Graph Attention Retrospective

Kimon Fountoulakis[*]   KIMON.FOUNTOULAKIS@UWATERLOO.CA
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

Amit Levi[*]   AMIT.LEVI@UWATERLOO.CA
Huawei Noah's Ark Lab
Montreal, Quebec, Canada

Shenghao Yang[*]   SHENGHAO.YANG@UWATERLOO.CA
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

Aseem Baranwal   ASEEM.BARANWAL@UWATERLOO.CA
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

Aukosh Jagannath   A.JAGANNATH@UWATERLOO.CA
Department of Statistics and Actuarial Science, Department of Applied Mathematics
University of Waterloo
Waterloo, Ontario, Canada

Editor: Joan Bruna

Abstract

Graph-based learning is a rapidly growing sub-field of machine learning with applications in social networks, citation networks, and bioinformatics. One of the most popular models is graph attention networks. They were introduced to allow a node to aggregate information from features of neighbor nodes in a non-uniform way, in contrast to simple graph convolution which does not distinguish the neighbors of a node. In this paper, we theoretically study the behaviour of graph attention networks. We prove multiple results on the performance of the graph attention mechanism for the problem of node classification for a contextual stochastic block model. Here, the node features are obtained from a mixture of Gaussians and the edges from a stochastic block model. We show that in an "easy" regime, where the distance between the means of the Gaussians is large enough, graph attention is able to distinguish inter-class from intra-class edges. Thus it maintains the weights of important edges and significantly reduces the weights of unimportant edges. Consequently, we show that this implies perfect node classification. In the "hard" regime, we show that every attention mechanism fails to distinguish intra-class from inter-class edges. In addition, we show that graph attention convolution cannot (almost) perfectly classify the nodes even if intra-class edges could be separated from inter-class edges. Beyond perfect node classification, we provide a positive result on graph attention's robustness against structural noise in the graph. In particular, our robustness result implies that graph attention

---

[*]. Equal contribution.

can be strictly better than both the simple graph convolution and the best linear classifier of node features. We evaluate our theoretical results on synthetic and real-world data.

Keywords: graph neural networks, attention mechanism, contextual stochastic block model, high-dimensional probability, node classification

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 and Towsley, 2016; Defferrard et al., 2016; Hamilton et al., 2017; Kipf and 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 *classification*, where the goal is to classify the nodes or edges of a graph given the graph and the features of the nodes. Two of the most fundamental mechanisms for classification, and graph learning in general, are graph convolution and graph attention. Graph convolution, usually defined using its spatial version, corresponds to averaging the features of a node with the features of its neighbors (Kipf and Welling, 2017).[1] Graph attention (Veličković et al., 2018) mechanisms augment this convolution by appropriately weighting the edges of a graph before spatially convolving the data. Graph attention is able to do this by using information from the given features for each node. Despite its wide adoption by practitioners (Fey and Lenssen, 2019; Wang et al., 2019; Hu et al., 2020) and its large academic impact as well, the number of works that rigorously study its effectiveness is quite limited.

One of the motivations for using a graph attention mechanism as opposed to a simple convolution is the expectation that the attention mechanism is able to distinguish inter-class edges from intra-class edges and consequently weights inter-class edges and intra-class edges differently before performing the convolution step. This ability essentially maintains the weights of important edges and significantly reduces the weights of unimportant edges, and thus it allows graph convolution to aggregate features from a subset of neighbor nodes that would help node classification tasks. In this work, we explore the regimes in which this heuristic picture holds in simple node classification tasks, namely classifying the nodes in a 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 2. 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. We briefly and informally summarize our contributions as follows:

1. In the "easy regime", i.e., when the distance between the means of the Gaussian mixture model is much larger than the standard deviation, we show that there exists a choice of attention architecture that distinguishes inter-class edges from intra-class

---

1. Although the model in Kipf and Welling (2017) is related to spectral convolutions, it is mainly a spatial convolution since messages are propagated along graph edges. More broadly, graph convolution can refer to different variants arising from different (approximations of) graph spectral filters. We provide details in Section 2.

edges with high probability (Theorem 3). In particular, we show that the attention coefficients for one class of edges are much higher than the other class of edges (Corollary 4). Furthermore, we show that these attention coefficients lead to a perfect node classification result (Corollary 5). However, in the same regime, we also show that the graph is not needed to perfectly classify the data (Proposition 7).

2. In the "hard regime", i.e., when the distance between the means is small compared to the standard deviation, we show that *every* attention architecture is unable to distinguish inter-class from intra-class edges with high probability (Theorem 9). Moreover, we show that using the original Graph Attention Network (GAT) architecture (Veličković et al., 2018), with high probability, most of the attention coefficients are going to have uniform weights, similar to those of simple graph convolution (Kipf and Welling, 2017) (Theorem 10). We also consider a setting where graph attention is provided with independent and perfect edge information, and we show that even in this case, if the distance between the means is sufficiently small, then graph attention would fail to (almost) perfectly classify the nodes with high probability (Theorem 13).

3. Beyond perfect node classification, we show that graph attention is able to assign a significantly higher weight to self-loops, irrespective of the parameters of the Gaussian mixture model that generates node features or the stochastic block model that generates the graph. Consequently, we show that with a high probability, graph attention convolution is at least as good as the best linear classifier for node features (Theorem 14). In a high structural noise regime when the graph is not helpful for node classification, i.e., when intra-class edge probability p is close to inter-class edge probability q, this means that graph attention is strictly better than simple graph convolution because it is able to essentially ignore the graph structural noise (Corollary 15). In addition, we obtain lower bounds on graph attention's performance for almost perfect node classification and partial node classification (Corollary 16).

4. We provide an extensive set of experiments both on synthetic data and on three popular real-world datasets that validates our theoretical results.

1.1 Limitations of our theoretical setting

In this work, to study the benefits and the limitations of the graph attention mechanism, we focus on node classification tasks using the CSBM generative model with two classes and Gaussian node features. This theoretical setting has a few limitations. First, we note that many real-world networks are much more complex and may exhibit different structures than the ones obtainable from a stochastic block model. Furthermore, real-world node attributes such as categorical features may not even follow a continuous probability distribution. Apparently, there are clear gaps between CSBM and the actual generative processes of real-world data. Nonetheless, we note that a good understanding of graph attention's behavior over CSBM would already provide us with useful insights. For example, it has been shown empirically that synthetic graph datasets based on CSBM constitute good benchmarks for evaluating various GNN architectures (Palowitch et al., 2022). To that end, in Section 6, we summarize some practical implications of our theoretical results which could be useful for practitioners working with GNNs.

Second, there are different levels of granularity for machine learning problems on graphs. Besides node-level tasks such as node classification, other important problems include edge-level tasks such as link prediction and graph-level tasks such as graph classification. While, empirically, variants of graph attention networks have been successfully applied to solve problems at all levels of granularity, our theoretical results mostly pertain to the effects of graph attention for node classification. This is a limitation on the scope of our study. On the other hand, we note that edge-level and graph-level tasks are typically solved by adding a pooling layer which combines node representations from previous graph (attention) convolution layers. Consequently, the quality of graph attention convolution's output for a node not only affects node classification but also has a major influence on link prediction and graph classification. Therefore, our results may be extended to study the effects of graph attention on link prediction and graph classification under the CSBM generative model. For example, link prediction is closely related to edge classification which we extensively study in Section 3. For graph classification, one may consider the problem of classifying graphs generated from different parameter settings of the CSBM. In this case, our results on graph attention's impact on node representations may be used to establish results on graph attention's impact on graph representations after certain pooling layers. In particular, if graph attention convolution generates good node representations that are indicative of node class memberships, then one can show that the graph representation obtained from appropriately aggregating node representations would be indicative of the clustering structure of the graph, and hence the graph representation would be useful for graph classification tasks.

1.2 Relevant 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 and Laurent, 2018; Veličković et al., 2018; Lee et al., 2019; Puny et al., 2020). A few papers have attempted to understand the attention mechanism in Veličković 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 generated by the attention mechanism in Veličković 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.

Other related work to ours, which does not focus on graph attention, comes from the field of statistical learning on random data models. Random graphs and the stochastic block model have been traditionally used in clustering and community detection (Abbe, 2018; Athreya et al., 2018; Moore, 2017). Moreover, the works by Binkiewicz et al. (2017); Deshpande et al. (2018), which also rely on CSBM are focused on the fundamental limits of unsupervised learning. Of particular relevance is the work by Baranwal et al. (2021), which studies the performance of graph convolution on CSBM as a semi-supervised learning problem. Within the context of random graphs, Keriven et al. (2021) studies the approximation power of graph neural networks on random graphs. In Maskey et al. (2022) the authors

derive generalization error of graph neural networks for graph classification and regression tasks. In our paper we are interested in understanding graph attention's capability for edge/node classification with respect to different parameter regimes of CSBM.

Finally, there are a few related theoretical works on understanding the generalization and representation power 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). For a recent survey in this direction see Jegelka (2022). Our work takes a statistical perspective which allows us to characterize the precise performance of graph attention compared to graph convolution and no convolution for CSBM, with the goal of answering the particular questions that we imposed above.

2. Preliminaries

In this section, we describe the *Contextual Stochastic Block Model (CSBM)* (Deshpande et al., 2018) which serves as our data model, and the *Graph Attention* mechanism (Veličković et al., 2018). We also define notations and terms that are frequently used throughout the paper.

For a vector $x \in \mathbb{R}^n$ and $n \in \mathbb{N}$, the norm $\|x\|$ denotes the Euclidean norm of x, i.e. $\|x\| \overset{\text{def}}{=} \sum_{i \in [n]} x_i^2$. We write $[n] \overset{\text{def}}{=} \{1, 2, \ldots, n\}$. We use $\text{Ber}(p)$ to denote the Bernoulli distribution, so $x \sim \text{Ber}(p)$ means the random variable x takes value 1 with probability p and 0 with probability $1 - p$. Let $d, n \in \mathbb{N}$, and $\epsilon_1, \ldots, \epsilon_n \sim \text{Ber}(1/2)$. 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 set the feature vector $\mathbf{X}_i \in \mathbb{R}^d$ as $\mathbf{X}_i \sim N((2\epsilon_i - 1)\boldsymbol{\mu}, \sigma^2 \mathbf{I})$, where $\boldsymbol{\mu} \in \mathbb{R}^d$, $\sigma \in \mathbb{R}$ and $\mathbf{I} \in \{0, 1\}^{d \times d}$ is the identity matrix.[2] 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 (i.e., *intra-class edge*), we set $a_{ij} \sim \text{Ber}(p)$, and if i, j are in different classes (i.e., *inter-class edge*), we set $a_{ij} \sim \text{Ber}(q)$. We denote by $(\mathbf{X}, \mathbf{A}) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ a sample obtained according to the above random process.

An advantage of CSBM is that it allows us to control the noise by controlling the parameters of the distributions of the model. In particular, CSBM allows us to control the distance of the means and the variance of the Gaussians, which are important for controlling the separability of the Gaussians. For example, fixing the variance, then the closer the means are the more difficult the separation of the node Gaussian features becomes. Another notable advantage of CSBM is that it allows us to control the structural noise and homophily level of the graph. The level of noise in the graph is controlled by increasing or reducing the gap between the intra-class edge probability p and the inter-class edge probability q, and the level of homophily in the graph is controlled by the relative magnitude between p and q. For example, when p is much larger than q, then a node is more likely to be connected with a node from the same class, and hence we obtain a homophilous graph; when q is much larger than p, then a node is more likely to be connected with a node from a different class, and hence we obtain a heterophilous graph. There are several recent works exploring the behavior of GNNs in heterophilous graphs (Lim et al., 2021; Yan et al., 2022; Bodnar et al., 2022; Luan et al., 2022). Interestingly, we note that our results for graph attention's

---

2. The means of the mixture of Gaussians are $\pm\boldsymbol{\mu}$. Our results can be easily generalized to general means. The current setting makes our analysis simpler without loss of generality.

behavior over the CSBM data model do not depend on whether the graph is homophilous or heterophilous.

Given node representations $\boldsymbol{h}_i \in \mathbb{R}^{F'}$ for $i \in [n]$, a (spatial) graph convolution for node i has output

$$\boldsymbol{h}'_i = \sum_{j \in [n]} \mathbf{A}_{ij} c_{ij} \mathbf{W} \boldsymbol{h}_j, \quad c_{ij} = \left(\sum_{\ell \in [n]} \mathbf{A}_{i\ell} \right)^{-1},$$

where $\mathbf{W} \in \mathbb{R}^{F \times F'}$ is a learnable matrix. Throughout this paper, we will refer to this operation as *simple graph convolution* or *standard graph convolution*. Our definition of graph convolution is essentially the mean aggregation step in a general GNN layer (Hamilton et al., 2017). The normalization constant c_{ij} in our definition is closely related to the symmetric normalization $c_{ij} = (\sum_{\ell \in [n]} \mathbf{A}_{i\ell})^{-1/2}(\sum_{\ell \in [n]} \mathbf{A}_{j\ell})^{-1/2}$ used in the original Graph Convolutional Network (GCN) introduced by Kipf and Welling (2017). Our definition does not affect the discussions or implications we have for GCN with symmetric normalization. More broadly, there are other forms of graph convolution in the literature (Bronstein et al., 2021; Defferrard et al., 2016; Levie et al., 2018), which we do not compare within this work.

A *single-head* graph attention applies some weight function on the edges based on their node features (or a mapping thereof). Given two representations $\boldsymbol{h}_i, \boldsymbol{h}_j \in \mathbb{R}^{F'}$ for two nodes $i, j \in [n]$, the *attention model/mechanism* is defined as the mapping

$$\Psi(\boldsymbol{h}_i, \boldsymbol{h}_j) \stackrel{\text{def}}{=} \alpha(\mathbf{W}\boldsymbol{h}_i, \mathbf{W}\boldsymbol{h}_j)$$

where $\alpha : \mathbb{R}^F \times \mathbb{R}^F \to \mathbb{R}$ and $\mathbf{W} \in \mathbb{R}^{F \times F'}$ is a learnable matrix. The *attention coefficient* for a node i and its neighbor j is defined as

$$\gamma_{ij} \stackrel{\text{def}}{=} \frac{\exp(\Psi(\boldsymbol{h}_i, \boldsymbol{h}_j))}{\sum_{\ell \in N_i} \exp(\Psi(\boldsymbol{h}_i, \boldsymbol{h}_\ell))}, \tag{1}$$

where N_i is the set of neighbors of node i that also includes node i itself. Let f be some element-wise activation function (which is usually nonlinear), the graph attention convolution output for a node $i \in [n]$ is given by

$$\begin{aligned} \boldsymbol{h}'_i &= \sum_{j \in [n]} \mathbf{A}_{ij} \gamma_{ij} \mathbf{W} \boldsymbol{h}_j, \\ \tilde{\boldsymbol{h}}_i &= f(\boldsymbol{h}'_i). \end{aligned} \tag{2}$$

A *multi-head* graph attention (Veličković et al., 2018) uses $K \in \mathbb{N}$ weight matrices $\mathbf{W}^1$, ..., $\mathbf{W}^K \in \mathbb{R}^{F \times F'}$ and averages their individual (single-head) outputs. We consider the most simplified case of a single graph attention layer (i.e., $F' = d$ and $F = 1$) where α is realized by an MLP using the LeakyRelu activation function. The LeakyRelu activation function is defined as $\text{LeakyRelu}(x) = x$ if $x \geq 0$ and $\text{LeakyRelu}(x) = \beta x$ for some constant $\beta \in [0, 1)$ if $x < 0$.

The CSBM model induces node features $\mathbf{X}$ which are correlated through the graph $G = ([n], E)$, represented by an adjacency matrix $\mathbf{A}$. A natural requirement of attention architectures is to maintain important edges in the graph and ignore unimportant edges. For example, important edges could be the set of intra-class edges and unimportant edges

could be the set of inter-class edges. In this case, if graph attention mains all intra-class edges and ignores all inter-class edges, then a node from a class will be connected only to nodes from its own class. More specifically, a node v will be connected to neighbor nodes whose associated node features come from the *same distribution* as node features of v. Given two sets A and B, we denote $A \times B \stackrel{\text{def}}{=} \{(i,j) : i \in A, j \in B\}$ and $A^2 \stackrel{\text{def}}{=} A \times A$.

3. Separation of edges and its implications for graph attention's ability for perfect node classification

In this section, we study the ability of graph attention to separate important edges from unimportant edges. In addition, we study the implications of graph attention's behavior for node classification. In particular, we are interested in whether or not the graph attention convolution is able to perfectly classify the nodes in the graph. Depending on the parameter regimes of the CSBM, we separate the results into two parts. In Section 3.1, we study graph attention's behavior when the distance between the means of the node features is large. We construct a specific attention architecture and demonstrate its ability to separate intra-class edges from inter-class edges. Consequently, we show that the corresponding graph attention convolution is able to perfectly classify the nodes. In Section 3.2, we study graph attention's behavior when the distance between the means of the node features is small. We show that with high probability no attention architecture is able to separate intra-class edges from inter-class edges. Consequently, we conjecture that no graph attention is able to perfectly classify the nodes. We provide evidence of this conjecture in Section 3.2.1, where we assume the existence of a strong attention mechanism that is able to perfectly classify the edges independently from node features. We show that using this "idealistic" attention mechanism still fails to (almost) perfectly classify the nodes, provided that the distance between the means of the node features is sufficiently small.

The two-parameter regimes of the CSBM that we consider in Section 3.1 and Section 3.2 are as follows. The first ("easy regime") is where $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$, and the second ("hard regime") is where $\|\boldsymbol{\mu}\| = \kappa\sigma$ for some $0 < \kappa \leq O(\sqrt{\log n})$. We start by defining edge separability and node separability.

Definition 1 (Edge separability)**.** *Given an attention model Ψ, we say that the model separates the edges, if the outputs satisfy*

$$\text{sign}(\Psi(\mathbf{X}_i, \mathbf{X}_j)) = -\text{sign}(\Psi(\mathbf{X}_k, \mathbf{X}_\ell))$$

whenever (i,j) is an intra-class edge, i.e. $(i,j) \in (C_1^2 \cup C_0^2) \cap E$, and (k,ℓ) is an inter-class edge, i.e. $(k,\ell) \in E \setminus (C_1^2 \cup C_0^2)$.

Definition 2 (Node separability)**.** *Given a classification model which outputs a scalar representation h_i' for node i, we say that the model separates the nodes if $h_i' > 0$ whenever $i \in C_1$ and $h_i' < 0$ whenever $i \in C_0$.*

Some of our results in this section rely on a mild assumption that lower bounds the sparsity of the graph generated by the CSBM model. This assumption says that the expected degree of a node in the graph is larger than $\log^2 n$. It is required to obtain a concentration of node degrees property used in the proofs. While this assumption may limit a direct

application of our results to very sparse graphs, for example, graphs whose node degrees are bounded by a constant, it is mainly an artifact of the proof techniques that we use. We expect that similar results still hold for sparser graphs, but to rigorously remove the assumption on graph density one may have to adopt a different proof technique.

Assumption 1. $p, q = \Omega(\log^2 n / n)$.

3.1 "Easy Regime"

In this regime $\left(\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})\right)$ we show that a two-layer MLP attention is able to separate the edges with high probability. Consequently, we show that the corresponding graph attention convolution is able to separate the nodes with high probability. At a high level, we transform the problem of classifying an edge $(i, j) \in E$ into the problem of classifying a point $[\tilde{\boldsymbol{w}}^T \mathbf{X}_i, \tilde{\boldsymbol{w}}^T \mathbf{X}_j]$ in $\mathbb{R}^2$, where $\tilde{\boldsymbol{w}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ is a unit vector that maximizes the total pairwise distances among the four means given below. When we consider the set of points $[\tilde{\boldsymbol{w}}^T \mathbf{X}_i, \tilde{\boldsymbol{w}}^T \mathbf{X}_j]$ for all $(i, j) \in E$, we can think of each point as a two-dimensional Gaussian vector whose mean is one of the following: $[\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}, \tilde{\boldsymbol{w}}^T \boldsymbol{\mu}]$, $[-\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}, \tilde{\boldsymbol{w}}^T \boldsymbol{\mu}]$, $[\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}, -\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}]$, $[-\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}, -\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}]$. The set of intra-class edges corresponds to the set of bivariate Gaussian vectors whose mean is either $[\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}, \tilde{\boldsymbol{w}}^T \boldsymbol{\mu}]$ or $[-\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}, -\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}]$, while the set of inter-class edges corresponds to the set of bivariate Gaussian vectors whose mean is either $[-\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}, \tilde{\boldsymbol{w}}^T \boldsymbol{\mu}]$ or $[\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}, -\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}]$. Therefore, in order to correctly classify the edges, we need to correctly classify the data corresponding to means $[\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}, \tilde{\boldsymbol{w}}^T \boldsymbol{\mu}]$ and $[-\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}, -\tilde{\boldsymbol{w}}^T \boldsymbol{\mu}]$ as 1, and classify the data corresponding to the other means as -1. This problem is known in the literature as the *"XOR problem"* (Minsky and Papert, 1969). To achieve this we consider a two-layer MLP architecture Ψ which separates the first and the third quadrants of the two-dimensional space from the second and the fourth quadrants. In particular, we consider the following specification of $\Psi(\mathbf{X}_i, \mathbf{X}_j)$,

$$\Psi(\mathbf{X}_i, \mathbf{X}_j) \stackrel{\text{def}}{=} \boldsymbol{r}^T \text{LeakyRelu} \left(\mathbf{S} \begin{bmatrix} \tilde{\boldsymbol{w}}^T \mathbf{X}_i \\ \tilde{\boldsymbol{w}}^T \mathbf{X}_j \end{bmatrix} \right), \tag{3}$$

where

$$\tilde{\boldsymbol{w}} \stackrel{\text{def}}{=} \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}, \quad \mathbf{S} \stackrel{\text{def}}{=} \begin{bmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad \boldsymbol{r} \stackrel{\text{def}}{=} R \cdot \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}, \tag{4}$$

where $R > 0$ is an arbitrary scaling parameter. The particular function Ψ has been chosen such that it is able to classify the means of the XOR problem correctly, that is,

$$\text{sign}(\Psi(\mathbf{E}[\mathbf{X}_i], \mathbf{E}[\mathbf{X}_j])) = \begin{cases} 1, & \text{if } (i, j) \text{ is an intra-class edge,} \\ -1, & \text{if } (i, j) \text{ is an inter-class edge.} \end{cases}$$

To see this, we can take look at the following simplified expression of Ψ, which is obtained by substituting the specifications of $\mathbf{S}$ and $\boldsymbol{r}$ from (4) to (3),

$$\Psi(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} -2R(1-\beta)\tilde{\boldsymbol{w}}^T \mathbf{X}_i, & \text{if } \tilde{\boldsymbol{w}}^T \mathbf{X}_j \leq -\left|\tilde{\boldsymbol{w}}^T \mathbf{X}_i\right|, \\ 2R(1-\beta)\text{sign}(\tilde{\boldsymbol{w}}^T \mathbf{X}_i)\tilde{\boldsymbol{w}}^T \mathbf{X}_j, & \text{if } -\left|\tilde{\boldsymbol{w}}^T \mathbf{X}_i\right| < \tilde{\boldsymbol{w}}^T \mathbf{X}_j < \left|\tilde{\boldsymbol{w}}^T \mathbf{X}_i\right|, \\ 2R(1-\beta)\tilde{\boldsymbol{w}}^T \mathbf{X}_i, & \text{if } \tilde{\boldsymbol{w}}^T \mathbf{X}_j \geq \left|\tilde{\boldsymbol{w}}^T \mathbf{X}_i\right|. \end{cases} \tag{5}$$

Then, by substituting $\tilde{\boldsymbol{w}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ into (5), one easily verifies that

$$\Psi(\mathbf{E}[\mathbf{X}_i], \mathbf{E}[\mathbf{X}_j]) = \left\{ \begin{array}{ll} 2R(1-\beta)\|\boldsymbol{\mu}\|, & \text{if } (i,j) \text{ is an intra-class edge,} \\ -2R(1-\beta)\|\boldsymbol{\mu}\|, & \text{if } (i,j) \text{ is an inter-class edge.} \end{array} \right. \tag{6}$$

On the other hand, our assumption that $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ guarantees that the distance between the means of the XOR problem is much larger than the standard deviation of the Gaussians, and thus with high probability there is no overlap between the distributions. More specifically, one can show that with a high probability,

$$\Psi(\mathbf{X}_i, \mathbf{X}_j) = \left\{ \begin{array}{ll} 2R(1-\beta)\|\boldsymbol{\mu}\|(1 \pm o(1)), & \text{if } (i,j) \text{ is an intra-class edge,} \\ -2R(1-\beta)\|\boldsymbol{\mu}\|(1 \pm o(1)), & \text{if } (i,j) \text{ is an inter-class edge.} \end{array} \right. \tag{7}$$

This property guarantees that with high probability,

$$\text{sign}(\Psi(\mathbf{X}_i, \mathbf{X}_j)) = \text{sign}(\Psi(\mathbf{E}[\mathbf{X}_i], \mathbf{E}[\mathbf{X}_j])), \text{ for all } (i,j) \in E,$$

which implies perfect separability of the edges. We state this result below in Theorem 3 and provide detailed arguments in the appendix.

