

# Graph Attention Retrospective

Kimon Fountoulakis<sup>\*†</sup>

Amit Levi<sup>\*‡</sup>

Shenghao Yang<sup>\*§</sup>

Aseem Baranwal<sup>††</sup>

Aukosh Jagannath<sup>‡‡</sup>

## 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 type of models is graph attention networks. These models were introduced to allow a node to aggregate information from the 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 study theoretically this expected 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 features of the nodes are obtained from a mixture of Gaussians and the edges from a stochastic block model where the features and the edges are coupled in a natural way. First, we show that in an “easy” regime, where the distance between the means of the Gaussians is large enough, graph attention maintains the weights of intra-class edges and significantly reduces the weights of the inter-class edges. As a corollary, we show that this implies perfect node classification independent of the weights of inter-class edges. However, a classical argument shows that in the “easy” regime, the graph is not needed at all to classify the data with high probability. In the “hard” regime, we show that *every* attention mechanism fails to distinguish intra-class from inter-class edges. We evaluate our theoretical results on synthetic and real-world data.

## 1 Introduction

Graph learning has received a lot of attention recently due to breakthrough learning models [GMS05, SGT<sup>+</sup>09, BZSL14, DMAI<sup>+</sup>15, HBL15, AT16, DBV16, HYL17, KW17] 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 popular mechanisms for classification and graph learning in general are the graph convolution and the graph attention. Graph convolution, usually defined using its spatial version, corresponds to averaging the features of a node with the features of its neighbors [KW17]<sup>1</sup>. Graph attention [VCC<sup>+</sup>18] 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

<sup>\*</sup>Equal contribution.

<sup>†</sup>David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada. [kimon.fountoulakis@uwaterloo.ca](mailto:kimon.fountoulakis@uwaterloo.ca).

<sup>‡</sup>Huawei Noah’s Ark Lab, Montreal, Canada. [amit.levi@huawei.com](mailto:amit.levi@huawei.com).

<sup>§</sup>David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada. [shenghao.yang@uwaterloo.ca](mailto:shenghao.yang@uwaterloo.ca).

<sup>††</sup>David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada. [aseem.baranwal@uwaterloo.ca](mailto:aseem.baranwal@uwaterloo.ca).

<sup>‡‡</sup>Department of Statistics and Actuarial Science, Department of Applied Mathematics, University of Waterloo, Waterloo, Canada. [a.jagannath@uwaterloo.ca](mailto:a.jagannath@uwaterloo.ca).

<sup>1</sup>Although the model in [KW17] is related to spectral convolutions, it is mainly a spatial convolution since messages are propagated along graph edges.

practitioners [FL19, WZY<sup>+</sup>19, HFZ<sup>+</sup>20] 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 relatively down-weight inter-class edges before performing the convolution step. 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) [BVR17, DSSM18]. 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 is much larger than the standard deviation, we show that there exists a choice of attention architecture that distinguishes the inter-edges from the intra-edges with high probability. In particular, we show that the attention coefficients for the intra-class edges are much higher than the ones for the inter-class edges. Furthermore, we show perfect node classification, where the performance of the attention architecture is independent of the intra-class edge coefficients. However, in the same regime, we show that the graph is not needed to perfectly classify the data.
2. In the “hard regime”, i.e., when the distance between the means is small compared to the standard deviation, we show that *any* attention architecture is unable to distinguish inter- from intra-class edges with high probability. Moreover, we show that using the original GAT architecture [VCC<sup>+</sup>18], with high probability, most of the attention coefficients are going to have uniform weights, similar to those of uniform graph convolution [KW17].
3. 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 Previous work

Recently the concept of attention for neural networks [BCB15, VSP<sup>+</sup>17] was transferred to graph neural networks [LZBT16, BL18, VCC<sup>+</sup>18, LRK<sup>+</sup>19, PBHL20]. A few papers have attempted to understand the mechanism in [VCC<sup>+</sup>18]. One work relevant to ours is [BAY22]. In this work, the authors show that graph attention can force all nodes to attend to the same neighbor. However, to prove this, the authors assume that all nodes have the same set of neighbors. This deviates significantly from the original graph attention mechanism where each node may have a different set of neighbors. Another related work is [KTA19], which presents an empirical study of the ability of graph attention to generalize on larger, complex, and noisy graphs. In addition, in [HZC<sup>+</sup>19] 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. In particular, random graphs and the stochastic block model have been traditionally used in clustering and community detection [Abb18, AFT<sup>+</sup>18, Moo17]. Moreover, the works by [BVR17, DSSM18], which also rely on CSBM are focused on the fundamental limits of unsupervised learning. Of particular relevance is the work by [BFJ21], which studies the performance of graph convolution on CSBM as a semi-supervised learning problem. Finally, there have been a few related theoretical works on understanding the performance and the universality of graph neural networks [CLB19, CPLM21, ZYZ<sup>+</sup>20, XHLJ19, GJJ20, Lou20a, Lou20b]. However, compared to these works, we provide theoretical results that 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)* [DSSM18] which serves as our data model, and the *Graph Attention* mechanism [VCC<sup>+</sup>18].

