_{ij} \stackrel{\text{def}}{=} (d_i d_j)^{-1/2}$, where d_i, d_j are the degrees of nodes i, j, respectively. Our conclusions are the same in that case too due to degree concentration.

classification tasks. However, in this paper we show that such matrix multiplication operation can be detrimental for perfect node classification.

4.1 Warm-up: Hard Instance for GCN

Consider a one layer network where the matrix $\mathbf{A}$ contains all the required information for classification. In particular, consider the case where $\mathbf{A}$ is constructed using SBM with $p = \Omega(\log^2 n/n)$ and $q = 0$. In such case, the community structure is obtained easily by running a connected component algorithm.

Proposition 1. *Assume that $\mathbf{A} \sim \mathit{SBM}(p, q)$, for $p = \Omega(\log^2 n/n)$ and $q = 0$. Then a connected component algorithm perfectly classifies the nodes with probability at least $1 - o_n(1)$.*

The proof for this is trivial. Given the particular p and q then each class will be a connected component with high probability (Erdős & Rényi, 1960) while there will be no edges between the connected components. Using an algorithm for finding the connected components will output the correct class for each node with high probability.

Now think of the feature matrix $\mathbf{X}$ having no useful information. In particular, think of the features drawn from a Gaussian distribution $N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. That is, $\forall i \in [n]$ $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. In such case, with high probability GCN can't perfectly distinguish the classes.

Proposition 2. *Fix any $\|\boldsymbol{w}\| = 1$. Assume that $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ $\forall i$ and that $\mathbf{A} \sim \mathit{SBM}(p, q)$ with $p = \Omega(\log^2 n/n)$ and $q = 0$. Then, with probability at least $1-o_n(1)$ over $\mathbf{A}$, for any $i, j \in C_0 \times C_1$ we have that $(\mathbf{D}^{-1}\mathbf{A}\mathbf{X}\boldsymbol{w})_i$ follows the same distribution as $(\mathbf{D}^{-1}\mathbf{A}\mathbf{X}\boldsymbol{w})_j$. This implies that GCN fails to achieve perfect classification with probability at least $1 - o_n(1)$.*

The proof of Proposition 2 is an application of the proof of Proposition 3 since the assumption about Ψ in Proposition 3 implies up to a constant uniform edge weights, similar to GCN. Furthermore, note that $\|\boldsymbol{w}\| = 1$ is assumed without loss of generality since all we care about is the sign of the prediction and not its magnitude. Propositions 1 and 2 show that the operation of matrix multiplication is detrimental for perfect node classification for GCN as the multiplication operation harms the information captured in the matrix $\mathbf{A}$. This means that a simple connected component method applied on the graph succeeds at predicting the classes of all nodes with high probability, but GCN fails to achieve perfect classification with high probability.

4.2 Regime where GAT is not better than community detection

In the following we present a negative result for perfect classification for one layer GAT. We consider the case where the graph contains enough information for perfect classification. In particular, we will consider the following result.

Lemma 2 (Exact recovery in (Abbe, 2018))**.** *Suppose that $p, q = \Omega(\log^2 n/n)$ and $|\sqrt{p} - \sqrt{q}| > \sqrt{2\log n/n}$. Then, there exists a classifier $\hat{\tau}$ taking as input the graph $\mathbf{A}$ and perfectly classifies the nodes with probability at least $1 - o_n(1)$.*

Note that Lemma 2 only uses the graph. Therefore, the result is independent of the distance between the means of the Gaussians, while in Theorem 1, we prove that GAT has a limitation in the distance between the means for achieving perfect classification. We show that even with an oracle access to ground-truth information for the edges that allows the attention function to achieve perfect edge separability, perfect node classification won't happen with high probability. We start with the definition of *perfect edge separability*.

Definition 2. *Given a graph $G = ([n], E)$ and a partition of the nodes $[n] = C_0 \cup C_1$, we say that a function $\Psi : E \to \mathbb{R}$* perfectly separates the edges *if*

$$\operatorname{sign}(\Psi(i,j)) = \operatorname{sign}(p-q)\ \forall (i,j) \in (C_0^2 \cup C_1^2) \cap E \text{ and } \operatorname{sign}(\Psi(i,j)) = -\operatorname{sign}(p-q)\ \forall (i,j) \in (C_0^2 \cup C_1^2)^c \cap E,$$

The above definition says that if $p > q$ then the Ψ function should be positive if i, j are in the same class and negative if i, j are in different class. As we show in the proof of Theorem 1, this means that if $p > q$ then intra-edges have much higher attention coefficient γ_{ij} than inter-edges. The opposite is true when $q > p$.