Theorem 3. *Suppose that $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$. Then with probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, the two-layer MLP attention architecture Ψ given in (3) and (4) separates intra-class edges from inter-class edges.*

Our analysis of edge separability by the attention architecture Ψ has two important implications. First, edge separability by Ψ implies a nice concentration result for the attention coefficients γ_{ij} defined in (1). In particular, one can show that the attention coefficients have the desirable property to maintain important edges and drop unimportant edges. Second, the nice distinguishing power of the attention coefficients in turn implies the exact recoverability of the node memberships using the graph attention convolution. We state these two results below in Corollary 4 and Corollary 5, respectively. Denote

$$\Psi' \stackrel{\text{def}}{=} (\mathbf{1}_{p \geq q} - \mathbf{1}_{p < q}) \cdot \Psi,$$

that is, $\Psi' = \Psi$ if $p \geq q$ and $\Psi' = -\Psi$ if $p < q$, where Ψ is the two-layer MLP attention architecture given in (3) and (4). As it will become clear later, the attention architecture Ψ' allows attention coefficients to assign more weights to either intra-class or inter-class edges, depending on if $p \geq q$ or $p < q$, and this will help us obtain a perfect separation of the nodes.

Corollary 4. *Suppose that $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ and that Assumption 1 holds. Then with probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, the attention architecture Ψ' yields attention coefficients γ_{ij} such that*

1. *If $p \geq q$, then $\gamma_{ij} = \frac{2}{np}(1 \pm o(1))$ if (i,j) is an intra-class edge and $\gamma_{ij} = o(\frac{1}{n(p+q)})$ otherwise;*

2. *If $p < q$, then $\gamma_{ij} = \frac{2}{nq}(1 \pm o(1))$ if (i,j) is an inter-class edge and $\gamma_{ij} = o(\frac{1}{n(p+q)})$ otherwise.*

Corollary 5. *Suppose that $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ and that Assumption 1 holds. Then with probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, using the attention architecture Ψ' with the graph attention convolution given in (2), where f is set to be the identify function, the model separates the nodes.*

Corollary 4 shows the desired behavior of the attention mechanism, namely it is able to assign significantly large weights to important edges while it drops unimportant edges. When $p \geq q$, the attention mechanism maintains intra-class edges and essentially ignores all inter-class edges; when $p < q$, it maintains inter-class edges and essentially ignores all intra-class edges. We now explain the high-level idea of the proof and leave detailed arguments to Appendix A.2. Corollary 4 follows from the concentration of $\Psi(\mathbf{X}_i, \mathbf{X}_j)$ around $\Psi(\mathbf{E}[\mathbf{X}_i], \mathbf{E}[\mathbf{X}_j])$ given by (7). Assume for a moment that $p \geq q$ so that $\mathbf{1}_{p\geq q} - \mathbf{1}_{p<q} = 1$. Then $\Psi' = \Psi$, and (7) implies that the value of $\exp(\Psi'(\mathbf{X}_i, \mathbf{X}_j))$ when (i, j) is an intra-class edge is exponentially larger than the value of $\exp(\Psi'(\mathbf{X}_i, \mathbf{X}_j))$ when (i, j) is an inter-class edge. Therefore, by the definition of the attention coefficients in (1), the denominator of γ_{ij} is dominated by terms (i, k) where k is in the same class as i. Moreover, using concentration of node degrees guaranteed by Assumption 1, each node i is connected to $\frac{2}{np}(1 \pm o(1))$ many intra-class nodes. By appropriately setting the scaling parameter R in (4), the values of $\Psi'(\mathbf{X}_i, \mathbf{X}_k)$ for all intra-class edges (i, k) can be made within $o(1)$ multiplicative factor from each other. Therefore we get $\gamma_{ij} = \frac{2}{np}(1 \pm o(1))$ when (i, j) is an intra-class edge. A similar reasoning applies to inter-class edges and yields the vanishing value of γ_{ij} when (i, j) is an inter-class edge. Finally, the argument for $p < q$ follows analogously.

The concentration result of attention coefficients in Corollary 4 implies the node classification result in Corollary 5, which holds for any p, q satisfying Assumption 1. That is, even when the graph structure is noisy, e.g., when $p \approx q$, it is still possible to exactly recover the node class memberships. Essentially, by applying Corollary 4 and carrying out some straightforward algebraic simplifications, one can show that with high probability for all $i \in [n]$,

$$h'_i = \sum_{j \in N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j = \begin{cases} -(\mathbf{1}_{p\geq q} - \mathbf{1}_{p<q})\|\boldsymbol{\mu}\|(1 \pm o(1)), & \text{if } i \in C_0, \\ (\mathbf{1}_{p\geq q} - \mathbf{1}_{p<q})\|\boldsymbol{\mu}\|(1 \pm o(1)), & \text{if } i \in C_1, \end{cases}$$

and hence the model separates the nodes with high probability. We provide details in the appendix.

While Corollary 5 provides a positive result for graph attention, it can be shown that a simple linear classifier which does not use the graph at all achieves perfect node separability with high probability. In particular, the Bayes optimal classifier for the node features without the graph is able to separate the nodes with high probability. This means that in this regime, using the additional graph information is unnecessary, as it does not provide additional power compared to a simple linear classifier for the node classification task.

Lemma 6 (Section 6.4 in Anderson (2003)). *Let $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$. Then the optimal Bayes classifier for $\mathbf{X}$ is realized by the linear classifier*

$$h(\mathbf{X}_i) = \begin{cases} 0 & \text{if } \boldsymbol{\mu}^T \mathbf{X}_i \leq 0 \\ 1 & \text{if } \boldsymbol{\mu}^T \mathbf{X}_i > 0 \end{cases}. \tag{8}$$

Proposition 7. *Suppose $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$. Then with probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, the linear classifier given in (8) separates the nodes.*

The proof of Proposition 7 is elementary. To see the claim one may show that the probability that the classifier in (8) misclassifies a node $i \in [n]$ is $o(1)$. To do this, let us fix $i \in [n]$ and write $\mathbf{X}_i = (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma\boldsymbol{g}_i$ where $\boldsymbol{g}_i \sim N(0, \mathbf{I})$. Assume for a moment $\epsilon_i = 0$. Then the probability of misclassification is

$$\mathbf{Pr}\left[\boldsymbol{\mu}^T\mathbf{X}_i > 0\right] = \mathbf{Pr}\left[\frac{\boldsymbol{\mu}^T\boldsymbol{g}_i}{\|\boldsymbol{\mu}\|} > \frac{\|\boldsymbol{\mu}\|}{\sigma}\right] = 1 - \Phi\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$ and the last equality follows from the fact that $\frac{\boldsymbol{\mu}^T\boldsymbol{g}_i}{\|\boldsymbol{\mu}\|} \sim N(0, 1)$. The assumption that $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ implies $\|\boldsymbol{\mu}\| \geq \sigma\sqrt{2\log n}$ for large enough n. Therefore, using standard tail bounds for normal distribution (Vershynin, 2018) we have that

$$1 - \Phi\left(\frac{\|\boldsymbol{\mu}\|}{\sigma}\right) \leq \frac{\sigma}{\sqrt{2\pi}\|\boldsymbol{\mu}\|}\exp\left(-\frac{\|\boldsymbol{\mu}\|^2}{2\sigma^2}\right) \leq \frac{n^{-1}}{\sqrt{4\pi\log n}}.$$

This means that the probability that there exists $i \in C_0$ which is misclassified is at most $\frac{1}{2\sqrt{4\pi\log n}} = o(1)$. A similar argument can be applied to the case where $\epsilon_i = 1$, and an application of a union bound on the events that there is $i \in [n]$ which is misclassified finishes the proof of Proposition 7.

3.2 "Hard Regime"

In this regime ($\|\boldsymbol{\mu}\| = \kappa\sigma$ for $\kappa \leq O(\sqrt{\log n})$), we show that *every* attention architecture Ψ fails to separate the edges if $\kappa < \sqrt{2\log n}$, and we conjecture that no graph attention convolution is able to separate the nodes. The conjecture is based on our node separability result in Section 3.2.1 which says that, even if we assume that there is an attention architecture which is able to separate the edges independently from node features, the corresponding graph attention convolution still fails to (almost) perfectly classify the nodes with high probability.

The goal of the attention mechanism is to decide whether an edge (i, j) is an inter-class edge or an intra-class edge based on the node feature vectors $\mathbf{X}_i$ and $\mathbf{X}_j$. Let $\mathbf{X}'_{ij}$ denote the vector obtained from concatenating $\mathbf{X}_i$ and $\mathbf{X}_j$, that is,

$$\mathbf{X}'_{ij} \stackrel{\text{def}}{=} \begin{pmatrix} \mathbf{X}_i \\ \mathbf{X}_j \end{pmatrix}. \tag{9}$$

We would like to analyze every classifier h' which takes as input $\mathbf{X}'_{ij}$ and tries to separate inter-class edges and intra-class edges. An ideal classifier would have the property

$$y = h'(\mathbf{X}'_{ij}) = \begin{cases} 0, & \text{if } (i, j) \text{ is an inter-class edge,} \\ 1, & \text{if } (i, j) \text{ is an intra-class edge.} \end{cases} \tag{10}$$

To understand the limitations of all such classifiers in this regime, it suffices to consider the Bayes optimal classifier for this data model, whose probability of misclassifying an arbitrary edge lower bounds that of every attention architecture which takes as input $(\mathbf{X}_i, \mathbf{X}_j)$.

Consequently, by deriving a misclassification rate for the Bayes classifier, we obtain a lower bound on the misclassification rate for every attention mechanism Ψ for classifying intra-class and inter-class edges. The following Lemma 8 describes the Bayes classifier for this classification task.

Lemma 8. *Let* $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ *and let* $\mathbf{X}'_{ij}$ *be defined as in (9). The Bayes optimal classifier for* $\mathbf{X}'_{ij}$ *is realized by the following function,*

$$h^*(\boldsymbol{x}) = \begin{cases} 0, & \text{if } p \cosh\left(\frac{\boldsymbol{x}^T \boldsymbol{\mu}'}{\sigma^2}\right) \leq q \cosh\left(\frac{\boldsymbol{x}^T \boldsymbol{\nu}'}{\sigma^2}\right), \\ 1, & \text{otherwise}, \end{cases} \tag{11}$$

where $\boldsymbol{\mu}' \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix}$ *and* $\boldsymbol{\nu}' \stackrel{\text{def}}{=} \begin{pmatrix} \boldsymbol{\mu} \\ -\boldsymbol{\mu} \end{pmatrix}$.

Using Lemma 8, we can lower bound the rate of misclassification of edges that every attention mechanism Ψ exhibits. Below we define $\Phi_c \stackrel{\text{def}}{=} 1 - \Phi$, where Φ is the cumulative distribution function of $N(0, 1)$.

Theorem 9. *Suppose* $\|\boldsymbol{\mu}\| = \kappa \sigma$ *for some* $\kappa > 0$ *and let* Ψ *be any attention mechanism. Then,*

1. *With probability at least* $1 - o(1)$, Ψ *fails to correctly classify at least* $2 \cdot \Phi_c(\kappa)^2$ *fraction of inter-class edges;*

2. *For any* $K > 0$ *if* $q > \frac{K \log^2 n}{n \Phi_c(\kappa)^2}$, *then with probability at least* $1 - O(n^{-8K\Phi_c(\kappa)^2 \log n})$, Ψ *misclassify at least one inter-class edge.*

Part 1 of Theorem 9 implies that if $\|\boldsymbol{\mu}\|$ is *linear* in the standard deviation σ, that is if $\kappa = O(1)$, then with overwhelming probability the attention mechanism fails to distinguish a constant fraction of inter-class edges from intra-class edges. Furthermore, part 2 of Theorem 9 characterizes a regime for the inter-class edge probability q where the attention mechanism fails to distinguish at least one inter-class edge. It provides a lower bound on q in terms of the scale at which the distance between the means grows compared to the standard deviation σ. This aligns with the intuition that as we increase the distance between the means, it gets easier for the attention mechanism to correctly distinguish inter-class and intra-class edges. However, if q is also increased with the right proportion, in other words, if the noise in the graph is increased, then the attention mechanism would still fail to correctly distinguish at least one inter-class edge. For instance, for $\kappa = \sqrt{2 \log \log n}$ and $K = \log^2 n$, we get that if $q > \Omega(\frac{\log^{6+o(1)} n}{n})$, then with probability at least $1 - o(1)$, Ψ misclassifies at least an inter-class edge.

The proof of Theorem 9 relies on analyzing the behavior of the Bayes optimal classifier in (11). We compute an upper bound on the probability with which the optimal classifier correctly classifies an arbitrary inter-class edge. Then the proof of part 1 of Theorem 9 follows from a concentration argument for the fraction of inter-class edges that are misclassified by the optimal classifier. For part 2, we use a similar concentration argument to choose a suitable threshold for q that forces the optimal classifier to fail on at least one inter-class edge. We provide formal arguments in the appendix.

As a motivating example of how the attention mechanism would fail and what exactly the attention coefficients would behave in this regime, we focus on one of the most popular attention architecture (Veličković et al., 2018), where α is a single-layer neural network parameterized by $(\boldsymbol{w}, \boldsymbol{a}, b) \in \mathbb{R}^d \times \mathbb{R}^2 \times \mathbb{R}$ with LeakyRelu activation function. Namely, the attention coefficients are defined by

$$\gamma_{ij} \stackrel{\text{def}}{=} \frac{\exp\left(\text{LeakyRelu}\left(\boldsymbol{a}^T \begin{bmatrix} \boldsymbol{w}^T \mathbf{X}_i \\ \boldsymbol{w}^T \mathbf{X}_j \end{bmatrix} + b\right)\right)}{\sum_{\ell \in N_i} \exp\left(\text{LeakyRelu}\left(\boldsymbol{a}^T \begin{bmatrix} \boldsymbol{w}^T \mathbf{X}_i \\ \boldsymbol{w}^T \mathbf{X}_\ell \end{bmatrix} + b\right)\right)}. \tag{12}$$

We show that, as a consequence of the inability of the attention mechanism to distinguish intra-class and inter-class edges, with overwhelming probability most of the attention coefficients γ_{ij} given by (12) are going to be $\Theta(1/|N_i|)$. In particular, Theorem 10 says that for the vast majority of nodes in the graph, the attention coefficients on most edges are uniform irrespective of whether the edge is inter-class or intra-class. As a result, this means that the attention mechanism is unable to assign higher weights to important edges and lower weights to unimportant edges.

Theorem 10. *Assume that $\|\boldsymbol{\mu}\| \leq K\sigma$ and $\sigma \leq K'$ for some absolute constants K and K'. Moreover, assume that the parameters $(\boldsymbol{w}, \boldsymbol{a}, b) \in \mathbb{R}^d \times \mathbb{R}^2 \times \mathbb{R}$ are bounded. Then, with probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, there exists a subset $\mathcal{A} \subseteq [n]$ with cardinality at least $n(1 - o(1))$ such that for all $i \in \mathcal{A}$ the following hold:*

1. *There is a subset $J_{i,0} \subseteq N_i \cap C_0$ with cardinality at least $\frac{9}{10}|N_i \cap C_0|$, such that $\gamma_{ij} = \Theta(1/|N_i|)$ for all $j \in J_{i,0}$.*

2. *There is a subset $J_{i,1} \subseteq N_i \cap C_1$ with cardinality at least $\frac{9}{10}|N_i \cap C_1|$, such that $\gamma_{ij} = \Theta(1/|N_i|)$ for all $j \in J_{i,1}$.*

Theorem 10 is proved by carefully computing the numerator and the denominator in (12). In this regime, $\|\boldsymbol{\mu}\|$ is not much larger than σ, that is, the signal does not dominate noise, so the numerator in (12) is not indicative of the class memberships of nodes i, j but rather acts like Gaussian noise. On the other hand, denote the denominator in (12) by δ_i and observe that it is the same for all $\gamma_{i\ell}$ where $\ell \in N_i$. Using concentration arguments about $\{\boldsymbol{w}^T \mathbf{X}_\ell\}_{\ell \in [n]}$ yields $\gamma_{ij} = \Theta(1/\delta_i)$ and $\delta_i = \Theta(|N_i|)$ finishes up the proof. We provide details in the appendix.

Compared to the easy regime, it is difficult to obtain a separation result for the nodes without additional assumptions. In the easy regime, the distance between the means was much larger than the standard deviation, which made the "signal" (the expectation of the convolved data) dominate the "noise" (i.e., the variance of the convolved data). In the hard regime the "noise" dominates the "signal". Thus, we conjecture the following.

Conjecture 11. *There is an absolute constant $M > 0$ such that, whenever $\|\boldsymbol{\mu}\| \leq M \cdot \sigma\sqrt{\frac{\log n}{n(p+q)}}(1 - \max(p, q)) \cdot \frac{p+q}{|p-q|}$, every graph attention model fails to perfectly classify the nodes with high probability.*

13

The above conjecture means that in the hard regime, the performance of the graph attention model depends on q as opposed to the easy regime, where in Corollary 5 we show that it doesn't. This property is verified by our synthetic experiments in Section 5. The quantity $\sigma\sqrt{\frac{\log n}{n(p+q)}}(1 - \max(p,q))$ in the threshold comes from our conjecture that the expected maximum "noise" of the graph attention convolved data over the nodes is at least $c\sigma\sqrt{\frac{\log n}{n(p+q)}}(1 - \max(p,q))$ for some constant $c > 0$. The quantity $\frac{p+q}{|p-q|}$ in the threshold comes from our conjecture that the distance between the means (i.e. "signal") of the graph attention convolved data is reduced to at most $|p-q|/(p+q)$ of the original distance. Proving Conjecture 11 would require delicate treatment of the correlations between the attention coefficients γ_{ij} and the node features $\mathbf{X}_i$ for $i \in [n]$.

3.2.1 Are good attention coefficients helpful in the "hard regime"?

In this subsection we are interested in understanding the implications of edge separability on node separability in the hard regime and when Ψ is restricted to a specific class of functions. In particular, we show that Conjecture 11 is true under an additional assumption that Ψ does not depend on the node features. In addition, we show that even if we were allowed to use an "extremely good" attention function $\tilde{\Psi}$ which separates the edges with an arbitrarily large margin, with high probability the graph attention convolution (2) will still misclassify at least one node as long as $\|\boldsymbol{\mu}\|/\sigma$ is sufficiently small.

We consider the class of functions $\tilde{\Psi}$ which can be expressed in the following form:

$$\tilde{\Psi}(i,j) = \begin{cases} \operatorname{sign}(p-q)t, & \text{if } (i,j) \text{ is an intra-class edge,} \\ -\operatorname{sign}(p-q)t, & \text{if } (i,j) \text{ is an inter-class edge,} \end{cases} \tag{13}$$

for some $t \geq 0$. The particular class of functions in (13) is motivated by the property of the ideal edge classifier in (10) and the behavior of Ψ in (6) when it is applied to the means of the Gaussians. There are a few possible ways to obtain a function $\tilde{\Psi}$ which satisfies (13). For example, in the presence of good edge features which reflect the class memberships of the edges, we can make $\tilde{\Psi}$ take as input the edge features. Moreover, if $|\sqrt{p} - \sqrt{q}| > \sqrt{2\log n/n}$, one such $\tilde{\Psi}$ may be easily realized from the eigenvectors of the graph adjacency matrix. By the exact spectral recovery result in Lemma 12, we know that there exists a classifier $\hat{\tau}$ which separates the nodes. Therefore, we can set $\tilde{\Psi}(i,j) = \operatorname{sign}(p-q)t$ if $\hat{\tau}(i) = \hat{\tau}(j)$ and $\tilde{\Psi}(i,j) = -\operatorname{sign}(p-q)t$ otherwise.

Lemma 12 (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 $1 - o(1)$.*

Theorem 13. *Suppose that p, q satisfy Assumption 1 and that p, q are bounded away from 1. For every $\epsilon > 0$, there are absolute constants $M, M' = O(\sqrt{\epsilon})$ such that, with probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, using the graph attention convolution in (2) and the attention architecture $\tilde{\Psi}$ in (13), the model misclassifies at least $\Omega(n^{1-\epsilon})$ nodes for any $\boldsymbol{w}$ such that $\|\boldsymbol{w}\| = 1$, if*

1. $t = O(1)$ and $\|\boldsymbol{\mu}\| \leq M\sigma\sqrt{\frac{\log n}{n(p+q)}}(1 - \max(p,q))\frac{p+q}{|p-q|}$;

2. $t = \omega(1)$ and $\|\boldsymbol{\mu}\| \leq M'\sigma\sqrt{\frac{\log n}{n(p+q)}}(1 - \max(p,q))$.