Let  $d, n \in \mathbb{N}$ , and  $\epsilon_1, \dots, \epsilon_n \sim \text{Ber}(1/2)$ <sup>2</sup>, and define two classes as  $C_k = \{j \in [n] \mid \epsilon_j = k\}$  for  $k \in \{0, 1\}$ . For each index  $i \in [n]$ , we set the feature vector  $\mathbf{X}_i \in \mathbb{R}^d$  as  $\mathbf{X}_i \sim N((2\epsilon_i - 1)\boldsymbol{\mu}, \mathbf{I} \cdot \sigma^2)$ <sup>3</sup>, where  $\boldsymbol{\mu} \in \mathbb{R}^d$ ,  $\sigma \in \mathbb{R}$  and  $\mathbf{I} \in \{0, 1\}^{d \times d}$  is the identity matrix. For a given pair  $p, q \in [0, 1]$  we consider the stochastic adjacency matrix  $\mathbf{A} \in \{0, 1\}^{n \times n}$  defined as follows. For  $i, j \in [n]$  in the same class (i.e., *intra-edge*), we set  $a_{ij} \sim \text{Ber}(p)$ , and if  $i, j$  are in different classes (i.e., *inter-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 separability of the Gaussians. For example, fixing the variance, then the closer the means are the more difficult the separability of the Gaussians becomes. Moreover, CSBM allows us to control the noise in the graph, namely the difference between intra- and inter-class edge probabilities.

A *single-head* graph attention applies some weight function on the edges based on their node features (or a mapping thereof). Given two representations  $\mathbf{h}_i, \mathbf{h}_j \in \mathbb{R}^{F'}$  for two nodes  $i, j \in [n]$ , let  $\Psi(\mathbf{h}_i, \mathbf{h}_j) \stackrel{\text{def}}{=} \alpha(\mathbf{W}\mathbf{h}_i, \mathbf{W}\mathbf{h}_j)$  where  $\alpha: \mathbb{R}^F \times \mathbb{R}^F \rightarrow \mathbb{R}$  and  $\mathbf{W} \in \mathbb{R}^{F \times F'}$  is a learnable matrix. We refer to the mapping  $\Psi$  as the *attention model/mechanism* with *attention coefficients*:

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

where  $N_i$  is the set of neighbors of node  $i$ . Letting  $f$  be some nonlinear activation function applied element-wise, the graph attention convolution output is

$$\tilde{\mathbf{h}}_i = f(\mathbf{h}'_i), \text{ where } \mathbf{h}'_i = \sum_{j \in [n]} \mathbf{A}_{ij} \gamma_{ij} \mathbf{W}\mathbf{h}_j \quad \forall i \in [n].$$

A *multi-head* graph attention [VCC<sup>+</sup>18] uses  $K \in \mathbb{N}$  weight matrices  $\mathbf{W}^1, \dots, \mathbf{W}^K \in \mathbb{R}^{F \times F'}$  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  $\alpha$  is realized by an MLP using the LeakyRelu<sup>4</sup> activation function.

The CSBM model induces dataset features  $\mathbf{X} \in \mathbb{R}^{n \times d}$  which are correlated through the graph  $G = ([n], E)$ , represented by an adjacency matrix  $\mathbf{A} \in \{0, 1\}^{n \times n}$ . A natural requirement of a good attention architecture is to maintain intra-class edges in the graph (i.e. edges in  $C_0^2$  or  $C_1^2$ )<sup>5</sup>, and ignore inter-class edges (i.e. edges in either  $C_0 \times C_1$  or  $C_1 \times C_0$ ), which makes a node from a class connected only to nodes from its own class. More specifically, a node  $v$  will be connected to nodes coming from the *same distribution* as  $v$ . To study this we will occasionally use the following definition of separability of edges. Given an attention model  $\Psi$ , we say that the model *separates the edges*, if the outputs  $\Psi(\mathbf{X}_i, \mathbf{X}_j)$  satisfy

$$\Psi(\mathbf{X}_i, \mathbf{X}_j) > 0, \text{ if } (i, j) \in (C_1^2 \cup C_0^2) \cap E, \quad \text{and} \quad \Psi(\mathbf{X}_i, \mathbf{X}_j) < 0, \text{ if } (i, j) \in E \setminus (C_1^2 \cup C_0^2).$$

<sup>2</sup>Ber( $\cdot$ ) denotes the Bernoulli distribution.

<sup>3</sup>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.

<sup>4</sup>LeakyRelu( $x$ ) is defined as  $x$  if  $x \geq 0$  and  $\beta x$  for some constant  $\beta \in [0, 1)$  otherwise.

<sup>5</sup>For two sets  $A$  and  $B$  we denote  $A \times B = \{(i, j) : i \in A, j \in B\}$ , e.g.,  $C_0^2 = C_0 \times C_0 = \{(i, j) \in [n] : i \in C_0, j \in C_0\}$ .

### 3 Results

We consider two parameter regimes: the first (“easy regime”) is where  $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ , and the second (“hard regime”) is where  $\|\boldsymbol{\mu}\| = K\sigma$  for some  $0 < K \leq O(\sqrt{\log n})$ . All of our results rely on a mild assumption that lower bounds the sparsity of the graph generated by the CSBM model.