Theorem 1. *Suppose $p, q < 1 - c$, where c is any arbitrary small constant. Fix any $\|\boldsymbol{w}\| = 1$, let $\Delta(p, q) = n\max(p, q)$ and assume that $\Psi(i, j)$ is a function that perfectly classifies the edges of any $\mathbf{A}$ with $|\Psi(i, j)| = \omega(1)$. If $\|\boldsymbol{\mu}\| \le K\sigma\sqrt{\frac{\log n}{\Delta(p,q)}}$ for some $K = O(1)$, then with probability at least $1 - o_n(1)$ over $(\mathbf{X}, \mathbf{A}) \sim \mathit{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, $\widehat{\boldsymbol{x}} = \sum_{j\in N_i} \gamma_{ij}\boldsymbol{w}^T\mathbf{X}_j$ fails to perfectly classify the nodes.*

Proof. Consider a fixed $\boldsymbol{w} \in \mathbb{R}^d$ with $\|\tilde{\boldsymbol{w}}\| = 1$. By the assumptions on Ψ we have that for any graph $G = ([n], E)$ represented by $\mathbf{A}$

$$\Psi(i,j) = \begin{cases} \operatorname{sign}(p-q)\omega(1) & \text{if } i,j \in (C_0^2 \cup C_1^2) \cap E \\ -\operatorname{sign}(p-q)\omega(1) & \text{otherwise} \end{cases}$$

By conditioning on event $\boldsymbol{\mathcal{E}}$, we have that with high probability each $i \in [n]$ has $np(1 \pm o_n(1))/2$ intra-class neighbors and $nq(1 \pm o_n(1))/2$ inter class neighbors. We focus on the case where $\Delta(p,q) = np$, that is $(p > q)$, as the converse is identical. Therefore, by definition of the attention coefficients

$$\gamma_{ij} = \begin{cases} \frac{2}{np}(1 \pm o_n(1)) & \text{if } \ i,j \in (C_0^2 \cup C_1^2) \cap E \\ o\left(\frac{1}{n(p+q)}\right) & \text{otherwise} \end{cases}.$$

Note that since Ψ is fixed, it is independent of $\mathbf{X}$. We consider the event $\tilde{\boldsymbol{\mathcal{E}}}$ for perfect classification for some fixed interceptor $b \in \mathbb{R}$ below

$$\left(-\boldsymbol{w}^T\boldsymbol{\mu}(1 \pm o(1)) + \sigma \sum_{j \in [n]} \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j + b\right) < 0 \qquad \forall i \in C_0$$

$$\left(\boldsymbol{w}^T\boldsymbol{\mu}(1 \pm o(1)) + \sigma \sum_{j \in [n]} \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j + b\right) > 0 \qquad \forall i \in C_1.$$

Observe that the above two conditions imply that at least one of them holds with $b = 0$. That is, if $b > 0$ the first condition holds with $b = 0$ and if $b < 0$ the second condition holds with $b = 0$.

Then, we bound

$$\mathbf{Pr}\left[\max_{i \in C_0} \sum_{j \in [n]} \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j < |\boldsymbol{w}^T\boldsymbol{\mu}|(1 \pm o(1))\right] \leq \mathbf{Pr}\left[\max_{i \in C_0} \sum_{j \in [n]} \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j < \|\boldsymbol{\mu}\|(1 \pm o(1))\right] \tag{1}$$

$$\mathbf{Pr}\left[\max_{i \in C_0} \sum_{j \in [n]} \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j < K\sqrt{\frac{\log n}{np}}(1 \pm o(1))\right] \tag{2}$$

Now we will use Sudakov minoration inequality (Vershynin, 2018) to obtain a lower bound on the expected supremum, and then apply Borell's inequality to upper bound the above probability.