Theorem 13 warrants some discussions. We start with the role of t in the attention function (13). One may think of t as the multiplicative margin of separation for intra-class and intra-class edges. When $t = O(1)$, the margin of separation is at most a constant. This includes the special case when $\tilde{\Psi}(i,j) = 0$ for all $(i,j) \in E$, i.e, the margin of separation is 0. In this case, the graph attention convolution in (2) reduces to the standard graph convolution with uniform averaging among the neighbors. Therefore, part 1 of Theorem 13 also applies to the standard graph convolution. On the other hand, when $t = \omega(1)$, the margin of separation is not only bounded away from 0 but also it grows with n.

Next, we discuss the additional assumption that p, q are bounded away from 1. This assumption is used to obtain a concentration result required for the proof of Theorem 13. It is also intuitive in the following sense. If both p and q are arbitrarily close to 1, then after the convolution the convolved node feature vectors collapse to approximately a single point, and thus this becomes a trivial case where no classifier is able to separate the nodes; on the other hand, if p is arbitrarily close to 1 and q is very small, then after the convolution the convolved node feature vectors collapse to approximately one of two points according to which class the node comes from, and in this case the nodes can be easily separated by a linear classifier.

We now focus on the threshold for $\|\boldsymbol{\mu}\|$ under which the model is going to misclassify at least one node with high probability. In part 1 of Theorem 13, $t = O(1)$, i.e., the attention mechanism $\tilde{\Psi}$ is either unable to separate the edges or unable to separate the edges with a large enough margin. In this case, one can show that all attention coefficients are $\Theta(\frac{1}{n(p+q)})$. Consequently, the quantity $|p - q|$ appears in denominator of the threshold for $\|\boldsymbol{\mu}\|$ in part 1 of Theorem 13. Because of that, if p and q are arbitrarily close, then the model is not able to separate the nodes irrespective of how large $\|\boldsymbol{\mu}\|$ is. For example, treating $1 - \max(p,q)$ as a constant since p and q are bounded away from 1 by assumption, we have that

$$|p - q| = o\left(\sqrt{\frac{p+q}{n}}\right) \text{ implies } M\sigma\sqrt{\frac{\log n}{n(p+q)}}(1 - \max(p,q))\frac{p+q}{|p-q|} = \omega(\sigma\sqrt{\log n}).$$

This means that if p and q are close enough, every attention function $\tilde{\Psi}$ in the form of (13) and $t = O(1)$ cannot help classify all nodes correctly even if $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$. On the contrary, recall that in the easy regime where $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$, the attention mechanism given in (3) and (4) helps separate the nodes with high probability. This illustrates the limitation of every attention mechanism in the form of (13) which have an insignificant margin of separation. According to Theorem 10, the vast majority of attention coefficients are uniform, and thus in Conjecture 11 we expect that graph attention in general shares similar limitations in the hard regime.

In part 2 of Proposition 13, $t = \omega(1)$, i.e., the attention mechanism $\tilde{\Psi}$ separates the edges with a large margin. In this case, one can show that the attention coefficients on important edges (e.g. intra-class edges) are exponentially larger than those on unimportant edges (e.g. inter-class edges). Consequently, the factor $(p+q)/|p-q|$ no longer appears in the threshold for $\|\boldsymbol{\mu}\|$ in part 2 of Theorem 13. However, at the same time, the threshold also implies that, even when we have a perfect attention mechanism that is able to separate

the edges with a large margin, as long as $\|\boldsymbol{\mu}\|/\sigma$ is small enough, then the model is going to misclassify at least one node with high probability.

Finally, notice that in Theorem 13 the parameter ϵ captures a natural tradeoff between the threshold for $\|\boldsymbol{\mu}\|$ and the lower bound on the number of misclassified nodes. Namely, the smaller the ϵ is, the smaller the threshold for $\|\boldsymbol{\mu}\|$ becomes, and hence the less signal there is in the node features, the more mistakes the model is going to make. This is precisely demonstrated by the scaling of $M, M' = O(\sqrt{\epsilon})$ and misclassification bound $\Omega(n^{1-\epsilon})$ with respect to ϵ. We leave the proof of Theorem 13 to the appendix.

4. Robustness to structural noise and its implications beyond perfect node classification

In this section, we provide a positive result on the capacity of graph attention convolution for node classification beyond the perfect classification regime. In particular, we show that independent of the parameters of CSBM, i.e., independent of p, q, $\boldsymbol{\mu}$ and σ, the two-layer MLP attention architecture Ψ from Section 3.1 is able to achieve the classification performance obtainable by the Bayes optimal classifier for node features. This is proved by showing that there is a parameter setting for Ψ where the attention coefficient on self-loops can be made exponentially large. Consequently, the corresponding graph attention convolution behaves like a linear function of node features. We provide two corollaries of this result. The first corollary provides a ranking of graph attention convolution, simple graph convolution, and linear function in terms of their ability to classify nodes in CSBM. More specifically, by noticing that the simple graph convolution is also realized by a specific parameter setting of the attention architecture Ψ, our result implies that the performance of graph attention convolution for node classification in CSBM is lower bounded by the best possible performance between a linear classifier and simple graph convolution. In particular, graph attention is strictly more powerful than simple graph convolution when the graph is noisy (e.g. when $p \approx q$), and it is strictly more powerful than a linear classifier when the graph is less noisy (e.g. when p and q are significantly different). The second corollary provides perfect classification, almost perfect classification, and partial classification results for graph attention convolution. It follows immediately from the reduction of graph attention convolution to a linear function under the specification of Ψ that we will discuss. In what follows we start with high-level ideas, then we present formal statements of the results, and we discuss the implications in detail.

Recall the two-layer MLP attention architecture Ψ from (3) and (4) is equivalently written in (5) as

$$\Psi(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} -2R(1-\beta)\tilde{\boldsymbol{w}}^T\mathbf{X}_i, & \text{if } \tilde{\boldsymbol{w}}^T\mathbf{X}_j \leq -\left|\tilde{\boldsymbol{w}}^T\mathbf{X}_i\right|, \\ 2R(1-\beta)\operatorname{sign}(\tilde{\boldsymbol{w}}^T\mathbf{X}_i)\tilde{\boldsymbol{w}}^T\mathbf{X}_j, & \text{if } -\left|\tilde{\boldsymbol{w}}^T\mathbf{X}_i\right| < \tilde{\boldsymbol{w}}^T\mathbf{X}_j < \left|\tilde{\boldsymbol{w}}^T\mathbf{X}_i\right|, \\ 2R(1-\beta)\tilde{\boldsymbol{w}}^T\mathbf{X}_i, & \text{if } \tilde{\boldsymbol{w}}^T\mathbf{X}_j \geq \left|\tilde{\boldsymbol{w}}^T\mathbf{X}_i\right|. \end{cases}$$

We make the following observations. Assuming that the scaling parameter $R > 0$,

- If $\tilde{\boldsymbol{w}}^T\mathbf{X}_i > 0$, then the function Ψ assigns more weight to an edge (i, j) such that $\tilde{\boldsymbol{w}}^T\mathbf{X}_j > 0$ than an edge (i, j') such that $\tilde{\boldsymbol{w}}^T\mathbf{X}_{j'} < 0$;

- If $\tilde{\boldsymbol{w}}^T\mathbf{X}_i < 0$, then the function Ψ assigns more weight to an edge (i, j) such that $\tilde{\boldsymbol{w}}^T\mathbf{X}_j < 0$ than an edge (i, j') such that $\tilde{\boldsymbol{w}}^T\mathbf{X}_{j'} > 0$;

- If $\tilde{\boldsymbol{w}}^T \mathbf{X}_i = 0$, then the function Ψ assigns uniform weight to every edge (i, j).

This means that the behavior of Ψ depends on which side of the hyperplane $\{\boldsymbol{x} : \tilde{\boldsymbol{w}}^T \boldsymbol{x} = 0\}$ that $\mathbf{X}_i$ comes from. In other words, for fixed $\mathbf{X}_j$, whether the attention function Ψ will up-weight or down-weight an edge (i, j) depends entirely on the classification of $\mathbf{X}_i$ based on the linear classifier $\tilde{\boldsymbol{w}}$. Moreover, note that the attention function value satisfies

$$2R(1 - \beta) \cdot \max\{-|\tilde{\boldsymbol{w}}^T \mathbf{X}_i|, -|\tilde{\boldsymbol{w}}^T \mathbf{X}_j|\} \le \Psi(\mathbf{X}_i, \mathbf{X}_j) \le 2R(1 - \beta) \cdot \min\{|\tilde{\boldsymbol{w}}^T \mathbf{X}_i|, |\tilde{\boldsymbol{w}}^T \mathbf{X}_j|\}.$$

Therefore, out of all unit norm vectors $\boldsymbol{w}$, our choice $\tilde{\boldsymbol{w}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ maximizes the range of values that Ψ can output. Recall from Lemma 6 that $\tilde{\boldsymbol{w}}$ also happens to characterize the Bayes optimal classifier for the node features. Finally, the attention function Ψ achieves minimum/maximum at self-attention, i.e.

$$\Psi(\mathbf{X}_i, \mathbf{X}_i) = 2R(1 - \beta)|\tilde{\boldsymbol{w}}^T \mathbf{X}_i| = \max_{j \in [n]} \Psi(\mathbf{X}_i, \mathbf{X}_j),$$

$$\Psi(\mathbf{X}_i, -\mathbf{X}_i) = -2R(1 - \beta)|\tilde{\boldsymbol{w}}^T \mathbf{X}_i| = \min_{j \in [n]} \Psi(\mathbf{X}_i, \mathbf{X}_j).$$

A consequence of these observations is that, by setting the scaling parameter R sufficiently large, one can make $\exp(\Psi(\mathbf{X}_i, \mathbf{X}_j))$ exponentially larger than $\exp(\Psi(\mathbf{X}_i, \mathbf{X}_{j'}))$ for any j, j' such that $\text{sign}(\tilde{\boldsymbol{w}}^T \mathbf{X}_j) = \text{sign}(\tilde{\boldsymbol{w}}^T \mathbf{X}_i)$ and $\text{sign}(\tilde{\boldsymbol{w}}^T \mathbf{X}_{j'}) \ne \text{sign}(\tilde{\boldsymbol{w}}^T \mathbf{X}_i)$. According to the definition of attention coefficients in (1), this means that the attention coefficients γ_{ij} where $\text{sign}(\tilde{\boldsymbol{w}}^T \mathbf{X}_j) = \text{sign}(\tilde{\boldsymbol{w}}^T \mathbf{X}_i)$ are going to be exponentially larger than the attention coefficients $\gamma_{ij'}$ where $\text{sign}(\tilde{\boldsymbol{w}}^T \mathbf{X}_{j'}) \ne \text{sign}(\tilde{\boldsymbol{w}}^T \mathbf{X}_i)$. Therefore, one could expect that in this case, if the linear classifier $\tilde{\boldsymbol{w}}$ correctly classifies $\mathbf{X}_i$ for some $i \in [n]$, e.g., for $i \in C_1$ this means that $\tilde{\boldsymbol{w}}^T \mathbf{X}_i > 0$, then graph attention convolution output $h_i' = \sum_{j \in N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j$ also satisfies $h_i' > 0$, due to sufficiently larger attention coefficients γ_{ij} for which $\tilde{\boldsymbol{w}}^T \mathbf{X}_j > 0$. We state the result below in Theorem 14 and leave detailed arguments in the appendix.

Theorem 14. *With probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, using the two-layer MLP attention architecture Ψ given in (3) and (4) with $R = \Omega(n \log^2 n / \sigma)$, the graph attention convolution output satisfies*

$$h_i' = \sum_{j \in N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j > 0 \text{ if and only if } \tilde{\boldsymbol{w}}^T \mathbf{X}_i > 0, \ \forall i \in [n],$$

$$h_i' = \sum_{j \in N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j < 0 \text{ if and only if } \tilde{\boldsymbol{w}}^T \mathbf{X}_i < 0, \ \forall i \in [n].$$

Theorem 14 means that there is a parameter setting for the attention architecture Ψ such that the performance of graph attention convolution matches with the performance of the Bayes optimal classifier for node features. This shows the ability of graph attention to "ignore" the graph structure, which can be beneficial when the graph is noisy. For example, if $p = q$, then it is easy to see that simple graph convolution completely mixes up the node features, making it not possible to achieve any meaningful node classification result. On the other hand, as long as there is some signal from the original node features, i.e. $\|\boldsymbol{\mu}\| > 0$, then graph attention will be able to pick that up and classify the nodes at least as good as the best classifier for the node features alone. It is also worth noting that by

setting $R = 0$, the attention function Ψ has a constant value, and hence graph attention convolution reduces to the standard graph convolution, which has been shown to be useful in the regime where the original node features are not very strong but the graph has a nice structure (Baranwal et al., 2021). For example, when there is a significant gap between p and q, $|p - q|/(p + q) = \Omega(1)$, then setting $R = 0$ could significantly improve the node separability threshold over the best linear classifier (Baranwal et al., 2021). This shows the robustness of graph attention against noise in one of the two sources of information, namely node features and edge connectivities. Unlike the Bayes optimal classifier for node features which is sensitive to feature noise or the simple graph convolution which is sensitive to structural noise, one can expect graph attention to work as long as at least one of the two sources of information has a good signal. Therefore, we obtain a ranking of node classification models among graph attention convolution, simple graph convolution, and a linear classifier. We state this below in Corollary 15.

Corollary 15. *The node classification performance obtainable by graph attention convolution is lower bounded by the best possible node classification performance between a simple graph convolution and a linear classifier.*

By a straightforward characterization of the performance of the linear classifier $\tilde{\boldsymbol{w}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$, we immediately obtain the following classification results in Corollary 16. Recall that we denoted Φ as the cumulative distribution function of the standard normal distribution. Write

$$\|\boldsymbol{\mu}\| = \kappa\sigma \ \text{ for some } \ \kappa > 0.$$

Corollary 16. *With probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim CSBM(n, p, q, \boldsymbol{\mu}, \sigma^2)$, using the two-layer MLP attention architecture Ψ given in (3) and (4) with $R = \Omega(n \log^2 n/\sigma)$, one has that*

- *(Perfect classification) If $\kappa \geq \sqrt{2 \log n}$ then all nodes are correctly classified;*

- *(Almost perfect classification) If $\kappa = \omega(1)$ then at least $1 - o(1)$ fraction of all nodes are correctly classified;*

- *(Partial classification) If $\kappa = O(1)$ then at least $\Phi(\kappa) - o(1)$ fraction of all nodes are correctly classified.*

Interestingly, we note that the perfect classification result in Corollary 16 is nearly equivalent (up to a small order in the threshold κ) to the perfect classification result in Corollary 5 from Section 3.1. They are obtained from different behaviors of the attention coefficients. This shows that there could be more than one type of "good" attention coefficients that are able to deliver good node classification performance.

5. Experiments

In this section, we demonstrate empirically our results on synthetic and real data. The parameters of the models that we experiment with are set by using an ansatz based on our theorems. The particular details are given in Section 5.1. For real datasets, we use the default splits which come from PyTorch Geometric (Fey and Lenssen, 2019) and OGB (Hu

et al., 2020). In all our experiments we use MLP-GAT, where the attention mechanism Ψ is set to be the two-layer network in (3) and (4) with $R = 1$. For synthetic experiments using CSBM with known p and q, we use the variant that takes p, q into account, $\Psi' = (\mathbf{1}_{p \geq q} - \mathbf{1}_{p < q})\Psi$. In Figure 2b and Figure 3b we additionally consider the original GAT architecture of Veličković et al. (2018) to demonstrate Theorem 10.

5.1 Ansatz for GAT, MLP-GAT and GCN

For the original GAT architecture we fix $\boldsymbol{w} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ and define the first head as $\boldsymbol{a}_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $b_1 = -\frac{1}{\sqrt{2}}\boldsymbol{w}^T \boldsymbol{\mu}$; The second head is defined as $\boldsymbol{a}_2 = -\boldsymbol{a}_1$ and $b_2 = -b_1$. We now discuss the choice of such an ansatz. The parameter $\boldsymbol{w}$ is picked based on the optimal Bayes classifier without a graph, and the attention is set such that the first head maintains intra-class edges in C_1 and the second head maintains intra-class edges in C_0. Note that for the original GAT (Veličković et al., 2018), due to the fact that the attention mechanism consists of just one layer (i.e. a nonlinear activation applied on a linear transformation, see (12)), it is not possible for the original GAT to keep only γ_{ij} which correspond to intra-class edges. More specifically, one may use the same techniques in the proof of Theorem 3 and Corollaries 4 and 5 to prove the node separability results for the original GAT. In this particular case, the result will *depend on* q in contrast to the result we get for MLP-GAT, where no dependence of q was needed. For MLP-GAT we use the ansatz $\Psi' = (\mathbf{1}_{p \geq q} - \mathbf{1}_{p < q})\Psi$ where Ψ is given in (3) and (4) with $R = 1$. This choice of two-layer network allows us to bypass the "XOR problem" (Minsky and Papert, 1969) and separate inter-class from intra-class edges as shown in Theorem 3. Note that no single-layer architecture will be able to separate the edges due to the "XOR problem". For GCN we used the ansatz from Baranwal et al. (2021) which is also $\boldsymbol{w} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$.

5.2 Synthetic data

We use the CSBM to generate the data. In a recent work (Palowitch et al., 2022), a simple variant of the CSBM was also chosen as the default generative model for benchmarking GNN performance for node classification tasks. We present two sets of experiments. In the first set, we fix the distance between the means and vary q, and in the second set, we fix q and vary the distance. We set $n = 1000$, $d = n/\log^2(n)$, $p = 0.5$ and $\sigma = 0.1$. Results are averaged over 10 trials.

5.2.1 FIXING THE DISTANCE BETWEEN THE MEANS AND VARYING q

We consider the two regimes separately, where for the "easy regime" we fix the mean $\boldsymbol{\mu}$ to be a vector where each coordinate is equal to $10\sigma\sqrt{\log n^2}/2\sqrt{d}$. This guarantees that the distance between the means is $10\sigma\sqrt{\log n^2}$. In the "hard regime" we fix the mean $\boldsymbol{\mu}$ to a vector where each coordinate is equal to $\sigma/\sqrt{d}$, and this guarantees that the distance is σ. We fix $p = 0.5$ and vary q from $\log^2(n)/n$ to $1 - \log^2(n)/n$.

In Figure 1 we illustrate Theorem 3 and Corollaries 4, 5 for the easy regime, and in Figure 2 we illustrate Theorem 9 and Theorem 10 for the hard regime. In particular, in Figure 1a we show Theorem 3, MLP-GAT is able to classify intra-class and inter-class edges perfectly. In Figure 1b we show that in the easy regime, as claimed in Corollary 4 for MLP-

GAT, when $p \geq q$, the γ that correspond to intra-class edges concentrate around $2/np$, while the γ for the inter-class edges concentrate to tiny values; when $p < q$, we see the opposite, that is the γ that correspond to intra-class edges concentrate to tiny values, while the γ for the inter-class edges concentrate around $2/nq$. In Figure 1c we observe that the performance of MLP-GAT for node classification is independent of q in the easy regime as claimed in Corollary 5. However, in this plot, we observe that not using the graph also achieves perfect node classification, a result which is proved in Proposition 7. In the same plot, we also show the performance of simple graph convolution, where its performance depends on q (see Baranwal et al. (2021)). In Figure 2a we show Theorem 9. MLP-GAT misclassifies a constant fraction of the intra and inter edges as proved in Theorem 9. In Figure 2b we show Theorem 10, where γ's in the hard regime concentrate around uniform (GCN) coefficients for both MLP-GAT and GAT. In Figure 2c we illustrate that node classification accuracy is a function of q for MLP-GAT. This is conjectured in Conjecture 11. On the other hand, note that the performance of MLP-GAT is lower bounded by the performance of not using the graph, as proved in Theorem 14.

(a) Edge classification

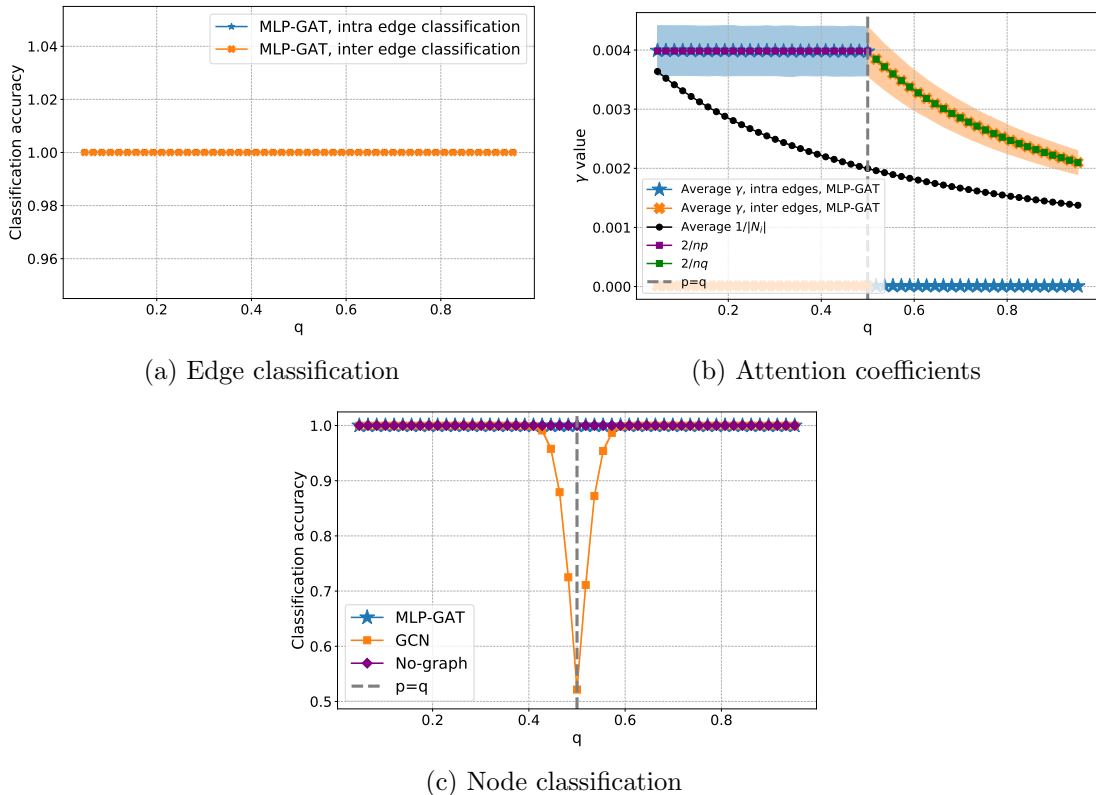

(b) Attention coefficients

(c) Node classification

Figure 1: Demonstration of Theorem 3 and Corollaries 4, 5 for the easy regime. The shaded areas in the plots show standard deviation. Unlike GCN, the performance of MLP-GAT does not degrade as we vary q.