**Assumption 1.**  $p, q = \Omega(\log^2 n/n)$ , and  $p \geq q$ .<sup>6</sup>

#### 3.1 “Easy Regime”

In this regime ( $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ ) we show that there exists a choice of attention architecture  $\Psi$ , which is able to separate the edges with high probability.

**Theorem 1.** *Suppose that  $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ . Then, there exists a choice of attention architecture  $\Psi$  such that with probability at least  $1 - o_n(1)$  over the data  $(\mathbf{X}, \mathbf{A}) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$  it holds that  $\Psi$  separates intra-edges from inter-edges.*

We present two corollaries of the above theorem. The first corollary characterizes the attention coefficients induced by using the architecture,  $\Psi$ , from Theorem 1. In this regime, separability of the edges implies high concentration for the attention coefficients,  $\gamma_{ij}$ . This shows the desired behavior of the attention mechanism—it maintains intra-class edges and essentially ignores all inter-edges.

**Corollary 2.** *Suppose that  $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ . Then there exists a choice of attention architecture  $\Psi$  such that with probability at least  $1 - o_n(1)$  over the data  $(\mathbf{X}, \mathbf{A}) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$  it holds that if  $(i, j)$  is intra-edge then  $\gamma_{ij} = \frac{2}{np}(1 \pm o_n(1))$ , and  $\gamma_{ij} = o\left(\frac{1}{n(p+q)}\right)$  otherwise.*

Note that the attention coefficients which correspond to the intra-edges are much larger than the coefficients of the inter-edges. This essentially means that the model ignores the inter-edges.

By using the above result on  $\gamma_{ij}$  we obtain a *node classification* result as well. In the following, we say that the model *separates the nodes* if  $\mathbf{h}'_i > 0$  when  $i \in C_1$  and  $\mathbf{h}'_i < 0$  when  $i \in C_0$ .

**Corollary 3.** *Suppose that  $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ . Then, there exists a choice of attention architecture  $\Psi$  such that with probability at least  $1 - o_n(1)$  over the data  $(\mathbf{X}, \mathbf{A}) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ , the model separates the nodes for any  $p, q$  satisfying Assumption 1.*

The proof of this corollary is a consequence of Corollary 2. Intuitively, Corollary 2 implies that the means of the convolved data should concentrate around the same means as the original data. On the other hand, by the choice of  $p, q$  we expect that each node will be connected to  $\Theta(np)$  many intra-class nodes so that the averaging operation reduces the variance significantly to  $\approx \sigma^2/np$ . However, since the distance between the new means is around  $2\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$  and the variance is much smaller than  $\sigma^2$ , we can expect to achieve perfect node separability.

Nonetheless, in this regime, we prove that a simple linear classifier, which does not use the graph at all, achieves perfect node separability with high probability. In particular, by a classical argument [And03], the Bayes optimal classifier for the node features (*without the graph*)<sup>7</sup> is realized by a simple linear classifier, which achieves perfect node separability with high probability.

<sup>6</sup>The assumption  $p \geq q$  is only required for the analyses for the “hard” regime.

<sup>7</sup>A Bayes optimal classifier makes the most probable prediction for a data point. Formally, for our scenario, such a classifier is given by  $h^*(\mathbf{x}) = \arg \max_{c \in \{0,1\}} \Pr[y = c \mid \mathbf{x}]$ .

**Proposition 4.** *Suppose  $\|\boldsymbol{\mu}\| = \Omega(\sigma\sqrt{\log n})$ , and let  $\mathbf{X}$  be sampled from the Gaussian mixture model. Then, the Bayes optimal classifier is realized by a linear classifier which achieves perfect node separability with probability at least  $1 - o_n(1)$  over  $\mathbf{X}$ .*

This implies that in the above regime, using the graph model is unnecessary, as it does not provide additional power compared to simple logistic regression for the node classification task.

### 3.2 “Hard Regime”

In this regime ( $\|\boldsymbol{\mu}\| = K\sigma$  for  $K \leq O(\sqrt{\log n})$ ), we show that any attention architecture  $\Psi$  will fail to separate the edges. The next theorem quantifies the misclassification of edge pairs that  $\Psi$  exhibits. Below we define  $\Phi_c(\cdot) = 1 - \Phi(\cdot)$ , where  $\Phi(\cdot)$  denotes standard Gaussian CDF.

**Theorem 5.** *Suppose  $\|\boldsymbol{\mu}\|_2 = K\sigma$  for some  $K > 0$  and let  $\Psi$  be any attention mechanism. Then,*

1. *For any  $c' > 0$ , with probability at least  $1 - O(n^{-c'})$ ,  $\Psi$  fails to distinguish at least a  $2 \cdot \Phi_c(K)^2$  fraction of the inter-edges from the intra-edges.*
2. *For any  $\kappa > 1$  if  $q > \frac{\kappa \log^2 n}{n \Phi_c(K)^2}$ , then with probability at least  $1 - O\left(\frac{1}{n^{\frac{\kappa}{4} \Phi_c(K)^2 \log n}}\right)$ ,  $\Psi$  fails to distinguish at least one inter-edge pair from the intra-edge pairs.*