In order to apply Sudakov's result we will need to define a canonical metric over the index set C_0. Let J_{ij} denote the set of uncommon neighbors of i and j, and note that

$$\rho \stackrel{\text{def}}{=} \sum_{k \in [n]} \gamma_{ik} a_{ik} \boldsymbol{w}^T \boldsymbol{g}_k - \sum_{k \in [n]} \gamma_{jk} a_{jk} \boldsymbol{w}^T \boldsymbol{g}_k = \frac{2(1 \pm o(1)}{np} \sum_{k \in J_{ij}} \boldsymbol{w}^T \boldsymbol{g}_k$$

For any $i, j \in C_0$:

$$d_\circ(i,j) = \sqrt{\mathbf{E}\left[\rho^2\right]} = \frac{2\sqrt{|J_{ij}|}}{np}(1 \pm o(1)).$$

Note that for any $i \neq j \in C_0$ and a node $k \in [n]$ the probability that k being a neighbor of exactly one of the nodes is $2p(1-p)$ if $k \in C_0$ and $2q(1-q)$ if $k \in C_1$. Therefore, by applying a Chernoff bound, we have that for any $\delta > 0$

$$\mathbf{Pr}[|J_{ij} - \mathbf{E}[|J_{ij}|]| > \delta\, \mathbf{E}[|J_{ij}|]] \leq 2\exp(-\delta^2\, \mathbf{E}[|J_{ij}|]/3).$$

By the fact that $p, q = \Omega(\log^2 n/n)$ we have that $\mathbf{E}[|J_{ij}|] = n(p(1-p) + q(1-q)) = \Omega(\log^2 n)$. Therefore, choosing $\delta = \frac{\sqrt{C \log n}}{\mathbf{E}[|J_{ij}|]}$ for any large constant $C > 0$, we get that the above holds with probability at least $1 - 1/\text{poly}(n)$. In addition, by the fact that $|J_{ij}| = \Omega(np)$ we have that $d_\circ(i,j) = \Omega(1/\sqrt{np})$. Under the concentration of J_{ij}s we let

$\epsilon_0 = \min_{i,j \in C_0} d_\circ(i,j)$. Then, for an ϵ_0-covering of the set, we need to have every point in the set (namely the cover $N(C_0, d_\circ, \epsilon_0) = |C_0|$. This implies that the expectation of the maxima is bounded below by

$$\mathbf{E}\left[\max_{i \in C_0} \sum_{j \in [n]} \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j\right] \geq c'' \epsilon_0 \sqrt{\log n} \geq c'' \sqrt{\frac{\log n}{np}},$$

for some $c'' > 0$. In addition, note that since Ψ is fixed, $\sum_j \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j$ is Gaussian with variance $O(1/np)$.

Now we can use Borell's inequality (Adler et al. (2007) chapter 2) to get that for any $t > 0$

$$\mathbf{Pr}\left[\max_{i \in C_0} \sum_{j \in [n]} \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j < \mathbf{E}\left[\max_{i \in C_0} \sum_{j \in [n]} \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j\right] - t\right] \leq 2\exp(-t^2 np).$$

By the lower bound on the expectation we have that the above implies that

$$\mathbf{Pr}\left[\max_{i \in C_0} \sum_{j \in [n]} \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j < c'' \sqrt{\frac{\log n}{np}} - t\right] \leq 2\exp(-t^2 np).$$

Pick $t = c'' \sqrt{\frac{\log n}{np}} - \frac{K\sqrt{\log n}}{\sqrt{np}} = \Omega(\sqrt{\log n / np})$, and combine with the event for the class and degree concentration to get

$$\mathbf{Pr}\left[\max_{i \in C_0} \sum_{j \in [n]} \gamma_{ij} a_{ij} \boldsymbol{w}^T \boldsymbol{g}_j \leq K \sqrt{\frac{\log n}{np}} \cdot (1 \pm o_n(1))\right] = \frac{1}{\text{poly}(n)} = o_n(1),$$

and the proof is complete. □

Remark 1. *Note that the setting described in Theorem 1 for the Ψ function can be easily realized when $|\sqrt{p} - \sqrt{q}| > \sqrt{2\log n / n}$. Under this setting we can use Lemma 2 to obtain classifier $\hat{\tau}$ and scale its values accordingly. In particular, by Lemma 2 we know that there exist a classifier $\hat{\tau}$ which perfectly classifies the nodes. Therefore, we can set $\Psi(i,j)$ as $\text{sign}(p-q)\omega(1)$ when $\hat{\tau}(i) = \hat{\tau}(j)$, and $-\text{sign}(p-q)\omega(1)$ otherwise. In addition, note that the case where p, q are close to 1 is degenerate–the graph drawn will be very close to the complete graph– which means that the graph holds no useful information.*