(a) Edge classification
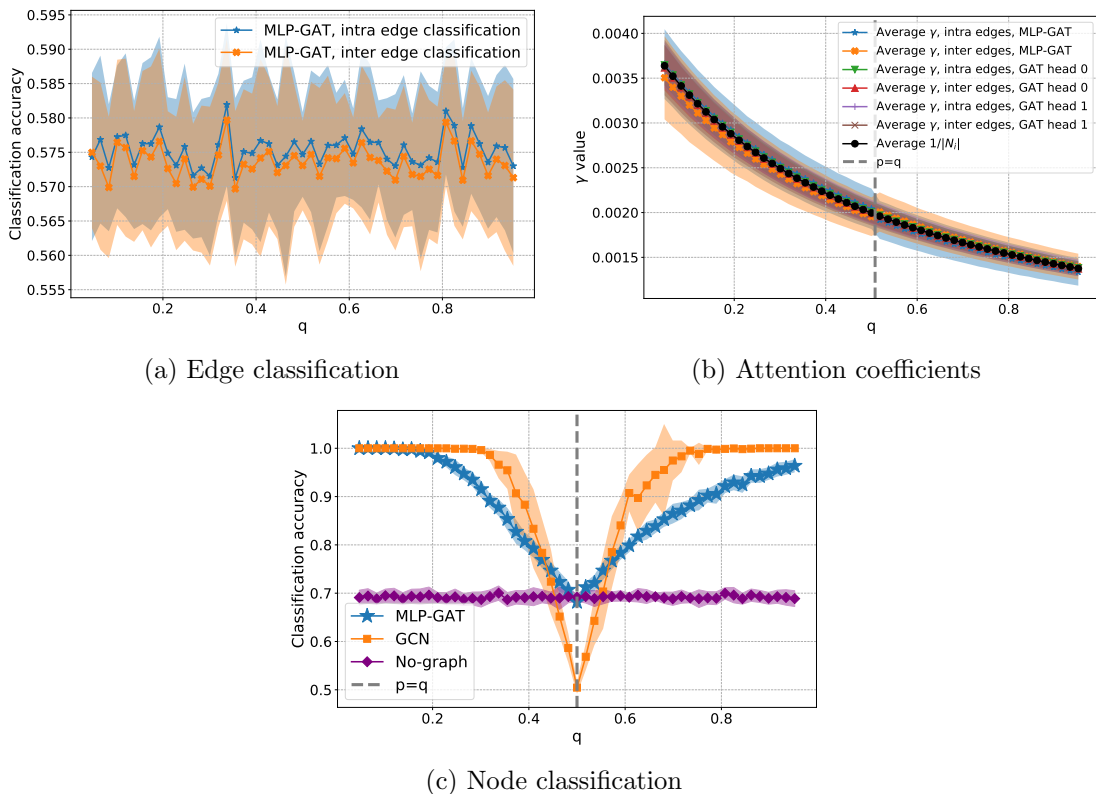
(b) Attention coefficients

(c) Node classification

Figure 2: Demonstration of Theorem 9 and Theorem 10 for the hard regime. The shaded areas in the plots show standard deviation. Unlike GCN, the performance of MLP-GAT is lower bounded by the performance of not using the graph.

5.2.2 FIXING q AND VARYING THE DISTANCE BETWEEN THE MEANS

We consider the case where $q = 0.1$. In Figure 3 we show how the attention coefficients of MLP-GAT and GAT, the node and edge classification depend on the distance between the means. We also add a vertical line at σ to approximately separate the easy (left of σ) and hard (right of σ) regimes. Figure 3c illustrates Theorems 3 and 9 in the hard and easy regimes, respectively. In particular, we observe that in the hard regime, MLP-GAT fails to distinguish intra from inter edges, while in the easy regime, it is able to do that perfectly for a large enough distance between the means.

In Figure 3a we observe that in the hard regime, γ concentrate around the uniform (GCN) coefficients, while in the easy regime, MLP-GAT is able to maintain the γ for intra-class edges, while it sets the γ to tiny values for inter-class edges. In Figure 3b. we observe that in the hard regime, the γ of GAT concentrate around the uniform coefficients (proved in Theorem 10), while in the easy regime although the γ concentrate, GAT is not able to distinguish intra from inter edges. This makes sense since the separation of edges can't be done by simple linear classifiers used by GAT, see the discussion below Theorem 10. Finally, in Figure 3d we show node classification results for MLP-GAT. In the easy regime,

we observe perfect classification as proved in Corollary 5. However, as the distance between the means decreases, we observe that MLP-GAT starts to misclassify nodes.

(a) Attention coefficients of MLP-GAT

(b) Attention coefficients of GAT

(c) Edge classification accuracy
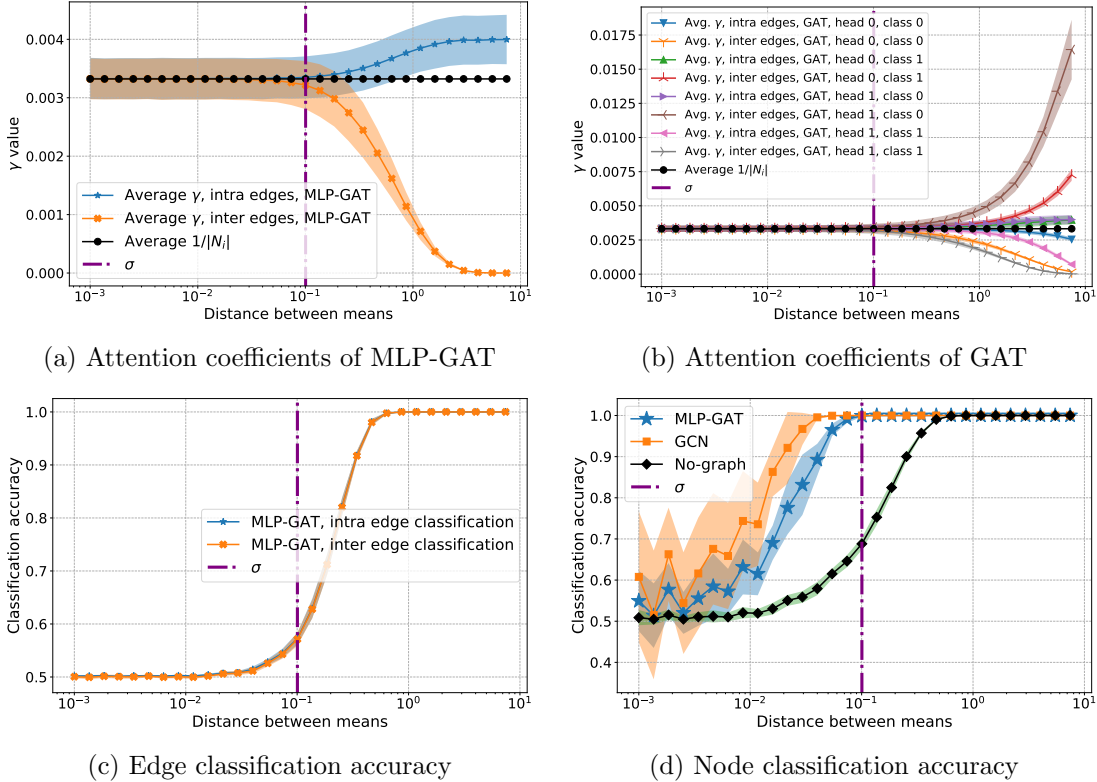
(d) Node classification accuracy

Figure 3: Attention coefficients of MLP-GAT and GAT, node and edge classification as a function of the distance between the means. Shaded areas show standard deviation. When there is a sufficient distance between the means, attention coefficients of MLP-GAT demonstrate a nice separation while GAT does not.

5.3 Real data

For the experiment on real data, we illustrate the attention coefficients, node and edge classification for MLP-GAT as a function of the distance between the means. We use popular real-world graph datasets Cora, PubMed, and CiteSeer collected by PyTorch Geometric (Fey and Lenssen, 2019) and ogbn-arxiv from Open Graph Benchmark (Hu et al., 2020). The datasets come with multiple classes, however, for each of our experiments, we do a one-v.s.-all classification for a single class. This is a semi-supervised problem, only a fraction of the training nodes have labels. The rest of the nodes are used for measuring prediction accuracy. To control the distance between the means of the problem we use the true labels to determine the class of each node and then we compute the empirical mean for each class. We subtract the empirical means from their corresponding classes and we also add means

$\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$ to each class, respectively. This modification can be thought of as translating the mean of the distribution of the data for each class.

(a) Edge class., Cora

(b) Attention coef., Cora

(c) Node class., Cora

(d) Edge class., PubMed

(e) Attention coef., PubMed

(f) Node class., PubMed

(g) Edge class., CiteSeer

(h) Attention coef., CiteSeer

(i) Node class., CiteSeer

(j) Edge class., ogbn-arxiv

(k) Attention coef., ogbn-arxiv
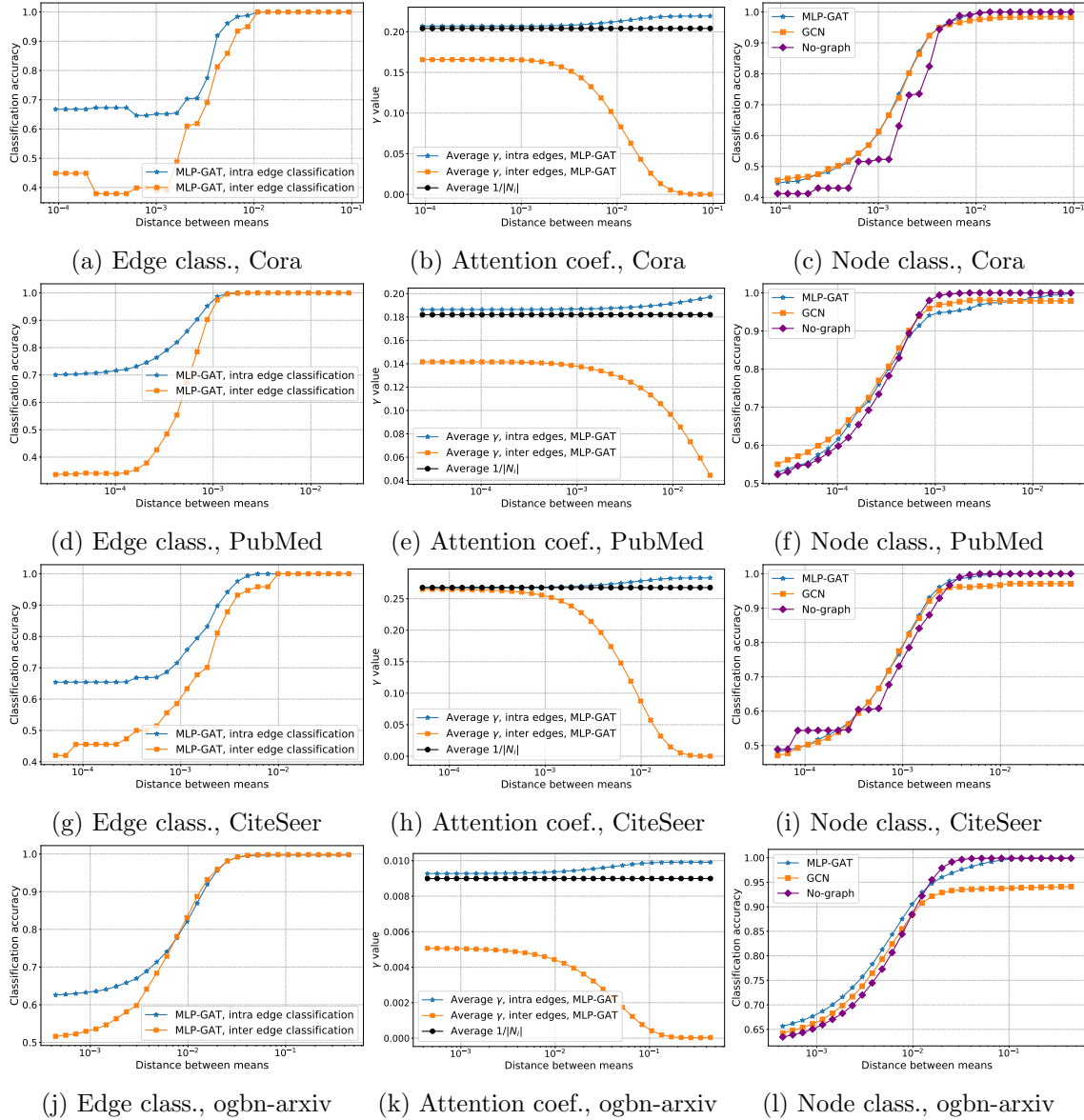
(l) Node class., ogbn-arxiv

Figure 4: Attention coefficients, node and edge classification for MLP-GAT as a function of the distance between the means for real data.

The results of this experiment are shown in Figure 4. For Cora, PubMed, and CiteSeer we show results for class 0 since these are small datasets, each dataset contains at most 7 classes, and the classes have similar sizes. In our experiments on other classes, we observed that the results are similar. For ogbn-arxiv we show results for the largest class (i.e. class 16) since this is a larger dataset which has 40 imbalanced classes. Picking a large class

makes the one-v.s.-all classification task more balanced. In our experiments on other classes having similar sizes, we obtained similar results. We note that in the real data, we also observe similar behavior of MLP-GAT in the easy and hard regimes as for the synthetic data. In particular, for all datasets as the distance of means increases, MLP-GAT is able to accurately classify intra-class and inter-class edges, see Figures 4a, 4d and 4g. Moreover, as the distance between the means increases, the average intra-class γ becomes much larger than the average inter-class γ, see Figures 4b, 4e, 4h, and 4k, and the model is able to classify the nodes accurately, see Figures 4c, 4f, 4i, and 4l. On the contrary, in the same figures, we observe that as the distance between the means decreases then MLP-GAT is not able to separate intra-class from inter-class edges, the averaged γ are very close to uniform coefficients and the model can't classify the nodes accurately.

Note that Figure 4 does not show the standard deviation for the attention coefficients γ. We show the standard deviation of γ in Figure 5. We observe that the standard deviation is higher than what we observed in the synthetic data. In particular, it can be more than half of the averaged γ. This is to be expected since for the real data the degrees of the nodes do not concentrate as well. In Figure 5 we show that the standard deviation of the uniform coefficients $1/|N_i|$ is also high. For Cora, PubMed, and CiteSeer, the standard deviation for intra-class γ is similar to that of $1/|N_i|$, while the deviation for inter-class γ is large for a small distance between the means, but it gets much smaller as the distance increases. For ognb-arxiv, the standard deviation of $1/|N_i|$ is particularly high. This implies that the degree distribution of nodes of ogbn-arxiv has a heavy tail, which could potentially result from the graph structure being noisier than other datasets and also explain GCN's relatively much worse performance when the distance between the means is large.

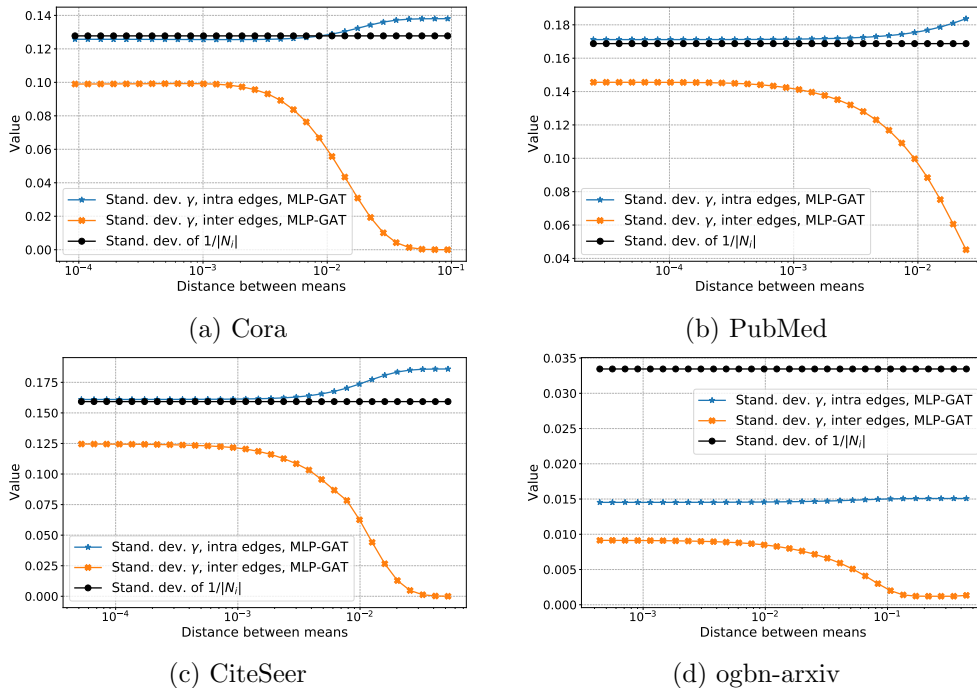(a) Cora (b) PubMed

(c) CiteSeer (d) ogbn-arxiv

Figure 5: Standard deviation for attention coefficients of MLP-GAT.

6. Summary of implications for practitioners

While this work focuses on theoretical understanding of graph attention's capability for edge and node classification using the CSBM generative model, our findings yield a series of potential suggestions for GNN practitioners. In this section we provide some interesting practical implications of our results.

6.1 Why graph attention? Benefits of graph attention's robustness to structural noise

When the graph is very noisy, e.g. when a node has a similar number of neighbors from the same class and from different classes, simple graph convolution will mix up node features and thus make nodes from different classes indistinguishable. In this case, simple graph convolution can do more harm than good. However, in practice, it is difficult to determine how noisy the graph is, or if the graph is even useful at all. This could pose a challenge in choosing an architecture. Our results in Theorem 14 and Corollary 15 imply that graph attention has the ability to dramatically reduce the impact of a noisy graph, in a way such that the output is at least as good as the output from the best linear classifier (on the input) which does not rely on the graph. This shows that graph attention convolution is more robust against structural noise in the graph, and hence on noisy graphs, it is strictly better than simple graph convolution.

6.2 Which attention architecture? Benefits of multi-layer attention architecture

In this work, we are able to obtain positive results for graph attention by using the two-layer MLP attention architecture in (3). This is different from the original GAT which uses a single-layer attention architecture (Veličković et al., 2018). In our analyses and empirical experiments, we found that the original single-layer attention does not have the important properties required for obtaining positive results (e.g. Theorem 3, Corollary 5, Theorem 14, Corollary 16) for graph attention convolution. Coincidentally, this aligns with the findings of Brody et al. (2022), where the authors study limitations of the original GAT architecture from a different perspective than ours, and they propose a new architecture termed GATv2. The two-layer MLP attention architecture that we consider in (3) encompasses GATv2 as a special case. Therefore, the two-layer MLP attention architecture can be a good candidate to consider when practitioners search for a suitable graph attention architecture for their specific downstream tasks. On the other hand, our results in Section 3.2 imply that even the two-layer MLP attention architecture (and hence GATv2) has limitations when the node features are noisy. To fix that, a potential solution is to incorporate additional information such as edge features, which we discuss next.

6.3 Will additional information help? Benefits of incorporating good edge features

Even though we do not consider edge features in our analyses, our results in Theorem 13 imply that good attention functions that are able to classify the edges independently from the node features can be very helpful, as they help reduce the threshold under which graph

attention convolution would fail to separate the nodes. One potential way to obtain good attention functions that behave like the one given in (13) is by incorporating good edge features. Furthermore, given our result in Theorem 9, which says that graph attention based on noisy node features cannot perfectly classify the edges, the importance of incorporating informative edge features that are more indicative of edge class memberships (if they are accessible in practice) into the attention mechanism is more pronounced.

7. Conclusion and future work

In this work, we study the impact of graph attention on edges and its implications for node classification. We show that graph attention improves robustness to noise in the graph structure. We also show that graph attention may not be very useful in a "hard" regime where the node features are noisy. Our work shows that single-layer graph attention convolution has limited power at distinguishing intra-class from inter-class edges. Given the empirical successes of graph attention and its many variants, a promising future work is to study the power of multi-layer graph attention convolutions for distinguishing intra-class and inter-class edges. Moreover, our negative results in Section 3.2 for edge/node classification pertains to perfect classification and almost perfect classification. In practice, misclassification of a small constant fraction of nodes/edges is often inevitable, but nonetheless useful. Therefore, an interesting future line of work is to characterize the threshold under which graph attention is going to misclassify a certain proportion of nodes. Finally, variants of graph attention networks have been successfully used in tasks other than node classification, such as link prediction and graph classification. These tasks are typically solved by architectures that add a final aggregation layer which combines node representations generated from graph attention convolution. It is an interesting future direction to develop a good understanding of the benefits and limitations of the graph attention mechanism for those tasks.

Appendix A. Proofs

We define the following high-probability events which will be used in some proofs. Each of these events holds with probability at least $1 - o(1)$, which follows from straightforward applications of Chernoff bound and union bound, e.g., see (Baranwal et al., 2021).

Definition 17. *Define the following events over the randomness of* $\mathbf{A}$ *and* $\{\epsilon_i\}_{i \in [n]}$ *and* $\{\mathbf{X}_i\}_{i \in [n]}$,

- $\mathcal{E}_1$ *is the event that* $|C_0| = \frac{n}{2} \pm O(\sqrt{n \log n})$ *and* $|C_1| = \frac{n}{2} \pm O(\sqrt{n \log n})$.

- $\mathcal{E}_2$ *is the event that for each* $i \in [n]$, $\mathbf{D}_{ii} = \frac{n(p+q)}{2} \left(1 \pm \frac{10}{\sqrt{\log n}} \right)$.

- $\mathcal{E}_3$ *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* $|C_1 \cap N_i| = \mathbf{D}_{ii} \cdot \frac{(1-\epsilon_i)q+\epsilon_i p}{p+q} \left(1 \pm \frac{10}{\sqrt{\log n}} \right)$.

- $\mathcal{E}_4$ *is the event that for each* $i \in [n]$, $\left| \tilde{\boldsymbol{w}}^T \mathbf{X}_i - \mathbf{E} \left[\tilde{\boldsymbol{w}}^T \mathbf{X}_i \right] \right| \leq 10\sigma\sqrt{\log n}$.

- $\mathcal{E}^*$ *is the intersection of the above 4 events.*

Lemma 18 (Baranwal et al. (2021)). *With probability at least* $1 - o(1)$ *event* $\mathcal{E}^*$ *holds.*

Some of our proofs also utilize the following simple observation on the mutual independence among $\{\boldsymbol{w}^T \boldsymbol{g}_i\}_{i \in [n]}$ when $\{\boldsymbol{g}_i\}_{i \in [n]}$ are i.i.d. Gaussian random vectors.

Observation 19. *Fix* $\boldsymbol{w} \neq 0$ *in* $\mathbb{R}^d$ *and let* $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n$ *be i.i.d. drawn from* $N(0, \mathbf{I})$. *Then* $\boldsymbol{w}^T \boldsymbol{g}_1, \boldsymbol{w}^T \boldsymbol{g}_2, \ldots, \boldsymbol{w}^T \boldsymbol{g}_n$ *are independent.*

Proof Note that since $\boldsymbol{w}^T \boldsymbol{g}_i \sim N(0, \|\boldsymbol{w}\|^2)$, it suffices to prove that the covariance $\mathbf{E}[\boldsymbol{w}^T \boldsymbol{g}_i \cdot \boldsymbol{w}^T \boldsymbol{g}_j] = 0$ for all $i \neq j$. By definition, for $i \neq j$,

$$\mathbf{E}[\boldsymbol{w}^T \boldsymbol{g}_i \cdot \boldsymbol{w}^T \boldsymbol{g}_j] = \mathbf{E}\left[\sum_{k \in [d]} \sum_{\ell \in [d]} \boldsymbol{w}_k \boldsymbol{w}_\ell \boldsymbol{g}_{ik} \boldsymbol{g}_{j\ell} \right] = \sum_{k \in [d]} \sum_{\ell \in [d]} \boldsymbol{w}_k \boldsymbol{w}_\ell \, \mathbf{E}[\boldsymbol{g}_{ik} \boldsymbol{g}_{j\ell}] = 0,$$

where the last equality follows from independence between $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$. ∎

A.1 Proof of Theorem 3

We restate Theorem 3 for convenience.