Part 1 of the theorem implies that if  $\|\boldsymbol{\mu}\|_2$  is *linear* in the standard deviation,  $\sigma$ , then with overwhelming probability the attention mechanism fails to distinguish a constant fraction of inter-edge pairs from the intra-edge pairs. Furthermore, part 2 of the theorem characterizes a regime for the inter-edge probability  $q$  where the attention mechanism fails to distinguish at least one inter-edge node pair, by providing a lower bound on  $q$  in terms of the scale at which the distance between the means grows compared to the standard deviation  $\sigma$ . This aligns with the intuition that as we increase the distance between the means, it gets easier for the attention mechanism to correctly distinguish the node pairs. However, if  $q$  is also increased in the right proportion (in other words, if the noise in the graph is increased), then the attention mechanism will still fail to correctly distinguish at least one of the inter-edge node pairs. For instance, setting  $K = \sqrt{2 \log \log n}$  and  $\kappa = 4$ , we get that if  $q > \Omega\left(\frac{\log^{4+o(1)} n}{n}\right)$ , then with probability at least  $1/2$ ,  $\Psi$  misclassifies at least one inter-edge.

As a motivating example, we focus our attention on one of the most popular attention architecture [VCC<sup>+</sup>18], where  $\alpha$  is a single layer neural network parametrized by  $(\mathbf{w}, \mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}^2 \times \mathbb{R}$  with LeakyRelu as activation. Namely, the attention coefficients are defined by

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

For the above attention mechanism, we show that with a very high probability most of the attention coefficients  $\gamma_{ij}$  are going to be  $\Theta(1/|N_i|)$ , which implies that the model fails to distinguish intra- and inter-edges.

**Theorem 6** (informal). *Assume that  $\|\boldsymbol{\mu}\|_2 \leq K\sigma$  and  $\sigma \leq K'$  for some constants  $K$  and  $K'$ . Moreover, assume that the parameters  $(\mathbf{w}, \mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}^2 \times \mathbb{R}$  are bounded by a constant. Then, with probability at least  $1 - o_n(1)$  over the data  $(\mathbf{X}, \mathbf{A}) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ , at least 90% of  $\gamma_{ij}$  are  $\Theta\left(\frac{1}{|N_i|}\right)$ .*

We explain the idea behind this result. When we consider the pair space  $(\mathbf{w}^T \mathbf{X}_i, \mathbf{w}^T \mathbf{X}_j)$ , we can think of each pair as a two-dimensional Gaussian vector, whose means are either  $(\mathbf{w}^T \boldsymbol{\mu}, \mathbf{w}^T \boldsymbol{\mu})$ ,  $(-\mathbf{w}^T \boldsymbol{\mu}, \mathbf{w}^T \boldsymbol{\mu})$ ,

$(\mathbf{w}^T \boldsymbol{\mu}, -\mathbf{w}^T \boldsymbol{\mu})$  or  $(-\mathbf{w}^T \boldsymbol{\mu}, -\mathbf{w}^T \boldsymbol{\mu})$ . We need to classify the data corresponding to means  $(\mathbf{w}^T \boldsymbol{\mu}, \mathbf{w}^T \boldsymbol{\mu})$  and  $(-\mathbf{w}^T \boldsymbol{\mu}, -\mathbf{w}^T \boldsymbol{\mu})$  as positive (i.e., intra-edges) and classify the data corresponding to the other means as negative (i.e., inter-edges). This problem is known in the literature as the “XOR problem” [MP69]. As the distance between the means is linear in the standard deviation, we can expect a high overlap in the high probability support of all four distributions. Therefore, any linear classifier will misclassify a large fraction of the pairs. In fact, it is quite easy to show that *for any distance between the means* a linear classifier fails on the XOR problem.

Compared to the easy regime, it seems difficult to obtain a separation result for the nodes. In the easy regime, the distance between the means was significantly 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 current regime where the distance between the means is rather small, we get that the “noise” dominates the “signal”. Nonetheless, we conjecture the following.

**Conjecture 7.** *Suppose that  $\|\boldsymbol{\mu}\|_2 \leq K\sigma$  and  $\sigma \leq K'$  for some constants  $K$  and  $K'$ . Then, any single layer graph attention model fails to perfectly classify the nodes with high probability when  $p-q = O\left(\sigma\sqrt{(\log n)/\Delta}\right)$ , where  $\Delta$  is the expected degree.*

The above conjecture means that if  $p$  and  $q$  are close then any graph attention model will fail to perfectly classify the nodes. This implies that in the hard regime the performance of the graph attention model depends on  $q$  as opposed to the easy regime, where in Theorem 3 we show that it doesn’t. This property is verified by our synthetic experiments in Figures 1c and 1f. The  $\sigma\sqrt{(\log n)/\Delta}$  bound comes from our conjecture that the expected maximum of the graph convolved data (with attention) over the nodes is at least  $c\sigma\sqrt{(\log n)/\Delta}$  for some constant  $c > 0$ .

## 4 Experiments

In this section, we demonstrate empirically our results in Section 3 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 the supplementary material. Also, in the supplementary material, we provide experiments on real data where PyTorch Geometric [FL19] is used to train the models. The results are similar to the main paper, we provide discussion when there are discrepancies. With two exemptions in Figures 1e and 2b, in all our experiments we use MLP-GAT, where the attention mechanism  $\Psi$  is set to be a two-layer network using the LeakyRelu activation function. The exemptions are made to demonstrate Theorem 6. We demonstrate more results for the original GAT [VCC+18] mechanism with two heads in the supplementary material.