5 Regime where GAT is better than community detection

Let us now discuss a positive result for one layer GAT in Theorem 2. In particular, part 1 of Theorem 2 shows that regardless of what p and q are, if the distance between the means is $\kappa\sigma\sqrt{\log n}$, where $\kappa = \omega(1)$, then GAT is able to perfect classify any input from CSBM with high probability. This is better than simple community detection which depends on the values of p and q, see Lemma 2. Unfortunately, as noted in Remark 2, a simple linear classifier without the graph data is enough to achieve perfect classification in this regime of the distance of the means.

Remark 2. *If $\|\boldsymbol{\mu}\| \geq \kappa\sigma\sqrt{\log n}$, a simple linear classifier on $\mathbf{X}$ obtains perfect classification (Anderson, 2003).*

In part 2 of Theorem 2 we show that GAT is able to achieve perfect classification for a limited range of the distance between the means and the same condition on p and q as in Lemma 2.

Theorem 2. *Suppose that $p, q = \Omega(\log^2 n / n)$, fix $\kappa = \omega(1)$, $K = O(1)$ and define $\Delta(p,q) = n\max(p,q)$. Then, there exists a choice of attention architecture Ψ such that with probability at least $1 - o_n(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim$ CSBM$(n, p, q, \boldsymbol{\mu}, \sigma^2)$, the estimator*

$$\widehat{x}_i \stackrel{\text{def}}{=} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \mathbf{X}_j \quad \textit{where } \boldsymbol{w} = \text{sign}(p-q)\boldsymbol{\mu}/\|\boldsymbol{\mu}\|,$$

satisfies the following

1. *If $\|\boldsymbol{\mu}\| \geq \kappa\sigma\sqrt{\log n}$, the estimator classifies correctly all the nodes.*

2. *If* $K\sigma\sqrt{\frac{\log n}{\Delta(p,q)}} < \|\boldsymbol{\mu}\| < \kappa\sigma\sqrt{\log n}$ *and* $|\sqrt{p}-\sqrt{q}| > \sqrt{2\log n/n}$*, then the estimator classifies correctly all the nodes.*

Proof. Define

$$\boldsymbol{w} \stackrel{\text{def}}{=} \operatorname{sign}(p-q)\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}, \qquad \mathbf{S} \stackrel{\text{def}}{=} \begin{bmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad \boldsymbol{r} \stackrel{\text{def}}{=} R \cdot \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

where $R > 0$ is an arbitrary scaling parameter, and consider the following definition of function α:

$$\alpha(\boldsymbol{w}^T\mathbf{X}_i, \boldsymbol{w}^T\mathbf{X}_j) = \operatorname{sign}(p-q)t(\|\boldsymbol{\mu}\|,\sigma,n)\cdot \boldsymbol{r}^T \cdot \operatorname{LeakyRelu}\left(\mathbf{S}\cdot\begin{bmatrix}\boldsymbol{w}^T\mathbf{X}_i \\ \boldsymbol{w}^T\mathbf{X}_j\end{bmatrix}\right),$$

where $t(\|\boldsymbol{\mu}\|,\sigma,n) = 1$ if $\|\boldsymbol{\mu}\| \geq \kappa\sigma\sqrt{\log n}$ and 0 otherwise.

We construct the edge features based on $\widehat{\tau}$ from Lemma 2, $\boldsymbol{e}_{ij} = (\widehat{\tau}(i), \widehat{\tau}(j))$ for every $i, j \in E$. Define

$$\alpha'(\boldsymbol{e}_{ij}) \stackrel{\text{def}}{=} (1 - t(p,q,n)) \cdot \begin{cases} \operatorname{sign}(p-q)R' & \text{if } \widehat{\tau}(i) = \widehat{\tau}(j) \\ -\operatorname{sign}(p-q)R' & \text{otherwise} \end{cases}, \tag{3}$$

where $R' > 0$ is some scaling constant. Finaly, we let

$$\Psi(\mathbf{X}_i, \mathbf{X}_j, \boldsymbol{e}_{ij}) = \alpha(\boldsymbol{w}^T\mathbf{X}_i, \boldsymbol{w}^T\mathbf{X}_j) + \alpha'(\boldsymbol{e}_{ij}).$$