Theorem. *Suppose that* $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$. *Then with probability at least* $1 - o(1)$ *over the data* $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, *the two-layer MLP attention architecture* Ψ *given in (3) and (4) separates intra-class edges from inter-class edges.*

Recall from (5) that

$$\Psi(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} -2R(1-\beta)\tilde{\boldsymbol{w}}^T \mathbf{X}_i, & \text{if } \tilde{\boldsymbol{w}}^T \mathbf{X}_j \leq -\left| \tilde{\boldsymbol{w}}^T \mathbf{X}_i \right|, \\ 2R(1-\beta)\operatorname{sign}(\tilde{\boldsymbol{w}}^T \mathbf{X}_i)\tilde{\boldsymbol{w}}^T \mathbf{X}_j, & \text{if } -\left| \tilde{\boldsymbol{w}}^T \mathbf{X}_i \right| < \tilde{\boldsymbol{w}}^T \mathbf{X}_j < \left| \tilde{\boldsymbol{w}}^T \mathbf{X}_i \right|, \\ 2R(1-\beta)\tilde{\boldsymbol{w}}^T \mathbf{X}_i, & \text{if } \tilde{\boldsymbol{w}}^T \mathbf{X}_j \geq \left| \tilde{\boldsymbol{w}}^T \mathbf{X}_i \right|. \end{cases}$$

We will condition on the event that $\left|\tilde{\boldsymbol{w}}^T\mathbf{X}_i - \mathbf{E}\left[\tilde{\boldsymbol{w}}^T\mathbf{X}_i\right]\right| \le 10\sigma\sqrt{\log n}$ for all $i \in [n]$, which holds with probability at least $1 - O(1/n^{99})$ following a union bound and the Gaussian tail probability. Under this event, because $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$, for all $i, j \in C_1$ we have

$$\text{sign}(\tilde{\boldsymbol{w}}^T\mathbf{X}_i) = \text{sign}(\tilde{\boldsymbol{w}}^T\mathbf{X}_j) = 1,$$
$$\min\{\tilde{\boldsymbol{w}}^T\mathbf{X}_i, \tilde{\boldsymbol{w}}^T\mathbf{X}_j\} \ge \|\boldsymbol{\mu}\| - 10\sigma\sqrt{\log n},$$
$$\max\{\tilde{\boldsymbol{w}}^T\mathbf{X}_i, \tilde{\boldsymbol{w}}^T\mathbf{X}_j\} \le \|\boldsymbol{\mu}\| + 10\sigma\sqrt{\log n},$$

and hence

$$\Psi(\mathbf{X}_i, \mathbf{X}_j) \ge 2R(1-\beta) \cdot \min\{\tilde{\boldsymbol{w}}^T\mathbf{X}_i, \tilde{\boldsymbol{w}}^T\mathbf{X}_j\} \ge 2R(1-\beta) \cdot (\|\boldsymbol{\mu}\| - O(\sigma\sqrt{\log n})),$$
$$\Psi(\mathbf{X}_i, \mathbf{X}_j) \le 2R(1-\beta) \cdot \max\{\tilde{\boldsymbol{w}}^T\mathbf{X}_i, \tilde{\boldsymbol{w}}^T\mathbf{X}_j\} \le 2R(1-\beta) \cdot (\|\boldsymbol{\mu}\| + O(\sigma\sqrt{\log n})),$$

which implies that $\Psi(\mathbf{X}_i, \mathbf{X}_j) = \|\boldsymbol{\mu}\| \pm O(\sigma\sqrt{\log n})$ for $i, j \in C_1$. Similarly, we get that

$$\Psi(\mathbf{X}_i, \mathbf{X}_j) = \begin{cases} \|\boldsymbol{\mu}\| \pm O(\sigma\sqrt{\log n}) = \|\boldsymbol{\mu}\|(1 \pm o(1)), & \text{if } (i,j) \text{ is an intra-class edge,} \\ -\|\boldsymbol{\mu}\| \pm O(\sigma\sqrt{\log n}) = -\|\boldsymbol{\mu}\|(1 \pm o(1)), & \text{if } (i,j) \text{ is an inter-class edge.} \end{cases} \tag{14}$$

Therefore, we probability at least $1 - o(1)$, we have that

$$\text{sign}(\Psi(\mathbf{X}_i, \mathbf{X}_j)) = \begin{cases} 1, & \text{if } (i,j) \text{ is an intra-class edge,} \\ -1, & \text{if } (i,j) \text{ is an inter-class edge,} \end{cases}$$

which means perfect separability of edges, and the proof is complete.

A.2 Proof of Corollary 4

We restate Corollary 4 for convenience.

Corollary. *Suppose that* $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ *and that Assumption 1 holds. Then with probability at least* $1 - o(1)$ *over the data* $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, *the attention architecture* Ψ' *yields attention coefficients* γ_{ij} *such that*

1. *If* $p \ge q$, *then* $\gamma_{ij} = \frac{2}{np}(1 \pm o(1))$ *if* (i,j) *is an intra-class edge and* $\gamma_{ij} = o(\frac{1}{n(p+q)})$ *otherwise;*

2. *If* $p < q$, *then* $\gamma_{ij} = \frac{2}{nq}(1 \pm o(1))$ *if* (i,j) *is an inter-class edge and* $\gamma_{ij} = o(\frac{1}{n(p+q)})$ *otherwise.*

The proof is straightforward by considering the cases $p \ge q$ and $p < q$ separately. When $p \ge q$, we have $\Psi' = \Psi$. Using the specification of Ψ in (3) and (4), the definition of attention coefficients in (1), the high probability event in Lemma 18, the expression of $\Psi(\mathbf{X}_i, \mathbf{X}_j)$ in (14), and picking R such that $1/R = \omega(\sigma\sqrt{\log n})$ and $1/R = o(\|\boldsymbol{\mu}\|)$, we obtain the claimed results. The result when $p < q$ is obtained in the same way.

A.3 Proof of Corollary 5

We restate Corollary 5 for convenience.

Corollary. *Suppose that $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ and that Assumption 1 holds. Then with probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, using the attention architecture Ψ' with the graph attention convolution given in (2), where f is set to be the identify function, the model separates the nodes.*

We prove the case $p \geq q$ and the case $p < q$ follows analogously. Consider the attention architecture $\Psi' = (\mathbf{1}_{p \geq q} - \mathbf{1}_{p < q}) \cdot \Psi = \Psi$, where Ψ is given in (3) and (4). Pick R such that $1/R = \omega(\sigma\sqrt{\log n})$ and $1/R = o(\|\boldsymbol{\mu}\|))$. Assume that $i \in C_1$, and denote the graph attention convolution output as

$$h_i' \overset{\text{def}}{=} \sum_{j \in N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j.$$

We will condition on the event $\mathcal{E}^*$, which holds with probability at least $1 - o(1)$. By using Corollary 4 we have

$$\sum_{j \in N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j = \sum_{j \in C_0 \cap N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j + \sum_{j \in C_1 \cap N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j$$

$$\leq |C_1 \cap N_i| \left(\frac{2}{np}(1 \pm o(1)) \left(\|\boldsymbol{\mu}\| + 10\sigma\sqrt{\log n} \right) \right)$$

$$+ |C_0 \cap N_i| \left(o\left(\frac{1}{n(p+q)} \right) \left(-\|\boldsymbol{\mu}\| + 10\sigma\sqrt{\log n} \right) \right)$$

$$= (1 \pm o(1)) \cdot \left(\|\boldsymbol{\mu}\| + 10\sigma\sqrt{\log n} \right) - \frac{nq(1 \pm o(1))}{\omega(n(p+q))} \cdot \left(\|\boldsymbol{\mu}\| - 10\sigma\sqrt{\log n} \right)$$

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

Similarly, we have that

$$\sum_{j \in N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j \geq (1 \pm o(1)) \cdot \left(\|\boldsymbol{\mu}\| - 10\sigma\sqrt{\log n} \right) - \frac{nq(1 \pm o(1))}{\omega(n(p+q))} \cdot \left(\|\boldsymbol{\mu}\| + 10\sigma\sqrt{\log n} \right)$$

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

This means that $h_i' = \|\boldsymbol{\mu}\|(1 \pm o(1))$ for $i \in C_1$. Applying the same reasoning we get that $h_i' = -\|\boldsymbol{\mu}\|(1 \pm o(1))$ for $i \in C_0$. Therefore, with probability at least $1 - o(1)$, the graph attention convolution separates the nodes.

A.4 Proof of Lemma 8

We restate Lemma 8 for convenience.

Lemma. *Let $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ and let $\mathbf{X}_{ij}'$ be defined as in (9). The Bayes optimal classifier for $\mathbf{X}_{ij}'$ is realized by the following function,*

$$h^*(\boldsymbol{x}) = \begin{cases} 0, & \text{if } p\cosh\left(\frac{\boldsymbol{x}^T \boldsymbol{\mu}'}{\sigma^2} \right) \leq q\cosh\left(\frac{\boldsymbol{x}^T \boldsymbol{\nu}'}{\sigma^2} \right), \\ 1, & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\mu}' \overset{\text{def}}{=} \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix}$ and $\boldsymbol{\nu}' \overset{\text{def}}{=} \begin{pmatrix} \boldsymbol{\mu} \\ -\boldsymbol{\mu} \end{pmatrix}$.

Proof Note that $\mathbf{X}'_{ij}$ is a mixture of $2d$-dimensional Gaussian distributions,

$$\mathbf{X}'_{ij} \sim \begin{cases} N(-\boldsymbol{\mu}', \sigma^2 \mathbf{I}) & i \in C_0, j \in C_0 \\ N(\boldsymbol{\mu}', \sigma^2 \mathbf{I}) & i \in C_1, j \in C_1 \\ N(-\boldsymbol{\nu}', \sigma^2 \mathbf{I}) & i \in C_0, j \in C_1 \\ N(\boldsymbol{\nu}', \sigma^2 \mathbf{I}) & i \in C_1, j \in C_0 \end{cases}.$$

The optimal classifier is then given by

$$h^*(\boldsymbol{x}) = \arg\max_{c \in \{0,1\}} \mathbf{Pr}[y = c \mid \boldsymbol{x}].$$

Note that $\mathbf{Pr}[y = 0] = \frac{q}{p+q}$ and $\mathbf{Pr}[y = 1] = \frac{p}{p+q}$. Thus, by Bayes rule, we obtain that

$$\mathbf{Pr}[y = c \mid \boldsymbol{x}] = \frac{\mathbf{Pr}[y = c] \cdot f_{\boldsymbol{x}|y}(\boldsymbol{x} \mid y = c)}{\mathbf{Pr}[y = 0] f_{\boldsymbol{x}|y=0}(\boldsymbol{x} \mid y = 0) + \mathbf{Pr}[y = 1] f_{\boldsymbol{x}|y=1}(\boldsymbol{x} \mid y = 1)}$$

$$= \frac{1}{1 + \frac{\mathbf{Pr}[y=1-c] \cdot f_{\boldsymbol{x}|y}(\boldsymbol{x}|y=1-c)}{\mathbf{Pr}[y=c] \cdot f_{\boldsymbol{x}|y}(\boldsymbol{x}|y=c)}}.$$

Suppose that $\boldsymbol{x} = \mathbf{X}'_{ij}$ such that $i \nsim j$. Then $h^*(\boldsymbol{x}) = 0$ if and only if $\mathbf{Pr}[y = 0 \mid \boldsymbol{x}] \geq \frac{1}{2}$. Hence, for $c = 0$ we require that

$$\frac{\mathbf{Pr}[y = 1 - c] \cdot f_{\boldsymbol{x}|y}(\boldsymbol{x} \mid y = 1 - c)}{\mathbf{Pr}[y = c] \cdot f_{\boldsymbol{x}|y}(\boldsymbol{x} \mid y = c)} = \frac{p}{q} \frac{f_{\boldsymbol{x}|y}(\boldsymbol{x} \mid y = 1)}{f_{\boldsymbol{x}|y}(\boldsymbol{x} \mid y = 0)} = \frac{p}{q} \frac{\cosh\left(\frac{1}{\sigma^2} \boldsymbol{x}^T \boldsymbol{\mu}'\right)}{\cosh\left(\frac{1}{\sigma^2} \boldsymbol{x}^T \boldsymbol{\nu}'\right)} \leq 1,$$

Similarly, we obtain the reverse condition for $h^*(\boldsymbol{x}) = 1$. ∎

A.5 Proof of Theorem 9

We restate Theorem 9 for convenience.

Theorem. *Suppose* $\|\boldsymbol{\mu}\| = \kappa \sigma$ *for some* $\kappa > 0$ *and let* Ψ *be any attention mechanism. Then,*

1. *With probability at least* $1 - o(1)$, Ψ *fails to correctly classify at least* $2 \cdot \Phi_c(\kappa)^2$ *fraction of inter-class edges;*

2. *For any* $K > 1$ *if* $q > \frac{K \log^2 n}{n \Phi_c(\kappa)^2}$, *then with probability at least* $1 - O(n^{-\frac{K}{4} \Phi_c(\kappa)^2 \log n})$, Ψ *misclassify at least one inter-class edge.*

We will write $i \sim j$ if node i and node j are in the same class and $i \nsim j$ otherwise. From Lemma 8, we observe that for successful classification by the optimal classifier, we need

$$p \cosh\left(\frac{\boldsymbol{x}^T \boldsymbol{\mu}'}{\sigma^2}\right) \leq q \cosh\left(\frac{\boldsymbol{x}^T \boldsymbol{\nu}'}{\sigma^2}\right) \quad \text{for } i \nsim j,$$

$$p \cosh\left(\frac{\boldsymbol{x}^T \boldsymbol{\mu}'}{\sigma^2}\right) > q \cosh\left(\frac{\boldsymbol{x}^T \boldsymbol{\nu}'}{\sigma^2}\right) \quad \text{for } i \sim j.$$

We will split the analysis into two cases. First, note that when $p \geq q$ we have for $i \nsim j$ that

$$p \cosh\left(\tfrac{\boldsymbol{x}^T \boldsymbol{\mu}'}{\sigma^2}\right) \leq q \cosh\left(\tfrac{\boldsymbol{x}^T \boldsymbol{\nu}'}{\sigma^2}\right) \implies \cosh\left(\tfrac{\boldsymbol{x}^T \boldsymbol{\mu}'}{\sigma^2}\right) \leq \cosh\left(\tfrac{\boldsymbol{x}^T \boldsymbol{\nu}'}{\sigma^2}\right) \implies |\boldsymbol{x}^T \boldsymbol{\mu}'| \leq |\boldsymbol{x}^T \boldsymbol{\nu}'|.$$

In the first implication, we used that $p \geq q$, while the second implication follows from the fact that $\cosh(a) \leq \cosh(b) \implies |a| \leq |b|$ for all $a, b \in \mathbb{R}$. Similarly, for $p < q$ we have for $i \sim j$ that

$$p \cosh\left(\tfrac{\boldsymbol{x}^T \boldsymbol{\mu}'}{\sigma^2}\right) > q \cosh\left(\tfrac{\boldsymbol{x}^T \boldsymbol{\nu}'}{\sigma^2}\right) \implies \cosh\left(\tfrac{\boldsymbol{x}^T \boldsymbol{\mu}'}{\sigma^2}\right) > \cosh\left(\tfrac{\boldsymbol{x}^T \boldsymbol{\nu}'}{\sigma^2}\right) \implies |\boldsymbol{x}^T \boldsymbol{\mu}'| > |\boldsymbol{x}^T \boldsymbol{\nu}'|.$$

Therefore, for each of the above cases, we can upper bound the probability for either $i \sim j$ or $i \nsim j$ that $\mathbf{X}'_{ij}$ is correctly classified, by the probability of the event $|\mathbf{X}'^T_{ij} \boldsymbol{\mu}'| \leq |\mathbf{X}'^T_{ij} \boldsymbol{\nu}'|$ or equivalently $|\mathbf{X}'^T_{ij} \boldsymbol{\mu}'| > |\mathbf{X}'^T_{ij} \boldsymbol{\nu}'|$. We focus on the former as the latter is equivalent and symmetric. Writing $\mathbf{X}_i = \boldsymbol{\mu} + \sigma \boldsymbol{g}_i$ and $\mathbf{X}_j = -\boldsymbol{\mu} + \sigma \boldsymbol{g}_j$, we have that for $i \in C_1$ and $j \in C_0$,

$$\begin{aligned}
\mathbf{Pr}[h^*(\mathbf{X}'_{ij}) = 0] &\leq \mathbf{Pr}\left[|\mathbf{X}'^T_{ij} \boldsymbol{\mu}'| \leq |\mathbf{X}'^T_{ij} \boldsymbol{\nu}'|\right] \\
&= \mathbf{Pr}\left[|\mathbf{X}^T_i \boldsymbol{\mu} + \mathbf{X}^T_j \boldsymbol{\mu}| \leq |\mathbf{X}^T_i \boldsymbol{\mu} - \mathbf{X}^T_j \boldsymbol{\mu}|\right] \\
&= \mathbf{Pr}\left[\sigma|\boldsymbol{g}^T_i \boldsymbol{\mu} + \boldsymbol{g}^T_j \boldsymbol{\mu}| \leq |\pm 2\|\boldsymbol{\mu}\|^2 + \sigma \boldsymbol{g}^T_i \boldsymbol{\mu} - \sigma \boldsymbol{g}^T_j \boldsymbol{\mu}|\right] \\
&\leq \mathbf{Pr}\left[|\boldsymbol{g}^T_i \hat{\boldsymbol{\mu}} + \boldsymbol{g}^T_j \hat{\boldsymbol{\mu}}| - |\boldsymbol{g}^T_i \hat{\boldsymbol{\mu}} - \boldsymbol{g}^T_j \hat{\boldsymbol{\mu}}| \leq \frac{2\|\boldsymbol{\mu}\|}{\sigma}\right] \\
&= \mathbf{Pr}\left[|\boldsymbol{g}^T_i \hat{\boldsymbol{\mu}} + \boldsymbol{g}^T_j \hat{\boldsymbol{\mu}}| - |\boldsymbol{g}^T_i \hat{\boldsymbol{\mu}} - \boldsymbol{g}^T_j \hat{\boldsymbol{\mu}}| \leq 2\kappa\right],
\end{aligned}$$

where we denote $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. In the second to last step above, we used triangle inequality to pull $2\|\boldsymbol{\mu}\|^2$ outside the absolute value, while in the last equation we use $\|\boldsymbol{\mu}\| = \kappa\sigma$.

We now denote $z_i = \boldsymbol{g}^T_i \hat{\boldsymbol{\mu}}$ for all $i \in [n]$. Then the above probability is $\mathbf{Pr}[|z_i + z_j| - |z_i - z_j| \leq 2\kappa]$, where $z_i, z_j \sim N(0,1)$ are independent random variables. Note that we have

$$\begin{aligned}
\mathbf{Pr}[h^*(\mathbf{X}'_{ij}) = 0] &\leq \mathbf{Pr}[|z_i + z_j| - |z_i - z_j| \leq 2\kappa] \\
&= \mathbf{Pr}[|z_i + z_j| - |z_i - z_j| \leq 2\kappa, |z_i| \leq \kappa] \\
&\quad + \mathbf{Pr}[|z_i + z_j| - |z_i - z_j| \leq 2\kappa, |z_i| > \kappa] \\
&= \mathbf{Pr}[|z_i| \leq \kappa] + \Phi(\kappa)\mathbf{Pr}[|z_i| > \kappa]. \quad (15)
\end{aligned}$$

To see how we obtain the last equation, observe that if $|z_i| \leq \kappa$ then we have

$$\begin{aligned}
|z_i + z_j| - |z_i - z_j| &= |z_i + z_j| - |z_j - z_i| \\
&\leq |z_i| + |z_j| - |z_j - z_i| && \text{by triangle inequality} \\
&\leq |z_i| + |z_j| - \big||z_j| - |z_i|\big| && \text{by reverse triangle inequality} \\
&\leq |z_i| + |z_j| - (|z_j| - |z_i|) = 2|z_i| \\
&\leq 2\kappa,
\end{aligned}$$

hence, $\mathbf{Pr}[|z_i + z_j| - |z_i - z_j| \leq 2\kappa, |z_i| \leq \kappa] = \mathbf{Pr}[|z_i| \leq \kappa]$. On the other hand, for $|z_i| > \kappa$, we look at each case, conditioned on the events $z_i > \kappa$ and $z_i < -\kappa$ for each of

the four cases based on the signs of $z_i + z_j$ and $z_i - z_j$. We denote by E the event that $|z_i + z_j| - |z_i - z_j| \le 2\kappa$, and analyze the cases in detail. First consider the case $z_i < -\kappa$:

$$\mathbf{Pr}[E, z_i + z_j \ge 0, z_i - z_j \ge 0 \mid z_i < -\kappa] = \mathbf{Pr}[z_j \le z_i, z_j \ge -z_i \mid z_i < -\kappa] = 0,$$
$$\mathbf{Pr}[E, z_i + z_j \ge 0, z_i - z_j < 0 \mid z_i < -\kappa] = \mathbf{Pr}[z_j > |z_i|, z_i \le \kappa \mid z_i < -\kappa] = \Phi(z_i),$$
$$\mathbf{Pr}[E, z_i + z_j < 0, z_i - z_j \ge 0 \mid z_i < -\kappa] = \mathbf{Pr}[z_j < -|z_i|, z_i \ge -\kappa \mid z_i < -\kappa] = 0,$$
$$\mathbf{Pr}[E, z_i + z_j < 0, z_i - z_j < 0 \mid z_i < -\kappa] = \mathbf{Pr}[z_i < z_j < -z_i, z_j > -\kappa \mid z_i < -\kappa]$$
$$= \Phi(\kappa) - \Phi(z_i).$$

The sum of the four probabilities in the above is $\mathbf{Pr}[E \mid z_i < -\kappa] = \Phi(\kappa)$. Similarly, we analyze the other case, $z_i > \kappa$:

$$\mathbf{Pr}[E, z_i + z_j \ge 0, z_i - z_j \ge 0 \mid z_i > \kappa] = \mathbf{Pr}[-z_i \le z_j \le z_i, z_j \le \kappa \mid z_i > \kappa]$$
$$= \Phi(\kappa) - \Phi_{\mathrm{c}}(z_i),$$
$$\mathbf{Pr}[E, z_i + z_j \ge 0, z_i - z_j < 0 \mid z_i > \kappa] = \mathbf{Pr}[z_j > |z_i|, z_i \le \kappa \mid z_i > \kappa] = 0,$$
$$\mathbf{Pr}[E, z_i + z_j < 0, z_i - z_j \ge 0 \mid z_i > \kappa] = \mathbf{Pr}[z_j < -|z_i|, z_i \ge -\kappa \mid z_i > \kappa] = \Phi_{\mathrm{c}}(z_i),$$
$$\mathbf{Pr}[E, z_i + z_j < 0, z_i - z_j < 0 \mid z_i > \kappa] = \mathbf{Pr}[z_j < -z_i, z_j > z_i \mid z_i > \kappa] = 0.$$

The sum of the four probabilities above is $\mathbf{Pr}[E \mid z_i > \kappa] = \Phi(\kappa)$. Therefore, we obtain that

$$\mathbf{Pr}[|z_i + z_j| - |z_i - z_j| \le 2\kappa \mid |z_i| > \kappa] = \Phi(\kappa),$$

which justifies (15).