### 4.1 Synthetic data

We use the CSBM to generate the data. 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.

#### 4.1.1 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  $\sigma$ . We vary  $q$  from  $\log^2(n)/n$  to  $p$ .



In Figure 1 we illustrate Theorem 1 and Corollaries 2, 3 for the easy regime, and Theorems 5, 6 for the hard regime. In particular, in Figure 1a we show Theorem 1, MLP-GAT is able to classify intra and inter edges perfectly. In Figure 1b we show that in the easy regime, the  $\gamma$  that correspond to intra-edges concentrate around  $2/np$  for MLP-GAT, while the  $\gamma$  for the inter-edges concentrate to tiny values, as proved in Corollary 2. In Figure 1c we observe that the performance of MLP-GAT for node classification is independent of  $q$  in the easy regime as is proved in Corollary 3. However, in this plot, we observe that not using the graph also achieves perfect node classification, a result which is proved in Proposition 4. In the same plot, we also show the performance of uniform graph convolution [KW17], where its performance depends on  $q$  (see [BFJ21]). In Figure 1d we show Theorem 5, MLP-GAT misclassifies a constant fraction of the intra and inter edges as proved in Theorem 5. In Figure 1e we show Theorem 6,  $\gamma$  in the hard regime concentrate around uniform (GCN) coefficients for both MLP-GAT and GAT. In Figure 1f we illustrate that node classification accuracy is a function of  $q$  for MLP-GAT. This is conjectured in Conjecture 7.

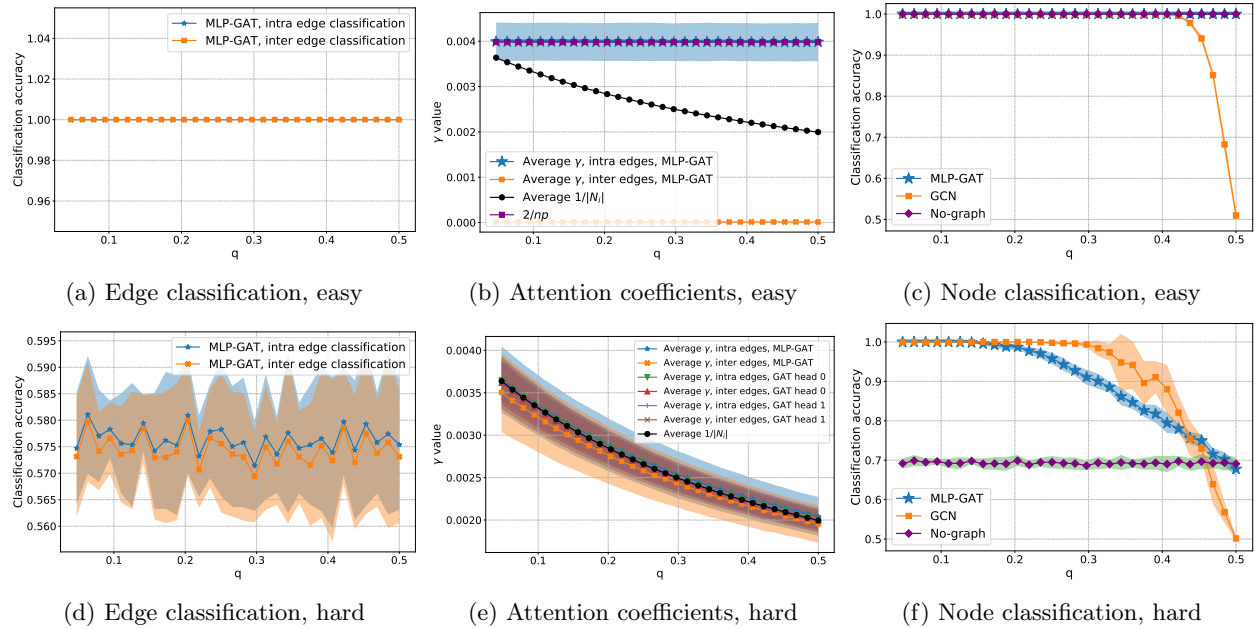


Figure 1: Demonstration of Theorem 1 and Corollaries 2, 3 for the easy regime, and Theorems 5, 6 for the hard regime. The first row of figures demonstrates the former results and the second row of figures demonstrates the latter results. The shaded areas show standard deviation.

#### 4.1.2 Fixing $q$ and varying the distance between the means

We consider the case where  $q = 0.1$ . In the supplementary material, we also demonstrate the case where  $q = 0.4$ . The results are similar in this case too. In Figure 2 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  $\sigma$  to approximately separate the easy (left of  $\sigma$ ) and hard (right of  $\sigma$ ) regimes. Figure 2a illustrates Theorems 1 and 5 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 2b we observe that in the hard regime  $\gamma$  concentrate around the uniform (GCN) coefficients, while in the easy regime MLP-GAT is able to maintain the  $\gamma$  for the intra edges, while it sets the  $\gamma$  to tiny values for the inter edges. In Figure 2c. we observe that in the hard regime  $\gamma$  of GAT concentrate around the

uniform coefficients (proved in Theorem 6), while in the easy regime although the  $\gamma$  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 that GAT is using, see the discussion below Theorem 6. Finally, in Figure 2d we show node classification results for MLP-GAT. In the easy regime we observe perfect classification as proved in Corollary 3. However, as the distance between the means decreases, we observe that MLP-GAT starts to misclassify nodes.