Suppose that $\|\boldsymbol{\mu}\| \geq \kappa\sigma\sqrt{\log n}$, so that

$$\Psi(\mathbf{X}_i, \mathbf{X}_j, \boldsymbol{e}_{ij}) = \operatorname{sign}(p-q)\boldsymbol{r}^T \cdot \operatorname{LeakyRelu}\left(\mathbf{S}\cdot\begin{bmatrix}\boldsymbol{w}^T\mathbf{X}_i \\ \boldsymbol{w}^T\mathbf{X}_j\end{bmatrix}\right).$$

However, this is exactly the setting dealt with in (Fountoulakis et al., 2022) (see Corollary 3). Therefore, when $\|\boldsymbol{\mu}\| \geq \kappa\sigma\sqrt{\log n}$ the estimator achieves perfect node classification with probability at least $1 - o_n(1)$. This concludes the first item of the theorem.

Now suppose that $K\sigma\sqrt{\frac{\log n}{\Delta(p,q)}} < \|\boldsymbol{\mu}\| < \kappa\sigma\sqrt{\log n}$, and $|\sqrt{p}-\sqrt{q}| > \sqrt{2\log n/n}$.

In this case, with probability at least $1 - o_n(1)$

$$\Psi(\mathbf{X}_i, \mathbf{X}_j, \boldsymbol{e}_{ij}) = \alpha'(\boldsymbol{e}_{ij}) = \begin{cases} \operatorname{sign}(p-q)R' & \text{if } i,j \in (C_1^2 \cup C_0^2) \cap E \\ -\operatorname{sign}(p-q)R' & \text{otherwise} \end{cases}, \tag{4}$$

since $t(\|\boldsymbol{\mu}\|,\sigma,n) = 0$ and the condition $|\sqrt{p}-\sqrt{q}| > \sqrt{2\log n/n}$ guarantees (by Lemma 2) that with probability at least $1 - o_n(1)$ $\widehat{\tau}$ achieves perfect classification. We let $R' = \omega(1)$ for the rest of the proof and note that Ψ is independent of $\mathbf{X}$.

We focus on the case where $\operatorname{sign}(p-q) > 0$, the proof for the opposite case is identical. By Lemma 1, conditioned on event $\boldsymbol{\mathcal{E}}$ which holds with probability at least $1 - o(1)$, the values of γ_{ij} satisfy

$$\gamma_{ij} = \begin{cases} \frac{1}{\frac{np}{2} + \frac{nq}{2}\exp(-2R')} & \text{if } i,j \in (C_0^2 \cup C_1^2) \cap E \\ \frac{1}{\frac{np}{2}\exp(2R') + \frac{nq}{2}} & \text{otherwise} \end{cases}. \tag{5}$$

By the fact that $p > q$ and $p, q = \Omega(\log^2 n/n)$, we get that $\gamma_{ij} = \frac{2}{np}(1 \pm o_n(1))$ for $i, j \in E$ in the same class and $1/\operatorname{poly}(n)$ otherwise. Note that this holds regardless of the size of $\|\boldsymbol{\mu}\|$.

Conditioned on $\boldsymbol{\mathcal{E}}$ and the fact that Ψ is independent of $\mathbf{X}$, the random variables $\widehat{x}_i$ are Gaussians with standard deviation $\tilde{\sigma} = O(\sigma/\sqrt{np})$. In addition, when $i \in C_1$

$$\mathbf{E}\left[\sum_{j\in N_i}\gamma_{ij}\boldsymbol{w}^T\mathbf{X}_j\middle|\boldsymbol{\mathcal{E}}\right] = \mathbf{E}\left[\sum_{j\in C_0\cap N_i}\gamma_{ij}\boldsymbol{w}^T\mathbf{X}_j+\sum_{j\in C_1\cap N_i}\gamma_{ij}\boldsymbol{w}^T\mathbf{X}_j\middle|\boldsymbol{\mathcal{E}}\right]$$

$$= |C_1\cap N_i|\left(\frac{2}{np}(1\pm o_n(1))\,\mathbf{E}[\boldsymbol{w}^T\mathbf{X}_j\mid\boldsymbol{\mathcal{E}}, j\in N_i\cap C_1]\right)+|C_0\cap N_i|\left(\frac{1}{\mathrm{poly}(n)}\,\mathbf{E}[\boldsymbol{w}^T\mathbf{X}_j\mid\boldsymbol{\mathcal{E}}, j\in N_i\cap C_0]\right)$$