Next, note that $\mathbf{Pr}[|z_i| \le \kappa] = \Phi(\kappa) - \Phi_{\mathrm{c}}(\kappa)$ and $\mathbf{Pr}[|z_i| > \kappa] = 2\Phi_{\mathrm{c}}(\kappa)$, so we have from (15) that

$$\mathbf{Pr}[h^*(\mathbf{X}'_{ij}) = 0] \le \Phi(\kappa) - \Phi_{\mathrm{c}}(\kappa) + 2\Phi_{\mathrm{c}}(\kappa)\Phi(\kappa)$$
$$= 1 - 2\Phi_{\mathrm{c}}(\kappa) + 2\Phi_{\mathrm{c}}(\kappa)\Phi(\kappa) = 1 - 2\Phi_{\mathrm{c}}(\kappa)^2.$$

Thus, $\mathbf{X}'_{ij}$ is misclassified with probability at least $2\Phi_{\mathrm{c}}(\kappa)^2$.

We will now construct sets of pairs with mutually independent elements, such that the union of those sets covers all inter-class edges. This will enable us to use a concentration argument that computes the fraction of the inter-class edges that are misclassified. Since the graph operations are permutation invariant, let us assume for simplicity that $C_0 = \{1, \ldots, \frac{n}{2}\}$ and $C_1 = \{\frac{n}{2} + 1, \ldots, n\}$ for an even number of nodes n. Also, define the function

$$m(i, l) = \begin{cases} i + l, & i + l \le \frac{n}{2}, \\ i + l - \frac{n}{2}, & i + l > \frac{n}{2}. \end{cases}$$

We now construct the following sequence of sets for all $l \in \{0, \ldots, \frac{n}{2} - 1\}$:

$$S_l = \{(X_{m(i,l)}, X_{i+\frac{n}{2}}) \text{ for all } i \in C_0 \text{ such that } (m(i,l), i + n/2) \in E\}.$$

Fix $l \in \{0, \ldots, \frac{n}{2} - 1\}$ and observe that the pairs in the set S_l are mutually independent. Define a Bernoulli random variable, β_i, to be the indicator that $(X_{m(i,l)}, X_{i+\frac{n}{2}})$ is misclassified. We have that $\mathbf{E}[\beta_i] \ge 2\Phi_{\mathrm{c}}(\kappa)^2$. Note that the fraction of pairs in the set S_l that are

misclassified is $\frac{1}{|S_l|}\sum_{i:(X_{m(i,l)},X_{i+n/2})\in S_l}\beta_i$, which is a sum of independent Bernoulli random variables. Hence, by the additive Chernoff bound, we obtain

$$\mathbf{Pr}\left[\sum_{i\in C_0\cap N_{m(i,l)}}\beta_i \geq 2|S_l|\Phi_c(\kappa)^2 - |S_l|t\right] \geq 1 - \exp(-2|S_l|t^2).$$

Since $p,q = \Omega(\frac{\log^2 n}{n})$, we have by the Chernoff bound and a union bound that with probability at least $1-1/\mathrm{poly}(n)$, $|S_l| = nq(1\pm o(1))$ for all l. We now choose $t = \sqrt{\frac{C\log n}{|S_l|}} = o(1)$ to obtain that on the event where $|S_l| = nq(1\pm o(1))$, we have the following for any large $C > 1$:

$$\mathbf{Pr}\left[\frac{1}{|S_l|}\sum_{i\in C_0\cap N_{m(i,l)}}\beta_i \geq 2\Phi_c(\kappa)^2 - o(1)\right] \geq 1 - n^{-C}.$$

Following a union bound over all $l \in \{0,\ldots,\frac{n}{2}-1\}$, we conclude that for any $c > 0$,

$$\mathbf{Pr}\left[\frac{1}{|S_l|}\sum_{i\in C_0\cap N_{m(i,l)}}\beta_i \geq 2\Phi_c(\kappa)^2 - o(1),\ \forall l \in \left\{0,\ldots,\frac{n}{2}-1\right\}\right] \geq 1 - O(n^{-c}).$$

Thus, out of all the pairs $\mathbf{X}'_{ij}$ with $j \nsim i$, with probability at least $1 - o(1)$, we have that at least a fraction $2\Phi_c(\kappa)^2$ of the pairs are misclassified by the attention mechanism. This concludes part 1 of the theorem.

For part 2, note that by the additive Chernoff bound we have for any $t \in (0,1)$,

$$\mathbf{Pr}\left[\sum_{i\in C_0\cap N_{m(i,l)}}\beta_i \geq 2|S_l|\Phi_c(\kappa)^2 - |S_l|t\right] \geq 1 - \exp(-2|S_l|t^2).$$

Since $|S_l| = \frac{nq}{2}(1\pm o(1))$ with probability at least $1/\mathrm{poly}(n)$, we choose $t = 2\sqrt{\frac{K\Phi_c(\kappa)^2\log^2 n}{nq}}$ to obtain

$$\mathbf{Pr}\left[\sum_{i\in C_0\cap N_{m(i,l)}}\beta_i \geq nq\Phi_c(\kappa)^2(1\pm o(1)) - \sqrt{Knq\Phi_c(\kappa)^2\log^2 n}\right] \geq 1 - O(n^{-8K\Phi_c(\kappa)^2\log n}).$$

Now note that if $q > \frac{K\log^2 n}{n\Phi_c(\kappa)^2}$ then we have $nq\Phi_c(\kappa)^2 > K\log^2 n$, which implies that

$$nq\Phi_c(\kappa)^2 - \sqrt{Knq\Phi_c(\kappa)^2\log^2 n} > 0.$$

Hence, in this regime of q,

$$\mathbf{Pr}\left[\sum_{i\in C_0\cap N_{m(i,l)}}\beta_i > 0\right] \geq 1 - O(n^{-8K\Phi_c(\kappa)^2\log n}),$$

and the proof is complete.

A.6 Proof of Theorem 10

We restate Theorem 10 for convenience

Theorem. *Assume that $\|\boldsymbol{\mu}\| \leq K\sigma$ and $\sigma \leq K'$ for some absolute constants K and K'. Moreover, assume that the parameters $(\boldsymbol{w}, \boldsymbol{a}, b) \in \mathbb{R}^d \times \mathbb{R}^2 \times \mathbb{R}$ are bounded. Then, with probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, there exists a subset $\mathcal{A} \subseteq [n]$ with cardinality at least $n(1 - o(1))$ such that for all $i \in \mathcal{A}$ the following hold:*

1. *There is a subset $J_{i,0} \subseteq N_i \cap C_0$ with cardinality at least $\frac{9}{10}|N_i \cap C_0|$, such that $\gamma_{ij} = \Theta(1/|N_i|)$ for all $j \in J_{i,0}$.*

2. *There is a subset $J_{i,1} \subseteq N_i \cap C_1$ with cardinality at least $\frac{9}{10}|N_i \cap C_1|$, such that $\gamma_{ij} = \Theta(1/|N_i|)$ for all $j \in J_{i,1}$.*

For $i \in [n]$ let us write $\mathbf{X}_i = (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma\boldsymbol{g}_i$ where $\boldsymbol{g}_i \sim N(0, \mathbf{I})$, $\epsilon_i = 0$ if $i \in C_0$ and $\epsilon_i = 1$ if $i \in C_1$. Moreover, since the parameters $(\boldsymbol{w}, \boldsymbol{a}, b) \in \mathbb{R}^d \times \mathbb{R}^2 \times \mathbb{R}$ are bounded, we can write $\boldsymbol{w} = R\hat{\boldsymbol{w}}$ and $\boldsymbol{a} = R'\hat{\boldsymbol{a}}$ such that $\|\hat{\boldsymbol{w}}\| = 1$ and $\|\hat{\boldsymbol{a}}\| = 1$ and R, R' are some constants. We define the following sets which will become useful later in our computation of γ_{ij}'s. Define

$$\mathcal{A} \stackrel{\text{def}}{=} \left\{ i \in [n] \ \middle| \ \begin{array}{l} |\hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i| \leq 10\sqrt{\log(n(p+q))}, \text{ and} \\ |\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_j| \leq 10\sqrt{\log(n(p+q))}, \ \forall j \in N_i \end{array} \right\}.$$

For $i \in [n]$ define

$$J_{i,0} \stackrel{\text{def}}{=} \left\{ j \in N_i \cap C_0 \mid |\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_j| \leq \sqrt{10} \right\},$$

$$J_{i,1} \stackrel{\text{def}}{=} \left\{ j \in N_i \cap C_1 \mid |\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_j| \leq \sqrt{10} \right\},$$

$$B_{i,0}^t \stackrel{\text{def}}{=} \left\{ j \in N_i \cap C_0 \mid 2^{t-1} \leq \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_j \leq 2^t \right\}, \ t = 1, 2, \ldots, T,$$

$$B_{i,1}^t \stackrel{\text{def}}{=} \left\{ j \in N_i \cap C_1 \mid 2^{t-1} \leq \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_j \leq 2^t \right\}, \ t = 1, 2, \ldots, T,$$

where $T \stackrel{\text{def}}{=} \left\lceil \log_2 \left(10\sqrt{\log(n(p+q))} \right) \right\rceil$.

We start with a few claims about the sizes of these sets.

Claim 20. *With probability at least $1 - o(1)$, we have that $|\mathcal{A}| \geq n(1 - o(1))$.*

Proof Because $|\hat{\boldsymbol{a}}_2| \leq 1$ we know that $\mathcal{A}$ is a superset of $\mathcal{A}'$ where

$$\mathcal{A}' \stackrel{\text{def}}{=} \left\{ i \in [n] \ \middle| \ \begin{array}{l} |\hat{\boldsymbol{w}}^T \boldsymbol{g}_i| \leq 10\sqrt{\log(n(p+q))}, \text{ and} \\ |\hat{\boldsymbol{w}}^T \boldsymbol{g}_j| \leq 10\sqrt{\log(n(p+q))}, \ \forall j \in N_i \end{array} \right\}.$$

We give a lower bound for $|\mathcal{A}'|$ and hence prove the result. First of all, note that if $p + q \geq \Omega(1/\log^2 n)$, then $\log(n(p+q)) = \log n(1 - o(1))$ and we easily get that with probability at least $1 - o(1)$, $|\hat{\boldsymbol{w}}^T \boldsymbol{g}_i| \leq 10\sqrt{\log(n(p+q))}$ for all $i \in [n]$, and thus $|\mathcal{A}| = |\mathcal{A}'| = n$. Therefore

34

let us assume without loss of generality that $p + q \leq O(1/\log^2 n)$. Consider the following sum of indicator random variables

$$S \overset{\text{def}}{=} \sum_{i \in [n]} \mathbf{1}_{\left\{|\hat{\boldsymbol{w}}^T \boldsymbol{g}_i| \geq 10\sqrt{\log(n(p+q))}\right\}}.$$

By the multiplicative Chernoff bound, for any $\delta > 0$ we have

$$\mathbf{Pr}\left[S \geq nb(1 + \delta)\right] \leq \exp\left(-\frac{\delta^2}{2 + \delta}nb\right)$$

where $b \overset{\text{def}}{=} \mathbf{Pr}(|\hat{\boldsymbol{w}}^T \boldsymbol{g}_i| \geq 10\sqrt{\log(n(p + q))})$. Moreover, by the standard upper bound on the Gaussian tail probability (Proposition 2.1.2, Vershynin (2018)) we know that $b < e^{-50\log(n(p+q))}$. Let us set

$$\delta \overset{\text{def}}{=} \frac{1}{bn(p + q)\log n}.$$

Then by the upper bound on b and the assumption that $p, q = \Omega(\log^2 n/n)$ we know that

$$\delta \geq \frac{(n(p+q))^{49}}{\log n} \geq \Omega(\log^{97} n) = \omega(1).$$

It follows that

$$\frac{\delta^2}{2 + \delta}nb \geq \Omega(\delta nb) = \Omega\left(\frac{1}{(p+q)\log n}\right) \geq \Omega(\log n).$$

Therefore, with probability at least $1 - o(1)$ we have that

$$S \leq nb(1 + \delta) \leq \frac{n}{(n(p+q))^{50}} + \frac{n}{n(p+q)\log n} = O\left(\frac{n}{n(p+q)\log n}\right).$$

Apply the concentration result of node degrees, this means that with probability at least $1 - o(1)$,

$$\left|\left\{i \in [n] \mid |\hat{\boldsymbol{w}}^T \boldsymbol{g}_i| \geq 10\sqrt{\log(n(p+q))} \text{ or } \exists j \in N_i \text{ such that } |\hat{\boldsymbol{w}}^T \boldsymbol{g}_j| \geq 10\sqrt{\log(n(p+q))}\right\}\right|$$

$$\leq \ S \cdot \frac{n}{2}(p+q)(1 \pm o(1)) = O\left(\frac{n}{n(p+q)\log n}\right) \cdot \frac{n}{2}(p+q)(1 \pm o(1)) = O\left(\frac{n}{\log n}\right).$$

Therefore we have

$$|\mathcal{A}'| \geq n - O(n/\log n) = n(1 - o(1)).$$

■

Claim 21. *With probability at least $1 - o(1)$, we have that for all $i \in [n]$,*

$$|J_{i,0}| \geq \frac{9}{10}|N_i \cap C_0| \quad and \quad |J_{i,1}| \geq \frac{9}{10}|N_i \cap C_1|.$$

Proof We prove the result for $J_{i,0}$, the result for $J_{i,1}$ follows analogously. First, fix $i \in [n]$. For each $j \in |N_i \cap C_0|$ we have that

$$\mathbf{Pr}[|\hat{\boldsymbol{a}}_2 \boldsymbol{w}^T \boldsymbol{g}_j| \geq \sqrt{10}] \leq \mathbf{Pr}[|\boldsymbol{w}^T \boldsymbol{g}_j| \geq \sqrt{10}] \leq e^{-50}.$$

Denote $J_{i,0}^c \stackrel{\text{def}}{=} (N_i \cap C_0) \setminus J_{i,0}$. We have that

$$\mathbf{E}[|J_{i,0}^c|] = \mathbf{E}\left[\sum_{j \in N_i \cap C_0} \mathbf{1}_{\{|\hat{\boldsymbol{a}}_2 \boldsymbol{w}^T \boldsymbol{g}_j| \geq \sqrt{10}\}} \right] \leq e^{-50}|N_i \cap C_0|,$$

Apply Chernoff's inequality (Theorem 2.3.4 in Vershynin (2018)) we have

$$
\begin{aligned}
\mathbf{Pr}\left[|J_{i,0}^c| \geq \frac{1}{10}|N_i \cap C_0| \right] &\leq e^{-\mathbf{E}[|J_{i,0}^c|]} \left(\frac{e\,\mathbf{E}[|J_{i,0}^c|]}{|N_i \cap C_0|/10} \right)^{|N_i \cap C_0|/10} \\
&\leq \left(\frac{e e^{-50}|N_i \cap C_0|}{|N_i \cap C_0|/10} \right)^{|N_i \cap C_0|/10} \\
&= \exp\left(-\left(\frac{1}{2} - \frac{\log 10}{10} - \frac{1}{10} \right) |N_i \cap C_0| \right) \\
&\leq \exp\left(-\frac{4}{25}|N_i \cap C_0| \right).
\end{aligned}
$$

Apply the union bound we get

$$
\begin{aligned}
\mathbf{Pr}\left[|J_{i,0}| \geq \frac{9}{10}|C_0 \cap N_i|, \forall i \in [n] \right] &\geq 1 - \sum_{i \in [n]} \exp\left(-\frac{4}{25}|N_i \cap C_0| \right) \\
&\geq \mathbf{Pr}(\mathcal{E}_3) \cdot \left(1 - \sum_{i \in [n]} \exp\left(-\frac{4}{25}\frac{n \min(p,q)(1 - o(1))}{2} \right) \right) \\
&= (1 - o(1)) \cdot \left(1 - n \exp\left(-\frac{2n \min(p,q)(1 - o(1))}{25} \right) \right) \\
&= 1 - o(1).
\end{aligned}
$$

The second inequality follows because $|N_i \cap C_0| \geq \frac{n}{2}\min(p,q)(1 - o(1))$ under the event $\mathcal{E}_3$ (cf. Definition 17) for all $i \in [n]$. The last equality is due to our assumption that $p, q = \Omega(\frac{\log^2 n}{n})$. ∎

Claim 22. *With probability at least $1 - o(1)$, we have that for all $i \in [n]$ and for all $t \in [T]$,*

$$|B_{i,0}^t| \leq \mathbf{E}[|B_{i,0}^t|] + \sqrt{T}|N_i \cap C_0|^{\frac{4}{5}} \quad and \quad |B_{i,1}^t| \leq \mathbf{E}[|B_{i,1}^t|] + \sqrt{T}|N_i \cap C_1|^{\frac{4}{5}}.$$

Proof We prove the result for $B_{i,0}^t$, and the result for $B_{i,1}^t$ follows analogously. First fix $i \in [n]$ and $t \in [T]$. By the additive Chernoff inequality, we have

$$\mathbf{Pr}\left(|B_{i,0}^t| \geq \mathbf{E}[|B_{i,0}^t|] + |N_i \cap C_0| \cdot \sqrt{T}|N_i \cap C_0|^{-\frac{1}{5}} \right) \leq e^{-2T|N_i \cap C_0|^{3/5}}.$$

Taking a union bound over all $i \in [n]$ and $t \in [T]$ we get

$$\mathbf{Pr}\left[\bigcup_{i \in [n]} \bigcup_{t \in [T]} \left\{|B_{i,0}^t| \geq \mathbf{E}[|B_{i,0}^t|] + \sqrt{T}|N_i \cap C_0|^{\frac{4}{5}}\right\}\right]$$

$$\leq nT \exp\left(-2T\left(\frac{n}{2}\min(p,q)(1-o(1))\right)^{3/5}\right) + o(1) = o(1),$$

where the last equality follows from Assumption 1 that $p, q = \Omega(\frac{\log^2 n}{n})$, and hence

$$nT \exp\left(-2T\left(\frac{n}{2}\min(p,q)(1-o(1))\right)^{3/5}\right) = nT\exp\left(-\omega\left(\sqrt{2T}\log n\right)\right) = O\left(n^{-c}\right)$$

for some absolute constant $c > 0$. Moreover, we have used degree concentration, which introduced the additional additive $o(1)$ term in the probability upper bound. Therefore we have

$$\mathbf{Pr}\left[|B_{i,0}^t| \leq \mathbf{E}[|B_{i,0}^t|] + \sqrt{T}|N_i \cap C_0|^{\frac{4}{5}}, \forall i \in [n] \; \forall t \in [T]\right] \geq 1 - o(1).$$

$\blacksquare$

We start by defining an event $\mathcal{E}^{\#}$ which is the intersection of the following events over the randomness of $\mathbf{A}$ and $\{\epsilon_i\}_i$ and $\mathbf{X}_i = (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma \boldsymbol{g}_i$,

- $\mathcal{E}_1'$ is the event that for each $i \in [n]$, $|C_0 \cap N_i| = \frac{n}{2}((1 - \epsilon_i)p + \epsilon_i q)(1 \pm o(1))$ and $|C_1 \cap N_i| = \frac{n}{2}((1 - \epsilon_i)q + \epsilon_i p)(1 \pm o(1))$.

- $\mathcal{E}_2'$ is the event that $|\mathcal{A}| \geq n - o(\sqrt{n})$.

- $\mathcal{E}_3'$ is the event that $|J_{i,0}| \geq \frac{9}{10}|N_i \cap C_0|$ and $|J_{i,1}| \geq \frac{9}{10}|N_i \cap C_1|$ for all $i \in [n]$.

- $\mathcal{E}_4'$ is the event that $|B_{i,0}^t| \leq \mathbf{E}[|B_{i,0}^t|] + \sqrt{T}|N_i \cap C_0|^{\frac{4}{5}}$ and $|B_{i,1}^t| \leq \mathbf{E}[|B_{i,1}^t|] + \sqrt{T}|N_i \cap C_1|^{\frac{4}{5}}$ for all $i \in [n]$ and for all $t \in [T]$.

By Claims 20, 21, 22, we get that with probability at least $1 - o(1)$, the event $\mathcal{E}^{\#} \overset{\text{def}}{=} \bigcap_{i=1}^{4} \mathcal{E}_i'$ holds. We will show that under event $\mathcal{E}^{\#}$, for all $i \in \mathcal{A}$, for all $j \in J_{i,c}$ where $c \in \{0, 1\}$, we have $\gamma_{ij} = \Theta(1/|N_i|)$. This will prove Theorem 10.

Fix $i \in \mathcal{A}$ and some $j \in J_{i,0}$. Let us consider

$$\gamma_{ij} = \frac{\exp\left(\text{LeakyRelu}(\boldsymbol{a}_1\boldsymbol{w}^T\mathbf{X}_i + \boldsymbol{a}_2\boldsymbol{w}^T\mathbf{X}_j + b)\right)}{\sum_{k \in N_i} \exp\left(\text{LeakyRelu}(\boldsymbol{a}_1\boldsymbol{w}^T\mathbf{X}_i + \boldsymbol{a}_2\boldsymbol{w}^T\mathbf{X}_k + b)\right)}$$

$$= \frac{\exp\left(\sigma RR'\,\text{LeakyRelu}(\kappa_{ij} + \hat{\boldsymbol{a}}_1\hat{\boldsymbol{w}}^T\boldsymbol{g}_i + \hat{\boldsymbol{a}}_2\hat{\boldsymbol{w}}^T\boldsymbol{g}_j + b')\right)}{\sum_{k \in N_i} \exp\left(\sigma RR'\,\text{LeakyRelu}(\kappa_{ik} + \hat{\boldsymbol{a}}_1\hat{\boldsymbol{w}}^T\boldsymbol{g}_i + \hat{\boldsymbol{a}}_2\hat{\boldsymbol{w}}^T\boldsymbol{g}_k + b')\right)}$$

$$= \frac{1}{\sum_{k \in N_i} \exp(\Delta_{ik} - \Delta_{ij})}$$

37

where for $l \in N_i$, we denote

$$\kappa_{il} \overset{\text{def}}{=} (2\epsilon_i - 1)\hat{w}^T \mu/\sigma + (2\epsilon_l - 1)\hat{w}^T \mu/\sigma,$$

$$\Delta_{il} \overset{\text{def}}{=} \sigma R R' \text{ LeakyRelu}(\kappa_{il} + \hat{a}_1 w^T g_i + \hat{a}_2 w^T g_l + b'),$$

and $b = \sigma R R' b'$. We will show that

$$\sum_{k \in N_i} \exp(\Delta_{ik} - \Delta_{ij}) = \Theta(|N_i|)$$

and hence conclude that $\gamma_{ij} = \Theta(1/|N_i|)$. First of all, note that since $\|\mu\| \le K\sigma$ for some absolute constant K, we know that

$$|\kappa_{il}| \le \sqrt{2}K = O(1).$$

Let us assume that $\hat{a}_1 \hat{w}^T g_i \ge 0$ and consider the following two cases regarding the magnitude of $\hat{a}_1 \hat{w}^T g_i$.