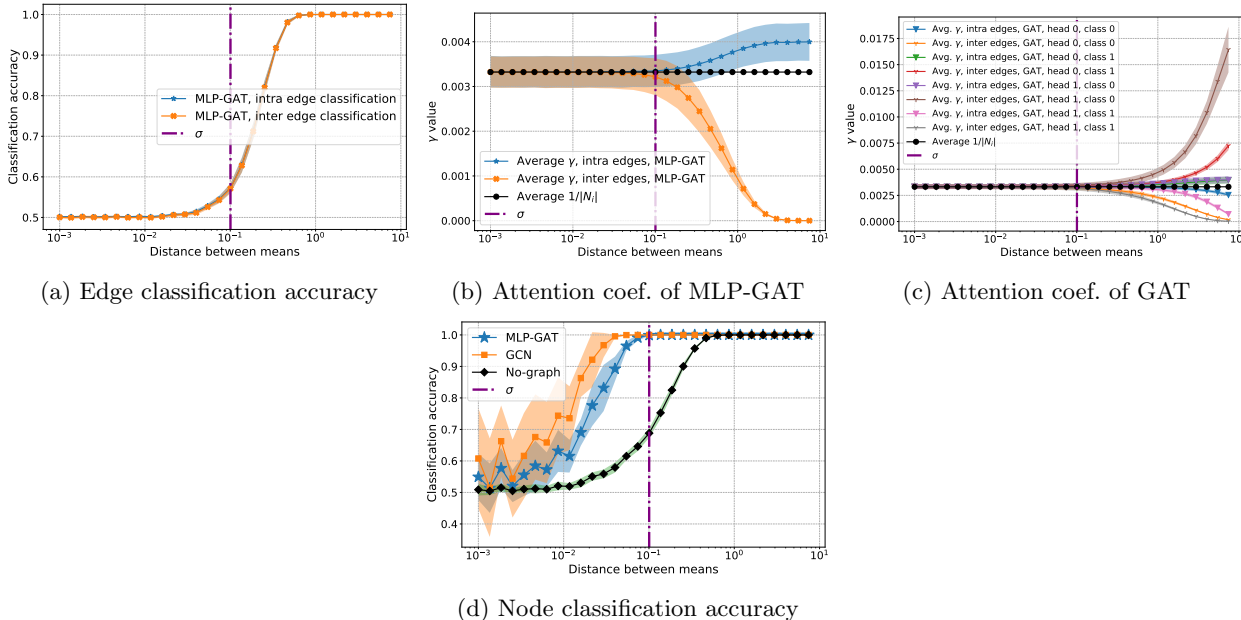


Figure 2: Attention coefficients of MLP-GAT and GAT, node and edge classification as a function of the distance between the means. The shaded areas in our plots show standard deviation.

## 4.2 Real data

In this experiment, we illustrate the attention coefficients, node and edge classification for MLP-GAT as a function of the distance between the means on real data. We use the popular real data Cora, PubMed, and CiteSeer. These data are publicly available and can be downloaded from [FL19]. 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 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  $\mu$  and  $-\mu$  to each class, respectively. This modification can be thought of as translating the mean of the distribution of the data for each class.

The results of this experiment are shown in Figure 3. In this figure we show results only for class 0 of each dataset, for results on other classes see the supplementary material. The results are similar. 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 the intra and inter edges, see Figures 3a, 3d and 3g. Moreover, as the distance between the means increases, the average intra  $\gamma$  becomes much larger than the average inter  $\gamma$ , see Figures 3b, 3e and 3h, and the model is able to classify the nodes accurately, see Figures 3c, 3f and 3i. On the contrary, in the same figures, we



observe that as the distance of the means decreases then MLP-GAT is not able to separate intra from inter edges, the averaged  $\gamma$  are very close to uniform coefficients and the model can't classify the nodes accurately.