$$= (1\pm o_n(1))\,\mathbf{E}[\boldsymbol{w}^T\mathbf{X}_j\mid\boldsymbol{\mathcal{E}}, j\in N_i\cap C_1]-\frac{q(1\pm o_n(1))}{\mathrm{poly}(n)}\|\boldsymbol{\mu}\| = (1\pm o_n(1))\frac{\mathbf{E}[\boldsymbol{w}^T\mathbf{X}_j\cdot\mathbf{1}_{\boldsymbol{\mathcal{E}}}\mid j\in N_i\cap C_1]}{\mathbf{Pr}[\boldsymbol{\mathcal{E}}]}\pm o_n(1)$$

$$= \|\boldsymbol{\mu}\|_2(1\pm o_n(1)).$$

Using the same reasoning, we get for $i\in C_0$, $\mathbf{E}[\widehat{x}_i|\boldsymbol{\mathcal{E}}] = -\|\boldsymbol{\mu}\|_2(1\pm o_n(1))$

Next, we bound the probability of misclassification for $i\in C_0$

$$\mathbf{Pr}\left[\max_{j\in C_0}\widehat{x}_j\geq 0\right]\leq\mathbf{Pr}\left[\max_{j\in C_0}\widehat{x}_j>t+\mathbf{E}[\widehat{x}_i]\right],$$

for $t<|\,\mathbf{E}[\widehat{x}_i]| = \|\boldsymbol{\mu}\|_2(1\pm o_n(1))$

Lemma 3 (Rigollet & Hütter (2015))**.** *Let $x_1,\ldots,x_n$ be sub-Gaussian random variables with the same mean and sub-Gaussian parameter $\tilde{\sigma}^2$. Then,*

$$\mathbf{E}\left[\max_{i\in[n]}(x_i-\mathbf{E}[x_i])\right]\leq\tilde{\sigma}\sqrt{2\log n}.$$

Moreover, for any $t>0$

$$\mathbf{Pr}\left[\max_{i\in[n]}(x_i-\mathbf{E}[x_i])>t\right]\leq 2n\exp\left(-\frac{t^2}{2\tilde{\sigma}^2}\right).$$

By $K\sigma\sqrt{\frac{\log n}{np}}<\|\boldsymbol{\mu}\|$, we can pick $t=\Theta\left(\frac{\sigma}{\sqrt{np}}\sqrt{\log|C_0|}\right)$ and applying Lemma 3 implies that the above probability is $1/\mathrm{poly}(n)$.

Similarly for $i\in C_1$ we have that the misclassification probability is

$$\mathbf{Pr}\left[\min_{j\in C_1}\widehat{x}_j\leq 0\right]=\mathbf{Pr}\left[-\max_{j\in C_1}(-\widehat{x}_j)\leq 0\right]=\mathbf{Pr}\left[\max_{j\in C_1}(-\widehat{x}_j)\geq 0\right]\leq\mathbf{Pr}\left[\max_{j\in C_1}-\widehat{x}_j>t-\mathbf{E}[\widehat{x}_i]\right],$$

for $t<\mathbf{E}[\hat{x}_i]$. Picking $t=\Theta\left(\frac{\sigma}{\sqrt{np}}\sqrt{\log|C_1|}\right)$ and applying Lemma 3 and a union bound over the misclassification probabilities of both classes conclude the proof.

□

6 GAT vs GCN

In this section we compare one layer GAT to GCN when the attention function Ψ is non-Lipschitz and Lipschitz.

6.1 Lipschitz GAT is not better than GCN

In this section we will show that considering bounded attention architectures (Lipschitz function Ψ) have no additional advantage over GCN in terms of improving the threshold where GAT fails to perfectly classify the nodes. Similarly to Theorem 1 we will assume that we have an oracle access to a perfect *bounded* attention architecture.