  Case 1. If $\kappa_{ij} + \hat{a}_1 \hat{w}^T g_i + \hat{a}_2 \hat{w}^T g_j + b' < 0$, then

$$\begin{aligned}
\Delta_{ik} - \Delta_{ij} &= \sigma R R' \Big(\text{LeakyRelu}(\kappa_{ik} + \hat{a}_1 \hat{w}^T g_i + \hat{a}_2 \hat{w}^T g_k + b') \\
&\quad - \text{LeakyRelu}(\kappa_{ij} + \hat{a}_1 \hat{w}^T g_i + \hat{a}_2 \hat{w}^T g_j + b') \Big) \\
&= \sigma R R' \Big(\text{LeakyRelu}(\hat{a}_1 \hat{w}^T g_i + \hat{a}_2 \hat{w}^T g_k \pm O(1)) \\
&\quad - \beta(\kappa_{ij} + \hat{a}_1 \hat{w}^T g_i + \hat{a}_2 \hat{w}^T g_j + b') \Big) \\
&= \sigma R R' \left(\text{LeakyRelu}(\hat{a}_2 \hat{w}^T g_k \pm O(1)) \pm O(1) \right) \\
&= \sigma R R' \left(\Theta(\hat{a}_2 \hat{w}^T g_k) \pm O(1) \right),
\end{aligned}$$

where β is the slope of LeakyRelu(x) for $x < 0$. Here, the second equality follows from $|\kappa_{ik} + b'| \le \sqrt{2}K + |b'| = O(1)$ and $\kappa_{ij} + \hat{a}_1 \hat{w}^T g_i + \hat{a}_2 \hat{w}^T g_j + b' < 0$. The third equality follows from

- We have $j \in J_{i,0}$ and hence $|\hat{a}_2 \hat{w}^T g_j| = O(1)$;

- We have $\kappa_{ij} + \hat{a}_1 \hat{w}^T g_i + \hat{a}_2 \hat{w}^T g_j + b' < 0$, so $\hat{a}_1 \hat{w}^T g_i < |\kappa_{ij}| + |\hat{a}_2 \hat{w}^T g_j| + |b'| = O(1)$, moreover, because $\hat{a}_1 \hat{w}^T g_i \ge 0$, we get that $|\hat{a}_1 \hat{w}^T g_i| = O(1)$;

- We have $|\kappa_{ij} + \hat{a}_1 \hat{w}^T g_i + \hat{a}_2 \hat{w}^T g_j + b'| \le |\hat{a}_1 \hat{w}^T g_i| + |\hat{a}_2 \hat{w}^T g_j| + |\kappa_{ij} + b'| = O(1) + O(1) + O(1) = O(1)$.

38

<u>Case 2.</u> If $\kappa_{ij} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_j + b' \geq 0$, then

$$
\begin{aligned}
\Delta_{ik} - \Delta_{ij} &= \sigma R R' \Big(\text{LeakyRelu}(\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b') \\
&\qquad - \text{LeakyRelu}(\kappa_{ij} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_j + b') \Big) \\
&= \sigma R R' \Big(\text{LeakyRelu}(\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b') \\
&\qquad - \kappa_{ij} - \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i - \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_j - b' \Big) \\
&= \sigma R R' \left(\text{LeakyRelu}(\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b') - \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \pm O(1) \right) \\
&\begin{cases} = \sigma R R' \left(\Theta(\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k) \pm O(1) \right), & \text{if } k \in J_{i,0} \cup J_{i,1} \\ \leq \sigma R R' \left(O(\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k) \pm O(1) \right), & \text{otherwise.} \end{cases}
\end{aligned}
$$

To see the last (in)equality in the above, consider the following cases:

1. If $k \in J_{i,0} \cup J_{i,1}$, then there are two cases depending on the sign of $\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b'$.

   - If $\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b' \geq 0$, then we have that

   $$
   \begin{aligned}
   &\text{LeakyRelu}(\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b') - \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \pm O(1) \\
   &= \kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b' - \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \pm O(1) \\
   &= \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + \kappa_{ik} + b' \pm O(1) \\
   &= \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k \pm O(1).
   \end{aligned}
   $$

   - If $\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b' < 0$, then because $\hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \geq 0$ and $|\kappa_{ik} + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b'| \leq |\kappa_{ik}| + |\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k| + |b'| = O(1)$, we know that $\hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i < |\kappa_{ik}| + |\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k| + |b'| = O(1)$ and $|\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b'| = O(1)$. Therefore it follows that

   $$
   \begin{aligned}
   &\text{LeakyRelu}(\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b') - \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \pm O(1) \\
   &= \text{LeakyRelu}(\pm O(1)) - O(1) \pm O(1) \\
   &= \pm O(1) \\
   &= \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k \pm O(1)
   \end{aligned}
   $$

   where the last equality is due to the fact that $k \in J_{i,0} \cup J_{i,1}$ so $|\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k| = O(1)$.

2. If $k \notin J_{i,0} \cup J_{i,1}$, then there are two cases depending on the sign of $\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b'$.

   - If $\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b' \geq 0$, then we have that

   $$
   \begin{aligned}
   &\text{LeakyRelu}(\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b') - \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \pm O(1) \\
   &= \kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b' - \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \pm O(1) \\
   &= \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + \kappa_{ik} + b' \pm O(1) \\
   &= \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k \pm O(1).
   \end{aligned}
   $$

- If $\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b' < 0$, then we have that,

$$
\begin{aligned}
&\text{LeakyRelu}(\kappa_{ik} + \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + b') - \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \pm O(1) \\
&= \quad \beta \kappa_{ik} + \beta \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i + \beta \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k + \beta b' - \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \pm O(1) \\
&= \quad \beta \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k - (1 - \beta) \hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \pm O(1) \\
&\leq \quad \beta \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k \pm O(1),
\end{aligned}
$$

where β is the slope of $\text{LeakyRelu}(\cdot)$.

Combining the two cases regarding the magnitude of $\hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i$ and our assumption that $\sigma, R, R = O(1)$, so far we have showed that, for any i such that $\hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i \geq 0$, for all $j \in J_{i,0}$, we have

$$
\Delta_{ik} - \Delta_{ij} = \begin{cases} \Theta(\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k) \pm O(1), & \text{if } k \in J_{i,0} \cup J_{i,1} \\ O(\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k) \pm O(1), & \text{otherwise.} \end{cases} \tag{16}
$$

By following a similar argument, one can show that Equation 16 holds for any i such that $\hat{\boldsymbol{a}}_1 \hat{\boldsymbol{w}}^T \boldsymbol{g}_i < 0$.

Let us now compute

$$
\sum_{k \in N_i} \exp(\Delta_{ik} - \Delta_{ij}) = \sum_{k \in N_i \cap C_0} \exp(\Delta_{ik} - \Delta_{ij}) + \sum_{k \in N_i \cap C_1} \exp(\Delta_{ik} - \Delta_{ij})
$$

for some $j \in J_{i,0}$. Let us focus on $\sum_{k \in N_i \cap C_0} \exp(\Delta_{ik} - \Delta_{ij})$ first. We will show that $\Omega(|N_i \cap C_0|) \leq \sum_{k \in N_i \cap C_0} \exp(\Delta_{ik} - \Delta_{ij}) \leq O(|N_i|)$.

First of all, we have that

$$
\begin{aligned}
\sum_{k \in N_i \cap C_0} \exp(\Delta_{ik} - \Delta_{ij}) &\geq \sum_{k \in J_{i,0}} \exp(\Delta_{ik} - \Delta_{ij}) = \sum_{k \in J_{i,0}} \exp\left(\Theta(\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k) \pm O(1)\right) \\
&\geq \sum_{k \in J_{i,0}} e^{c_1} = |J_{i,0}| e^{c_1} = \Omega(|N_i \cap C_0|),
\end{aligned} \tag{17}
$$

where c_1 is an absolute constant (possibly negative). On the other hand, consider the following partition of $N_i \cap C_0$:

$$
\begin{aligned}
P_1 &\stackrel{\text{def}}{=} \{k \in N_i \cap C_0 \mid \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k \leq 1\}, \\
P_2 &\stackrel{\text{def}}{=} \{k \in N_i \cap C_0 \mid \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k \geq 1\}.
\end{aligned}
$$

It is easy to see that

$$
\sum_{k \in P_1} \exp(\Delta_{ik} - \Delta_{ij}) \leq \sum_{k \in P_1} \exp\left(O(\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k) \pm O(1)\right) \leq \sum_{k \in P_1} e^{c_2} = |P_1| e^{c_2} = O(|N_i \cap C_0|),
$$

$$
\tag{18}
$$

where c_2 is an absolute constant. Moreover, because $i \in \mathcal{A}$ we have that $P_2 \subseteq \bigcup_{t \in [T]} B_{i,0}^t$. It follows that

$$\sum_{k \in P_2} \exp(\Delta_{ik} - \Delta_{ij}) = \sum_{t \in [T]} \sum_{k \in B_{i,0}^t} \exp(\Delta_{ik} - \Delta_{ij})$$

$$\leq \sum_{t \in [T]} \sum_{k \in B_{i,0}^t} \exp\left(O(\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k) \pm O(1)\right) \tag{19}$$

$$\leq \sum_{t \in [T]} |B_{i,0}^t| e^{c_3 2^t},$$

where c_3 is an absolute constant. We can upper bound the above quantity as follows. Under the Event $\mathcal{E}^*$, we have that

$$|B_{i,0}^t| \leq m_t + \sqrt{T} |N_i \cap C_0|^{\frac{4}{5}}, \text{ for all } t \in [T],$$

where

$$m_t \stackrel{\text{def}}{=} \mathbf{E}[|B_{i,0}^t|] = \sum_{k \in N_i \cap C_0} \mathbf{Pr}(2^{t-1} \leq \hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k \leq 2^t) \leq \sum_{k \in N_i \cap C_0} \mathbf{Pr}[\hat{\boldsymbol{a}}_2 \hat{\boldsymbol{w}}^T \boldsymbol{g}_k \geq 2^{t-1}]$$

$$\leq \sum_{k \in N_i \cap C_0} \mathbf{Pr}[\hat{\boldsymbol{w}}^T \boldsymbol{g}_k \geq 2^{t-1}] \leq |N_i \cap C_0| e^{-2^{2t-3}}.$$

It follows that

$$\sum_{t \in [T]} |B_{i,0}^t| e^{c_3 2^t} \leq \sum_{t \in [T]} \left(|N_i \cap C_0| e^{-2^{2t-3}} + \sqrt{T} |N_i \cap C_0|^{\frac{4}{5}}\right) e^{c_3 2^t}$$

$$\leq |N_i \cap C_0| \sum_{t=1}^{\infty} e^{-2^{2t-3}} e^{c_3 2^t} + \sum_{t \in [T]} \sqrt{T} |N_i \cap C_0|^{\frac{4}{5}} e^{c_3 2^T} \tag{20}$$

$$\leq c_4 |N_i \cap C_0| + o(|N_i|)$$

$$\leq O(|N_i|),$$

where c_4 is an absolute constant. The third inequality in the above follows from

- The series $\sum_{t=1}^{\infty} e^{-2^{2t-3}} e^{c_3 2^t}$ converges absolutely for any constant c_3;

- The sum $\sum_{t \in [T]} \sqrt{T} |N_i \cap C_0|^{\frac{4}{5}} e^{c_3 2^T} = T^{\frac{3}{2}} |N_i \cap C_0|^{\frac{4}{5}} e^{c_3 2^T} = o(|N_i|)$ because

$$\log\left(T^{\frac{3}{2}} e^{c_3 2^T}\right) = \frac{3}{2} \log\left[\log_2\left(10\sqrt{\log(n(p+q))}\right)\right] + c_3 2^{\left\lceil \log_2\left(10\sqrt{\log(n(p+q))}\right)\right\rceil}$$

$$\leq \frac{3}{2} \log\left[\log_2\left(10\sqrt{\log(n(p+q))}\right)\right] + 20c_3 \sqrt{\log(n(p+q))}$$

$$\leq O\left(\frac{1}{c} \log(n(p+q))\right),$$

for any $c > 0$. In particular, by picking $c > 5$ we see that $T^{\frac{3}{2}} e^{c_3 2^T} \leq O((n(p+q))^{\frac{1}{c}}) \leq o(|N_i|^{\frac{1}{5}})$, and hence we get $T^{\frac{3}{2}} e^{c_3 2^T} |N_i \cap C_0|^{\frac{4}{5}} \leq |N_i|^{\frac{4}{5}} \cdot o(|N_i|^{\frac{1}{5}}) = o(|N_i|)$.

Combining Equations 19 and 20 we get

$$\sum_{k \in P_2} \exp(\Delta_{ik} - \Delta_{ij}) \leq O(|N_i|), \tag{21}$$

and combining Equations 18 and 21 we get

$$\sum_{k \in N_i \cap C_0} \exp(\Delta_{ik} - \Delta_{ij}) = \sum_{k \in P_1} \exp(\Delta_{ik} - \Delta_{ij}) + \sum_{k \in P_1} \exp(\Delta_{ik} - \Delta_{ij}) \leq O(|N_i|). \tag{22}$$

Now, by Equations 17 and 22 we get

$$\Omega(|N_i \cap C_0|) \leq \sum_{k \in N_i \cap C_0} \exp(\Delta_{ik} - \Delta_{ij}) \leq O(|N_i|). \tag{23}$$

It turns out that repeating the same argument for $\sum_{k \in N_i \cap C_1} \exp(\Delta_{ik} - \Delta_{ij})$ yields

$$\Omega(|N_i \cap C_1|) \leq \sum_{k \in N_i \cap C_1} \exp(\Delta_{ik} - \Delta_{ij}) \leq O(|N_i|). \tag{24}$$

Finally, Equations 23 and 24 give us

$$\sum_{k \in N_i} \exp(\Delta_{ik} - \Delta_{ij}) = \Theta(|N_i|),$$

which readily implies

$$\gamma_{ij} = \frac{1}{\sum_{k \in N_i} \exp(\Delta_{ik} - \Delta_{ij})} = \Theta(1/|N_i|)$$

as required. We have showed that for all $i \in \mathcal{A}$ and for all $j \in J_{i,0}$, $\gamma_{ij} = \Theta(1/|N_i|)$. Repeating the same argument we get that the same result holds for all $i \in \mathcal{A}$ and for all $j \in J_{i,1}$, too. Hence, by Claims 20 and 21 about the cardinalities of $\mathcal{A}$, $J_{i,0}$ and $J_{i,1}$ we have thus proved Theorem 10.

A.7 Proof of Proposition 13

We restate Proposition 13 for convenience.

Proposition 23. *Suppose that p, q satisfy Assumption 1 and that p, q are bounded away from 1. For every $\epsilon > 0$, there are absolute constants $M, M' = O(\sqrt{\epsilon})$ such that, with probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, using the graph attention convolution in (2) and the attention architecture $\tilde{\Psi}$ in (13), the model misclassifies at least $\Omega(n^{1-\epsilon})$ nodes for any $\boldsymbol{w}$ such that $\|\boldsymbol{w}\| = 1$, if*

1. *$t = O(1)$ and $\|\boldsymbol{\mu}\| \leq M\sigma\sqrt{\frac{\log n}{n(p+q)}}(1 - \max(p,q))\frac{p+q}{|p-q|}$;*

2. *$t = \omega(1)$ and $\|\boldsymbol{\mu}\| \leq M'\sigma\sqrt{\frac{\log n}{n(p+q)}}(1 - \max(p,q))$.*

We start with part 1 of the proposition. Let us assume that $p \geq q$. The result when $p < q$ follows analogously. We will condition on the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ defined in Definition 17. These events are concerned with the concentration of class sizes $|C_0|$ and $|C_1|$ and the concentration of the number of intra-class and inter-class edges, i.e. $|C_0 \cap N_i|$ and $C_1 \cap N_i|$ for all i. By Lemma 18, the probability that these events hold simultaneously is at least $1 - o(1)$. Fix any $\boldsymbol{w} \in \mathbb{R}^d$ such that $\|\boldsymbol{w}\| = 1$. Without loss of generality, assume that $\boldsymbol{w}^T \boldsymbol{\mu} > 0$. Because $t = O(1)$, by the definition of $\tilde{\Psi}$ in (13) and the attention coefficients in (1) we have that

$$\gamma_{ij} = \begin{cases} \frac{c_1}{n(p+q)}(1 \pm o(1)), & \text{if } (i,j) \text{ is an intra-class edge,} \\ \frac{c_2}{n(p+q)}(1 \pm o(1)), & \text{if } (i,j) \text{ is an inter-class edge,} \end{cases} \tag{25}$$

for some positive constants $c_1 \geq 1$ and $c_2 \leq 1$. Let us write $\mathbf{X}_i = (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma \boldsymbol{g}_i$ where $\boldsymbol{g}_i \sim N(0, \mathbf{I})$, $\epsilon_i = 0$ if $i \in C_0$ and $\epsilon_i = 1$ if $i \in C_1$. Let us consider the classification of nodes in C_0 based on the model output $\sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T X_j$ for node i. Note that, depending on whether $i \in C_0$ or $i \in C_1$, the expectation of the model output is symmetric around 0, and therefore we consider the decision boundary at 0. Let $0 < \epsilon < 1$ and fix any partition of C_0 into a set of disjoint subsets,

$$C_0 = \bigcup_{h=1}^{\ell} C_0^{(h)}$$

such that $\ell = |C_0|^{1-\epsilon}$ and $|C_0^{(h)}| = |C_0|^{\epsilon}$ for every $h = 1, 2, \ldots, \ell$. In what follows we will consider the classification of nodes in each $C_0^{(h)}$ separately. We will show that, with probability at least $1 - o(1)$, for each one of more than half of the subsets $C_0^{(h)}$ where $h = 1, 2, \ldots, \ell$, the model is going to misclassify at least one node, and consequently, the model is going to misclassify at least $\ell/2 = |C_0|^{1-\epsilon}/2 = \Omega(n^{1-\epsilon})$ nodes, giving the required result on misclassification rate.

Fix any $h' \in \{1, 2, \ldots, \ell\}$. Using (25) we get that, for large enough n, the event that the model correctly classifies all nodes in $C_0^{(h)}$ satisfies

$$\left\{ \max_{i \in C_0^{(h')}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \mathbf{X}_j < 0 \right\}$$

$$= \left\{ \max_{i \in C_0^{(h')}} \left(\sum_{j \in N_i \cap C_1} \gamma_{ij} - \sum_{j \in N_i \cap C_0} \gamma_{ij} \right) \boldsymbol{w}^T \boldsymbol{\mu} + \sigma \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \boldsymbol{g}_j < 0 \right\}$$

$$\subseteq \left\{ c_3 \left(\frac{q-p}{p+q} \right) \boldsymbol{w}^T \boldsymbol{\mu} + \sigma \max_{i \in C_0^{(h')}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \boldsymbol{g}_j < 0 \right\}$$

for some absolute constant $c_3 > 0$, and hence the probability that the model correctly classifies all nodes in $C_0^{(h')}$ satisfies, for large enough n,

$$
\mathbf{Pr}\left(\max_{i \in C_0^{(h')}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \mathbf{X}_j < 0\right) \leq \mathbf{Pr}\left(\max_{i \in C_0^{(h')}} \sum_{j \in N_i} \boldsymbol{w}^T \boldsymbol{g}_j < c_3 \left(\frac{p-q}{p+q}\right) \frac{|\boldsymbol{w}^T \boldsymbol{\mu}|}{\sigma}\right)
$$

$$
\leq \mathbf{Pr}\left(\max_{i \in C_0^{(h')}} \sum_{j \in N_i} \boldsymbol{w}^T \boldsymbol{g}_j < \tilde{M} \sqrt{\frac{\log n}{n(p+q)}(1 - \max(p,q))}\right)
$$

where the last inequality follows from our assumption on $\|\boldsymbol{\mu}\|$ and we denote $\tilde{M} \overset{\text{def}}{=} c_3 M > 0$. Now we will use Sudakov's minoration inequality (Vershynin, 2018) to obtain a lower bound on the expected maximum, 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 metric over the node index set $[n]$. Let $\boldsymbol{z}_i \overset{\text{def}}{=} \sum_{j \in N_i} \boldsymbol{w}^T \boldsymbol{g}_j$. For $i, j \in C_0$, $i \neq j$, consider the canonical metric $d_\circ(i,j)$ given by

$$
\begin{aligned}
d_\circ(i,j)^2 &\overset{\text{def}}{=} \mathbf{E}[(\boldsymbol{z}_i - \boldsymbol{z}_j)^2] \\
&= \sum_{k \in N_i} \gamma_{ik}^2 + \sum_{k \in N_j} \gamma_{jk}^2 - 2 \sum_{k \in N_i \cap N_j} \gamma_{ik}\gamma_{jk} \\
&\geq c_4 \sum_{k \in J_{ij}} \frac{1}{n^2(p+q)^2} \\
&= \frac{c_4|J_{ij}|}{n^2(p+q)^2},
\end{aligned}
$$

where $J_{ij} \overset{\text{def}}{=} (N_i \cup N_j) \setminus (N_i \cap N_j)$ is the symmetric difference of the neighbors of i and j, $c_4 > 0$ is an absolute constant, and the inequality is due to (25). We lower bound $|J_{ij}|$ as follows. For $i, j \in C_0$, $i \neq j$, and a node $k \in [n]$, the probability that k is a neighbor of exactly one of i and j is $2p(1-p)$ if $k \in C_0$ and $2q(1-q)$ if $k \in C_1$. Therefore we have $\mathbf{E}[|J_{ij}|] = n(p(1-p) + q(1-q))$. It follows from the multiplicative Chernoff bound that for any $0 < \delta < 1$,

$$
\mathbf{Pr}[|J_{ij}| < \mathbf{E}[|J_{ij}|](1-\delta)] \leq \exp(-\delta^2 \mathbf{E}[|J_{ij}|]/3).
$$