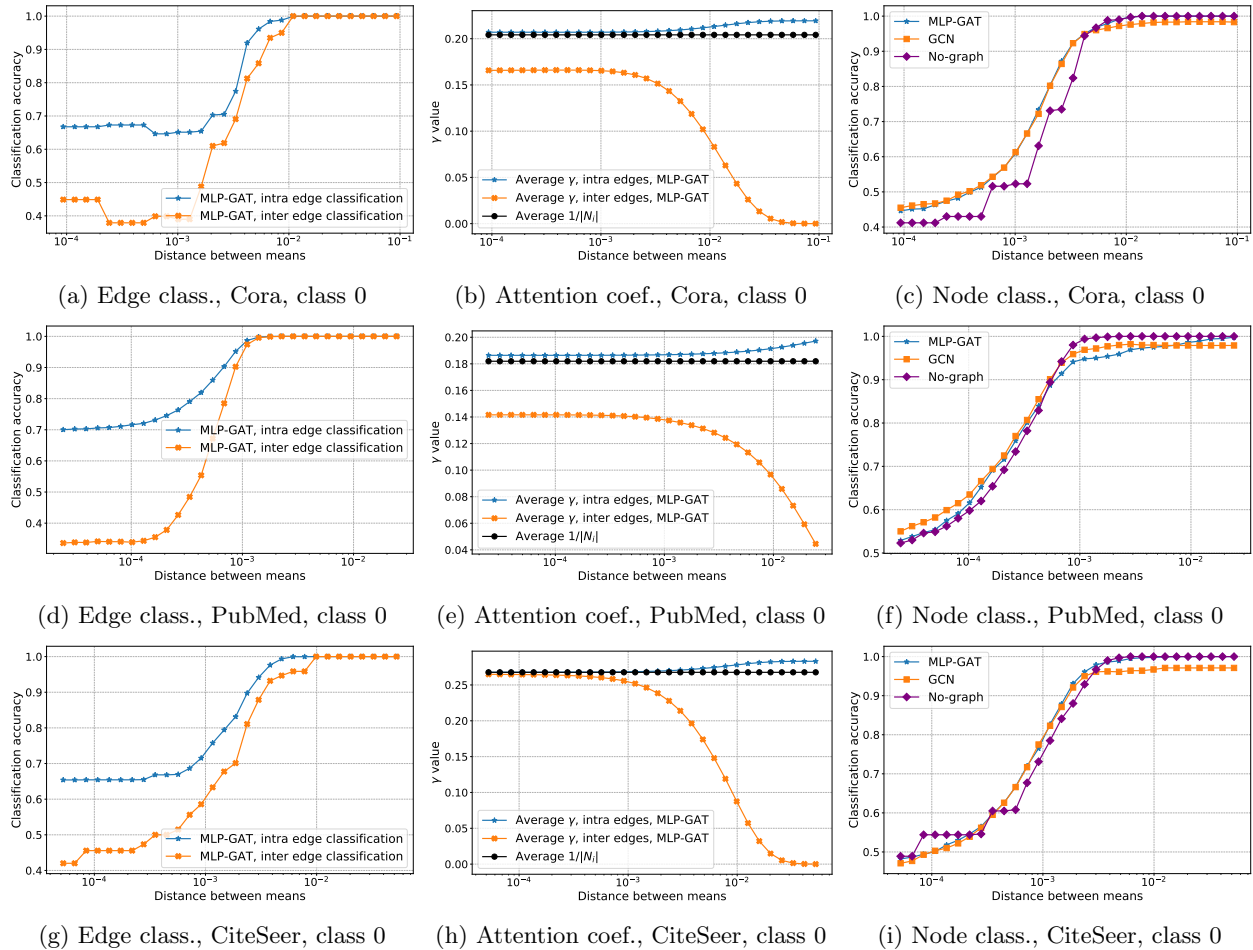


Figure 3: Attention coefficients, node and edge classification for MLP-GAT as a function of the distance between the means for real data.

Note that Figure 3 does not show the standard deviation for the attention coefficients  $\gamma$ . We show the standard deviation of  $\gamma$  in Figure 4. 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  $\gamma$ . This is to be expected since for the real data the degrees of the nodes do not concentrate as well. In Figure 4 we show that the standard deviation of the uniform coefficients  $1/|N_i|$  is also high and that the standard deviation of  $\gamma$  is similar to that of  $1/|N_i|$  for intra-class  $\gamma$ , while the deviation for inter-class  $\gamma$  is large for a small distance between the means, but it gets much smaller as the distance increases.

## 5 Conclusion and future work

We show that graph attention improves robustness to noise in graph structure in an “easy” regime, where the graph is not needed at all. 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 has limited

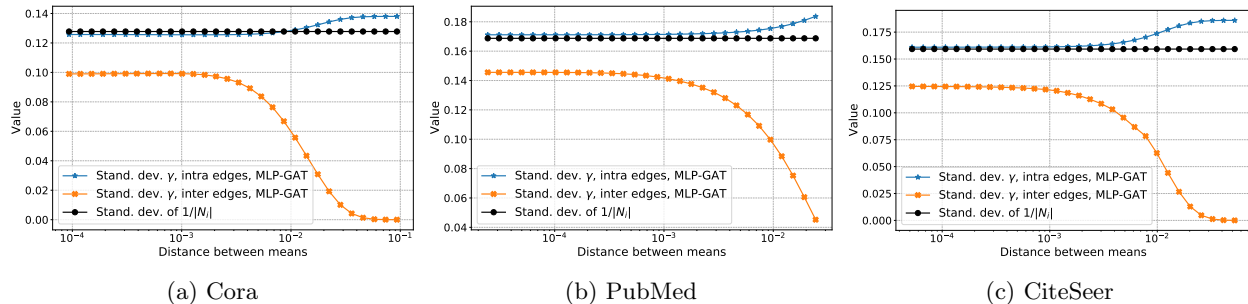


Figure 4: Standard deviation for attention coefficients of MLP-GAT.

power at distinguishing intra- 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 mechanisms for distinguishing intra- and inter-class edges.