Proposition 3. *Suppose $p,q<1-c$, where c is any arbitrary small constant. Fix any $\|\boldsymbol{w}\|=1$ and assume $\Psi(i,j)$ is a function that perfectly separates the edges of any graph $\mathbf{A}$ with $|\Psi(i,j)|=O(1)$ for all $i,j\in E$. If $\|\boldsymbol{\mu}\|\leq K\sigma\sqrt{\frac{\log n}{n(p+q)}}\cdot\frac{p+q}{|p-q|}$ and $K=O(1)$, then with probability at least $1-o_n(1)$ over $(\mathbf{X},\mathbf{A})\sim$ CSBM$(n,p,q,\boldsymbol{\mu},\sigma^2)$, $\widehat{\boldsymbol{x}}=\sum_{j\in N_i}\gamma_{ij}\boldsymbol{w}^T\mathbf{X}_j$ fails to perfectly classify the nodes.*

Proof. The proof follows the same techniques as Theorem 1. Fix $\boldsymbol{w} \in \mathbb{R}^d$ with $\|\boldsymbol{w}\| = 1$. By the assumptions on Ψ we have that for any graph $G = ([n], E)$ represented by $\mathbf{A}$

$$\Psi(i,j) = \begin{cases} \text{sign}(p-q)O(1) & \text{if } i,j \in (C_0^2 \cup C_1^2) \cap E \\ -\text{sign}(p-q)O(1) & \text{otherwise.} \end{cases}.$$

We focus in the case where $p > q$ since the other case is identical. By conditioning on event $\boldsymbol{\mathcal{E}}$, we have that with high probability each $i \in [n]$ has $np(1 \pm o_n(1))/2$ intra-class neighbors and $nq(1 \pm o_n(1))/2$ inter class neighbors. Therefore, by definition of the attention coefficients, for every $i, j \in E$

$$\gamma_{ij} = \Theta\left(\frac{1}{n(p+q)}\right) = \Theta\left(\frac{1}{|N_i|}\right).$$

We consider the conditions for perfect classification for some fixed interceptor $b \in \mathbb{R}$

$$\Theta\left(-\frac{p-q}{p+q}\boldsymbol{w}^T\boldsymbol{\mu}(1 \pm o_n(1)) + \max_{i \in C_0} \frac{\sigma}{|N_i|} \sum_{j \in [n]} a_{ij}\boldsymbol{w}^T\boldsymbol{g}_j\right) + b < 0$$

$$\Theta\left(\frac{p-q}{p+q}\boldsymbol{w}^T\boldsymbol{\mu}(1 \pm o_n(1)) + \min_{i \in C_1} \frac{\sigma}{|N_i|} \sum_{j \in [n]} a_{ij}\boldsymbol{w}^T\boldsymbol{g}_j\right) + b > 0.$$

The rest of the proof proceeds as Theorem 1. □

From Proposition 3 we have that for bounded functions Ψ if $\|\boldsymbol{\mu}\| < K\sigma\sqrt{\log n/n(p+q)} \cdot (p+q/|p-q|)$ for some $K = O(1)$, then with high probability GAT fails to perfectly classify the nodes. However, this condition is up to a constant the condition for GCN to fail to perfectly classify the nodes.

Remark 3. *For any $\|\boldsymbol{w}\| = 1$ the condition $\|\boldsymbol{\mu}\| \leq K\sigma\sqrt{\log n/n(p+q)} \cdot (p+q/|p-q|)$ and $K = O(1)$ is the condition for GCN to fail to perfectly classify the nodes. The proof is an application of the proof of Proposition 3 since the assumption about Ψ in Proposition 3 implies up to a constant uniform edge weights, similar to GCN.*

Note that the condition for one layer GAT and GCN for perfect classification is $\|\boldsymbol{\mu}\| \geq \omega(\sigma\sqrt{\log n/n(p+q)} \cdot (p+q/|p-q|))$. The proof for both GAT and GCN is identical since their edge coefficients are the same up to a constant, and it has appeared in (Baranwal et al., 2021). The statement can also be proved easily by following similar arguments as in the proof of Theorem 2. Therefore, the conditions for achieving and failing perfect classification of GAT are up to a constant the same as that of GCN.