Choose

$$
\delta = 3\sqrt{\frac{\log n}{\mathbf{E}[|J_{ij}|]}} = 3\sqrt{\frac{\log n}{n(p(1-p) + q(1-q))}} = o(1),
$$

where the last equality follows from $n(p(1-p) + q(1-q)) = \Omega(\log^2 n)$, which is in turn due to the assumptions that $p, q = \Omega(\log^2 n/n)$ and p, q are bounded away from 1. Apply a union bound over all $i, j \in C_0$, we get that with probability at least $1 - o(1)$, the size of J_{ij} satisfies

$$
|J_{ij}| \geq n(p(1-p) + q(1-q))(1 - o(1)), \forall i, j \in C_0. \tag{26}
$$

Therefore it follows that, for large enough n,

$$
d_\circ(i,j) \geq \sqrt{\frac{c_4|J_{ij}|}{n^2(p+q)^2}} = \sqrt{\frac{c_4 n(p(1-p) + q(1-q))(1 - o(1))}{n^2(p+q)^2}} = \Omega\left(\sqrt{\frac{1 - \max(p,q)}{n(p+q)}}\right).
$$

44

We condition on the event that the inequality (26) holds for all $i, j \in C_0$, which happens with probability at least $1 - o(1)$. Apply Sudakov's minoration with metric $d_\circ(i, j)$, we get that for large enough n, for all $h = 1, 2, \ldots, \ell$,

$$\mathbf{E}\left[\max_{i \in C_0^{(h)}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \boldsymbol{g}_j\right] \geq c_5 d_\circ(i, j) \sqrt{\log |C_0^{(h)}|} \geq c_6 \sqrt{\epsilon} \sqrt{\frac{\log n}{n(p+q)}} (1 - \max(p, q)) \quad (27)$$

for some absolute constants $c_5, c_6 > 0$. The last inequality in the above follows from $|C_0^{(h)}| = |C_0|^\epsilon = \Omega(n^\epsilon)$. In addition, note that since by assumption Ψ is independent from the node features, using (25) we have that $\sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \boldsymbol{g}_j$ is Gaussian with variance $O(\frac{1}{n(p+q)})$. We use Borell's inequality (Adler and Taylor (2007) Chapter 2) to get that for any $t > 0$ and large enough n,

$$\mathbf{Pr}\left[\max_{i \in C_0^{(h')}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \boldsymbol{g}_j < \mathbf{E}\left[\max_{i \in C_0^{(h')}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \boldsymbol{g}_j\right] - t\right] \leq \exp(-c_7 t^2 n(p+q)).$$

for some absolute constant $c_7 > 0$. By the lower bound of the expectation (27), we get that

$$\mathbf{Pr}\left[\max_{i \in C_0^{(h')}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \boldsymbol{g}_j < c_6 \sqrt{\epsilon} \sqrt{\frac{\log n}{n(p+q)}} (1 - \max(p, q)) - t\right] \leq \exp(-c_7 t^2 n(p+q)).$$

Now, let $M > 0$ be any constant that also satisfies $M < c_6 \sqrt{\epsilon}/c_3$. Recall that we defined $\tilde{M} = c_3 M$, and hence $\tilde{M} < c_6 \sqrt{\epsilon}$. Set

$$t = (c_6 \sqrt{\epsilon} - \tilde{M}) \sqrt{\frac{\log n}{n(p+q)}} (1 - \max(p, q)) = \Omega\left(\sqrt{\frac{\log n}{n(p+q)}} (1 - \max(p, q))\right), \quad (28)$$

then combine with the events we have conditioned so far we get

$$\mathbf{Pr}\left[\max_{i \in C_0^{(h')}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \boldsymbol{g}_j \leq \tilde{M} \sqrt{\frac{\log n}{n(p+q)}} (1 - \max(p, q))\right] = o(1).$$

Recall that the above probability is the probability of correctly classifying all nodes in $C_0^{(h')}$. Since our choice of h' was arbitrary, this applies to every $h \in \{1, 2, \ldots, h\}$. Let $I_0^{(h)}$ denote the indicator variable of the event that there is at least one node in $C_0^{(h)}$ that is misclassified, then

$$\mathbf{E}\left[\sum_{h=1}^\ell I_0^{(h)}\right] = \ell \cdot \mathbf{E}[I_0^{(1)}] = \ell \cdot (1 - o(1)) = \ell - o(\ell).$$

Apply the reverse Markov inequality, we get

$$\mathbf{Pr}\left[\sum_{h=1}^\ell I_0^{(h)} \leq \frac{1}{2}\ell\right] \leq \frac{\ell - \mathbf{E}\left[\sum_{h=1}^\ell I_0^{(h)}\right]}{\ell - \frac{1}{2}\ell} = \frac{o(\ell)}{\frac{1}{2}\ell} = o(1).$$

Therefore, with probability at least $1 - o(1)$, we have $\sum_{h=1}^{\ell} I_0^{(h)} \geq \frac{1}{2}\ell = \Omega(n^{1-\epsilon})$. This implies the required result that the model misclassifies at least $\Omega(n^{1-\epsilon})$ nodes.

The proof of part 2 is similar to the proof of part 1. Let us assume that $p \geq q$ since the result when $p < q$ can be proved analogously. We condition on the events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ defined in Definition 17 which simultaneous hold with probability at least $1 - o(1)$ by Lemma 18. Fix any $\boldsymbol{w} \in \mathbb{R}^d$ such that $\|\boldsymbol{w}\| = 1$. Because $t = \omega(1)$, by the definition of $\tilde{\Psi}$ in (13) and the attention coefficients in (1) we have that

$$\gamma_{ij} = \begin{cases} \frac{2}{np}(1 \pm o(1)), & \text{if } (i,j) \text{ is an intra-class edge,} \\ o(\frac{1}{n(p+q)}), & \text{if } (i,j) \text{ is an inter-class edge.} \end{cases} \tag{29}$$

Write $\mathbf{X}_i = (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma \boldsymbol{g}_i$ where $\boldsymbol{g}_i \sim N(0, \mathbf{I})$, $\epsilon_i = 0$ if $i \in C_0$ and $\epsilon_i = 1$ if $i \in C_1$. We consider the classification of nodes in C_0 based on the model output $\sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T X_j$ for node i and the decision boundary at 0. As before, let $0 < \epsilon < 1$ and fix any partition of C_0 into a set of disjoint subsets $C_0 = \bigcup_{h=1}^{\ell} C_0^{(h)}$ such that $\ell = |C_0|^{1-\epsilon}$ and $|C_0^{(h)}| = |C_0|^{\epsilon}$ for every $h = 1, 2, \ldots, \ell$. We proceed to show that, with high probability, for each one of more than half of the subsets $C_0^{(h)}$ where $h = 1, 2, \ldots, \ell$, the model is going to misclassify at least one node, and consequently, the model is going to misclassify at least $\ell/2 = \Omega(n^{1-\epsilon})$ nodes as required. Fix any $h' \in \{1, 2, \ldots, \ell\}$. Using (29) we get that, for large enough n, the event that the model correctly classifies all nodes in $C_0^{(h')}$ satisfies

$$\left\{ \max_{i \in C_0^{(h')}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \mathbf{X}_j < 0 \right\} \subseteq \left\{ c_1 \boldsymbol{w}^T \boldsymbol{\mu} + \sigma \max_{i \in C_0^{(h')}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \boldsymbol{g}_j < 0 \right\}$$

for some absolute constant $c_1 > 0$, and hence the probability that the model classifies all nodes in $C_0^{h'}$ correctly satisfies, for large enough n,

$$\mathbf{Pr}\left(\max_{i \in C_0^{(h')}} \sum_{j \in N_i} \gamma_{ij} \boldsymbol{w}^T \mathbf{X}_j < 0 \right) \leq \mathbf{Pr}\left(\max_{i \in C_0^{(h')}} \sum_{j \in N_i} \boldsymbol{w}^T \boldsymbol{g}_j < c_1 \frac{|\boldsymbol{w}^T \boldsymbol{\mu}|}{\sigma} \right)$$

$$\leq \mathbf{Pr}\left(\max_{i \in C_0^{(h')}} \sum_{j \in N_i} \boldsymbol{w}^T \boldsymbol{g}_j < \tilde{M} \sqrt{\frac{\log n}{n(p+q)}(1 - \max(p,q))} \right)$$

where the last inequality follows from our assumption on $\|\boldsymbol{\mu}\|$ and we denote $\tilde{M} \stackrel{\text{def}}{=} c_1 M' > 0$. The rest of the proof of part 2 proceeds as the proof of part 1.

A.8 Proof of Theorem 14

We restate Theorem 14 for convenience. Recall that we write $\|\boldsymbol{\mu}\| = \kappa\sigma$ for some $\kappa > 0$.

Theorem. *With probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim CSBM(n, p, q, \boldsymbol{\mu}, \sigma^2)$, using the two-layer MLP attention architecture Ψ given in (3) and (4) with $R = \Omega(n \log^2 n\sigma)$, the graph attention convolution output satisfies*

$$h_i' = \sum_{j \in N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j > 0 \text{ if and only if } \tilde{\boldsymbol{w}}^T \mathbf{X}_i > 0, \ \forall i \in [n],$$

$$h_i' = \sum_{j \in N_i} \gamma_{ij} \tilde{\boldsymbol{w}}^T \mathbf{X}_j < 0 \text{ if and only if } \tilde{\boldsymbol{w}}^T \mathbf{X}_i < 0, \ \forall i \in [n].$$

We will condition on the following two events concerning the values of $\tilde{\boldsymbol{w}}^T\mathbf{X}_i$ for $i \in [n]$. Both events hold with probability at least $1 - o(1)$. First, for $\epsilon > 0$ consider the event

$$\left\{\tilde{\boldsymbol{w}}^T\mathbf{X}_i \notin [-\epsilon\sigma\max\{1,\kappa\}, \epsilon\sigma\max\{1,\kappa\}] \text{ for all } i \in [n]\right\}.$$

Since $\tilde{\boldsymbol{w}} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ and $\|\boldsymbol{\mu}\| = \kappa\sigma$, each $\tilde{\boldsymbol{w}}^T\mathbf{X}_i$ follows a normal distribution with mean $\epsilon_i\kappa\sigma$ and variance σ^2 (recall that $\epsilon_i \in \{0,1\}$ generates the node memberships in CSBM), we have that

$$\mathbf{Pr}\left(-\epsilon\sigma\max\{1,\kappa\} \leq \tilde{\boldsymbol{w}}^T\mathbf{X}_i \leq \epsilon\sigma\max\{1,\kappa\}\right) = \int_{\kappa-\epsilon\max\{1,\kappa\}}^{\kappa+\epsilon\max\{1,\kappa\}} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx$$

$$\leq \begin{cases} \sqrt{\frac{2}{\pi}}\epsilon, & \text{if } \kappa \leq 1, \\ \sqrt{\frac{2}{\pi}}e^{-\kappa^2(1-\epsilon)^2/2}\epsilon\kappa, & \text{if } \kappa > 1. \end{cases}$$

Let us pick ϵ such that

$$\epsilon = \frac{1}{n\sqrt{\log n}} \text{ if } \kappa \leq 2\sqrt{\log n}, \quad \text{and} \quad \epsilon = \frac{1}{4} \text{ if } \kappa > 2\sqrt{\log n}.$$

This gives that

$$\mathbf{Pr}\left(\tilde{\boldsymbol{w}}^T\mathbf{X}_i \notin [-\epsilon\sigma\max\{1,\kappa\}, \epsilon\sigma\max\{1,\kappa\}] \text{ for all } i \in [n]\right) \geq 1 - n \cdot o(1/n) = 1 - o(1).$$

Second, consider the event that $|\tilde{\boldsymbol{w}}^T\mathbf{X}_i - \mathbf{E}[\tilde{\boldsymbol{w}}^T\mathbf{X}_i]| \leq 10\sigma\sqrt{\log n}$. By Lemma 18 we know that it holds with probability at least $1 - o(1)$. Under these two events, we have that

$$\epsilon\sigma\max\{1,\kappa\} \leq \tilde{\boldsymbol{w}}^T\mathbf{X}_i \leq \kappa\sigma + 10\sigma\sqrt{\log n} \quad \text{if} \quad \tilde{\boldsymbol{w}}^T\mathbf{X}_i > 0,$$

$$-\kappa\sigma - 10\sigma\sqrt{\log n} \leq \tilde{\boldsymbol{w}}^T\mathbf{X}_i \leq -\epsilon\sigma\max\{1,\kappa\} \quad \text{if} \quad \tilde{\boldsymbol{w}}^T\mathbf{X}_i < 0.$$

Recall from (5) that

$$\Psi(\mathbf{X}_i,\mathbf{X}_j) = \begin{cases} -2R(1-\beta)\tilde{\boldsymbol{w}}^T\mathbf{X}_i, & \text{if } \tilde{\boldsymbol{w}}^T\mathbf{X}_j \leq -\left|\tilde{\boldsymbol{w}}^T\mathbf{X}_i\right|, \\ 2R(1-\beta)\operatorname{sign}(\tilde{\boldsymbol{w}}^T\mathbf{X}_i)\tilde{\boldsymbol{w}}^T\mathbf{X}_j, & \text{if } -\left|\tilde{\boldsymbol{w}}^T\mathbf{X}_i\right| < \tilde{\boldsymbol{w}}^T\mathbf{X}_j < \left|\tilde{\boldsymbol{w}}^T\mathbf{X}_i\right|, \\ 2R(1-\beta)\tilde{\boldsymbol{w}}^T\mathbf{X}_i, & \text{if } \tilde{\boldsymbol{w}}^T\mathbf{X}_j \geq \left|\tilde{\boldsymbol{w}}^T\mathbf{X}_i\right|. \end{cases}$$

Consider an arbitrary node i where $\tilde{\boldsymbol{w}}^T\mathbf{X}_i > 0$. We will show that $h_i = \sum_{j\in N_i}\gamma_{ij}\tilde{\boldsymbol{w}}^T\mathbf{X}_j > 0$. By the definition of attention coefficients in (1), we know that

$$\sum_{j\in N_i}\gamma_{ij}\tilde{\boldsymbol{w}}^T\mathbf{X}_j > 0 \iff \sum_{j\in N_i}\exp(\Psi(\mathbf{X}_i,\mathbf{X}_j))\tilde{\boldsymbol{w}}^T\mathbf{X}_j > 0.$$

Using the above expression for Ψ, we have that

$$\sum_{j\in N_i}\exp(\Psi(\mathbf{X}_i,\mathbf{X}_j))\tilde{\boldsymbol{w}}^T\mathbf{X}_j$$

$$= \exp(\Psi(\mathbf{X}_i,\mathbf{X}_i))\tilde{\boldsymbol{w}}^T\mathbf{X}_i + \sum_{\substack{j\in N_i \\ j\neq i}}\exp(\Psi(\mathbf{X}_i,\mathbf{X}_j))\tilde{\boldsymbol{w}}^T\mathbf{X}_j \tag{30}$$

$$\geq \exp(\Psi(\mathbf{X}_i,\mathbf{X}_i))\tilde{\boldsymbol{w}}^T\mathbf{X}_i - (n-1)\exp(\Psi(\mathbf{X}_i,-\mathbf{X}_i))(\kappa\sigma + 10\sigma\sqrt{\log n})$$

$$\geq e^{2R(1-\beta)\epsilon\sigma\max\{1,\kappa\}}\epsilon\sigma\max\{1,\kappa\} - (n-1)e^{-2R(1-\beta)\epsilon\sigma\max\{1,\kappa\}}(\kappa\sigma + 10\sigma\sqrt{\log n}),$$

where the second last inequality follows from the event that $|\tilde{\boldsymbol{w}}^T \mathbf{X}_j| \leq \kappa\sigma + 10\sigma\sqrt{\log n}$ for all j, and $|N_i| \leq n$, and the last inequality follows from the fact that the function

$$f(x) = xe^{2R(1-\beta)x} - (n-1)(\kappa\sigma + 10\sigma\sqrt{\log n})e^{-2R(1-\beta)x}$$

is increasing with respect to x for $x \geq 0$. To see that $h_i' = \sum_{j \in N_i} \exp(\Psi(\mathbf{X}_i, \mathbf{X}_j))\tilde{\boldsymbol{w}}^T \mathbf{X}_j > 0$, consider the following cases separately.

- If $\kappa \leq 2\sqrt{\log n}$, recall that we picked $\epsilon = 1/(n\log n)$, then because $R = \Omega(n\log^2 n/\sigma)$, for large enough n we have that

$$4R(1-\beta)\sigma > \frac{n\sqrt{\log n}}{\max\{1,\kappa\}}\left(\log n + \log\left(\frac{n\sqrt{\log n}}{\max\{1,\kappa\}}\right) + \log(\kappa + 10\sqrt{\log n})\right)$$

$$\iff 4R(1-\beta)\sigma > \frac{1}{\epsilon\max\{1,\kappa\}}\left(\log n + \log\left(\frac{1}{\epsilon\max\{1,\kappa\}}\right) + \log(\kappa + 10\sqrt{\log n})\right)$$

$$\iff e^{4R(1-\beta)\sigma\epsilon\max\{1,\kappa\}} > \frac{n(\kappa + 10\sqrt{\log n})}{\epsilon\max\{1,\kappa\}}$$

$$\iff e^{2R(1-\beta)\sigma\epsilon\max\{1,\kappa\}}\sigma\epsilon\max\{1,\kappa\} > ne^{-2R(1-\beta)\sigma\epsilon\max\{1,\kappa\}}(\kappa\sigma + 10\sigma\sqrt{\log n}).$$

By (30) this means that $h_i' > 0$.

- If $\kappa > 2\sqrt{\log n}$, then $\epsilon = 1/4$. Because $R = \Omega(n\log^2 n/\sigma)$, for large enough n we have

$$4R(1-\beta)\sigma > \frac{\log n + \log(\kappa + 10\sqrt{\log n}) - \log(\kappa/4)}{\kappa/4}$$

$$\iff 4R(1-\beta)\sigma > \frac{\log n + \log(\kappa + 10\sqrt{\log n}) - \log(\epsilon\kappa)}{\epsilon\kappa}$$

$$\iff e^{4R(1-\beta)\sigma\epsilon\kappa} > \frac{n(\kappa + 10\sqrt{\log n})}{\epsilon\kappa}$$

$$\iff e^{2R(1-\beta)\sigma\epsilon\kappa}\sigma\epsilon\kappa > n(\kappa\sigma + 10\sigma\sqrt{\log n})$$

By (30) this means that $h_i' > 0$.

This shows that $h_i' > 0$ whenever $\tilde{\boldsymbol{w}}^T \mathbf{X}_i > 0$. Similarly, one easily gets that $h_i' < 0$ whenever $\tilde{\boldsymbol{w}}^T \mathbf{X}_i < 0$. Therefore the proof is complete.

A.9 Proof of Corollary 16

We restate Corollary 16 for convenience. Recall that we write $\|\boldsymbol{\mu}\| = \kappa\sigma$ for some $\kappa > 0$.

Corollary. *With probability at least $1 - o(1)$ over the data $(\mathbf{X}, \mathbf{A}) \sim \mathsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$, using the two-layer MLP attention architecture Ψ given in (3) and (4) with $R = \Omega(n\log^2 n\sigma)$, one has that*

- *(Perfect classification) If $\kappa \geq \sqrt{2\log n}$ then all nodes are correctly classified;*

- *(Almost perfect classification) If $\kappa = \omega(1)$ then at least $1 - o(1)$ fraction of all nodes are correctly classified;*

48

- *(Partial classification) If $\kappa = O(1)$ then at least $\Phi(\kappa) - o(1)$ fraction of all nodes are correctly classified.*

Fix $i \in [n]$ and write $\tilde{\boldsymbol{w}}^T \mathbf{X}_i = (2\epsilon_i - 1)\|\boldsymbol{\mu}\| + \sigma/\|\boldsymbol{\mu}\| \cdot \boldsymbol{\mu}^T \boldsymbol{g}_i$ where $\boldsymbol{g}_i \sim N(0, \mathbf{I})$, where we recall that $\epsilon_i \in \{0, 1\}$ defines the class membership of node i. We have that $\tilde{\boldsymbol{w}}^T \mathbf{X}_i$ follows a normal distribution with mean $(2\epsilon_i - 1)\kappa\sigma$ and standard deviation σ. Part 1 of the corollary is exactly Proposition 7 whose proof is given in the main text. We consider the other cases for κ. For both classes $\epsilon_i = 0$ and $\epsilon_i = 1$, the probability of correct classification is $\Phi(\kappa)$ (using 0 as the decision boundary). Therefore, by applying additive Chernoff bound, we have that

$$\mathbf{Pr}\left[\text{At most } n\Phi(\kappa) - \sqrt{n \log n} \text{ nodes are correctly classified}\right]$$

$$= \mathbf{Pr}\left[\sum_{i \in [n]} \mathbf{1}_{\{\text{node } i \text{ is correctly classified}\}} \leq n\Phi(\kappa) - \sqrt{n \log n}\right] \leq \frac{2}{n}.$$

The proof is complete by noticing that $n\Phi(\kappa) - \sqrt{n \log n} = n(1 - o(1))$ when $\kappa = \omega(1)$ and $n\Phi(\kappa) - \sqrt{n \log n} = n(\Phi(\kappa) - o(1))$ when $\kappa = O(1)$.

References

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

R. J. Adler and J. E. Taylor. *Random fields and geometry*. Springer, 2007.

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

A. Athreya, D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, Y. Qin, and D. L. Sussman. Statistical inference on random dot product graphs: A survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018.

J. Atwood and D. Towsley. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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 *International Conference on Machine Learning (ICML)*, 2021.

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

C. Bodnar, F. Di Giovanni, B. Chamberlain, P. Lio, and M. M. Bronstein. Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in gnns. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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.

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

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)*, 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)*, 2015.

M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

V. Garg, S. Jegelka, and T. Jaakkola. Generalization and representational limits of graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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)*, 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.

W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec. Open graph benchmark: datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

N. Keriven, A. Bietti, and S. Vaiter. On the universality of graph neural networks on large random graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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)*, 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.

R. Levie, F. Monti, X. Bresson, and M. M. Bronstein. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing*, 67(1):97–109, 2018.

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

D. Lim, F. Hohne, X. Li, S. L. Huang, V. Gupta, O. Bhalerao, and S. N. Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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.

S. Luan, C. Hua, Q. Lu, J. Zhu, M. Zhao, S. Zhang, X.-W. Chang, and D. Precup. Revisiting heterophily for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

S. Maskey, R. Levie, Y. Lee, and G. Kutyniok. Generalization analysis of message passing neural networks on large random graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

M. Minsky and S. Papert. Perceptron: an introduction to computational geometry, 1969.

C. Moore. The computer science and physics of community detection: Landscapes, phase transitions, and hardness. *Bulletin of The European Association for Theoretical Computer Science*, 1(121), 2017.

J. Palowitch, A. Tsitsulin, B. Mayer, and B. Perozzi. Graphworld: Fake graphs bring real insights for gnns. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.

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

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)*, 2017.

P. Veličković, 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.

M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, T. Xiao, T. He, G. Karypis, J. Li, and Z. Zhang. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019.

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

Y. Yan, M. Hashemi, K. Swersky, Y. Yang, and D. Koutra. Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks. In *IEEE International Conference on Data Mining (ICDM)*, 2022.

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.