6.2 Non-Lipschitz GAT is better than GCN

From part 2 of Theorem 2 we have that the threshold on the distance between the means that makes the data perfectly separable for GAT is up to a constant $\sigma\sqrt{\log n/\Delta(p,q)}$ when $|\sqrt{p} - \sqrt{q}| > \sqrt{2\log n/n}$. While the threshold for GCN to fail perfect classification is up to a constant $\sigma\sqrt{\log n/n(p+q)} \cdot (p+q/|p-q|)$. If $p > q$, then the threshold for perfect classification of GAT is lower than that of GCN when $3p > q$. Since we assumed that $p > q$ then GAT achieves perfect classification when GCN fails. The same conclusion is reached if we assume that $q > p$. On the other hand if $|\sqrt{p} - \sqrt{q}| \leq \sqrt{2\log n/n}$, then GAT can be reduced to GCN by setting the Ψ function to zero. In this case, GAT is equivalent to GCN. Overall, there exist a parameter regime where GAT perfectly classifies the data, while for the rest of the parameters GAT can be equivalent to GCN.

7 Conclusion and limitations of our work

Our work shows the benefits and limitations of GAT versus a simple community detection method and GCN on the CSBM. There are two limitations to our analysis that we leave for future work. First, the negative result in Theorem 1 holds for a fixed $\boldsymbol{w}$. Although, we do show in our proofs for the positive result that there exist fixed parameters that can perform well, it would be useful to obtain a negative result where $\boldsymbol{w}$ is not fixed and it is a function of the training data. Second, the negative result in Theorem 1 holds for a fixed attention function Ψ. We do set the Ψ function according to ground-truth labels of the nodes to demonstrate an ideal scenario of picking Ψ, but it would be useful to obtain a similar result for a Ψ that depends on the data as we did for the positive results in Theorem 2.

REFERENCES

E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18:1–86, 2018.

Robert J Adler, Jonathan E Taylor, et al. *Random fields and geometry*, volume 80. Springer, 2007.

T.W. Anderson. *An introduction to multivariate statistical analysis*. John Wiley & Sons, 2003.

J. Atwood and D. Towsley. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2001–2009, 2016.

D. Bahdanau, K. H. Cho, , and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.

A. Baranwal, K. Fountoulakis, and A. Jagannath. Graph convolution for semi-supervised classification: Improved linear separability and out-of-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 684–693, 2021.

N. Binkiewicz, J. T. Vogelstein, and K. Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104:361–377, 2017.

X. Bresson and T. Laurent. Residual gated graph convnets. In *arXiv:1711.07553*, 2018.

S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2022.

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

Z. Chen, L. Li, and J. Bruna. Supervised community detection with line graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

E. Chien, J. Peng, P. Li, and O. Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations (ICLR)*, 2021.

M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3844–3852, 2016.

Y. Deshpande, A. Montanari S. Sen, and E. Mossel. Contextual stochastic block models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 45, pp. 2224–2232, 2015.

P. Erdős and A. Rényi. On the evolution of random graphs. In *Publication of the Mathematical Institute of the Hungarian Academy Sciences*, pp. 17–61, 1960.

Kimon Fountoulakis, Amit Levi, Shenghao Yang, Aseem Baranwal, and Aukosh Jagannath. Graph attention retrospective. *arXiv preprint arXiv:2202.13060*, 2022.

V. Garg, S. Jegelka, and T. Jaakkola. Generalization and representational limits of graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 119, pp. 3419–3430, 2020.

M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2005.

W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1025–1035, 2017.

M. Henaff, J. Bruna, and Y. LeCun. Deep convolutional networks on graph-structured data. In *arXiv:1506.05163*, 2015.

Y. Hou, J. Zhang, J. Cheng, K. Ma, R. T. B. Ma, H. Chen, and M.-C. Yang. Measuring and improving the use of graph information in graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

S. Jegelka. Theory of graph neural networks: Representation and learning. In *arXiv:2204.07697*, 2022.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

B. Knyazev, G. W. Taylor, and M. Amer. Understanding attention and generalization in graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4202–4212, 2019.

B. J. Lee, R. A. Rossi, S. Kim, K. N. Ahmed, and E. Koh. Attention models in graphs: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2019.

Y. Li, R. Zemel, M. Brockschmidt, and D. Tarlow. Gated graph sequence neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.

A. Loukas. How hard is to distinguish graphs with graph neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

A. Loukas. What graph neural networks cannot learn: Depth vs width. In *International Conference on Learning Representations (ICLR)*, 2020b.

O. Puny, H. Ben-Hamu, and Y. Lipman. Global attention improves graph networks generalization. In *arXiv:2006.07846*, 2020.

P. Rigollet and J.-C. Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813:814, 2015.

F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1), 2009.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6000–6010, 2017.

P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, and D. Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.