## 6 Acknowledgements

K. Fountoulakis would like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Waterloo-Huawei Joint Innovation Laboratory. Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [RGPIN-2019-04067, DGEGR-2019-00147].

A. Jagannath acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [RGPIN-2020-04597, DGEGR-2020-00199].

## References

- [Abb18] E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18:1–86, 2018.
- [AFT<sup>+</sup>18] 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.
- [And03] T.W. Anderson. *An introduction to multivariate statistical analysis*. John Wiley & Sons, 2003.
- [AS04] N. Alon and J. H. Spencer. *The Probabilistic Method*. John Wiley & Sons, 2004.
- [AT16] J. Atwood and D. Towsley. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, page 2001–2009, 2016.
- [BAY22] S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2022.
- [BCB15] 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.

- [BFJ21] A. Baranwal, K. Fountoulakis, and A. Jagannath. Graph convolution for semi-supervised classification: Improved linear separability and out-of-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 684–693, 2021.
- [BL18] X. Bresson and T. Laurent. Residual gated graph convnets. In *arXiv:1711.07553*, 2018.
- [BVR17] N. Binkiewicz, J. T. Vogelstein, and K. Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104:361–377, 2017.
- [BZSL14] 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.
- [CLB19] Z. Chen, L. Li, and J. Bruna. Supervised community detection with line graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.
- [CPLM21] E. Chien, J. Peng, P. Li, and O. Milenkovic. Adaptive universal generalized pagerank graph neural network. In *International Conference on Learning Representations (ICLR)*, 2021.
- [DBV16] 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)*, page 3844–3852, 2016.
- [DMAI<sup>+</sup>15] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 45, page 2224–2232, 2015.
- [DSSM18] Y. Deshpande, A. Montanari S. Sen, and E. Mossel. Contextual stochastic block models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [FL19] M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- [GJJ20] V. Garg, S. Jegelka, and T. Jaakkola. Generalization and representational limits of graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 119, pages 3419–3430, 2020.
- [GMS05] 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.
- [Han14] R. Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.
- [HBL15] M. Henaff, J. Bruna, and Y. LeCun. Deep convolutional networks on graph-structured data. In *arXiv:1506.05163*, 2015.
- [HFZ<sup>+</sup>20] 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.
- [HYL17] W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1025–1035, 2017.
- [HZC<sup>+</sup>19] 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.

- [KTA19] B. Knyazev, G. W. Taylor, and M. Amer. Understanding attention and generalization in graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4202–4212, 2019.
- [KW17] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.
- [Lou20a] A. Loukas. How hard is to distinguish graphs with graph neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [Lou20b] A. Loukas. What graph neural networks cannot learn: Depth vs width. In *International Conference on Learning Representations (ICLR)*, 2020.
- [LRK<sup>+</sup>19] 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.
- [LZBT16] Y. Li, R. Zemel, M. Brockschmidt, and D. Tarlow. Gated graph sequence neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.
- [Moo17] 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.
- [MP69] M. Minsky and S. Papert. Perceptron: an introduction to computational geometry, 1969.
- [PBHL20] O. Puny, H. Ben-Hamu, and Y. Lipman. Global attention improves graph networks generalization. In *arXiv:2006.07846*, 2020.
- [RH15] P. Rigollet and J.-C. Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813:814, 2015.
- [SGT<sup>+</sup>09] 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.
- [VCC<sup>+</sup>18] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.
- [Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- [VSP<sup>+</sup>17] 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)*, page 6000–6010, 2017.
- [WZY<sup>+</sup>19] 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.
- [XHLJ19] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.
- [YZY<sup>+</sup>20] 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.

# A Proofs

## A.1 General Results

We first start by stating some standard definitions and probability tools which will be used throughout this work. The first definition is regarding *sub-Gaussian* random variables. Those random variables are characterized by their tail decay.

**Definition A.1** ([Ver18]). *We say that a random variable  $z$  follows sub-Gaussian distribution if there are positive constants  $C, v$  such that for every  $t > 0$*

$$\Pr[|z - \mathbf{E}[z]| > t] \leq C \exp(-vt^2).$$

*Equivalently,  $z$  is sub-Gaussian if  $\mathbf{E}[\exp(a(z - \mathbf{E}[z])^2)] \leq 2$  for some  $a > 0$  (this condition is known as  $\psi_2$ -condition).*

The following lemma discuss the behavior of the maxima of sub-Gaussian random variables.

**Lemma A.2** ([RH15]). *Let  $\mathbf{x}_1, \dots, \mathbf{x}_n$  be sub-Gaussian random variables with the same mean and sub-Gaussian parameter  $\tilde{\sigma}^2$ . Then,*

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

*Moreover, for any  $t > 0$*

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

Next, we define *Lipschitz* functions, and state the LeakyRelu is Lipschitz.

**Definition A.3.** *Let  $(\mathcal{X}, d_{\mathcal{X}})$  and  $(\mathcal{Y}, d_{\mathcal{Y}})$  be metric spaces. A function  $f : \mathcal{X} \rightarrow \mathcal{Y}$  is called  $L$ -Lipschitz if there is  $L \in \mathbb{R}$  such that for every  $u, v \in \mathcal{X}$*

$$d_{\mathcal{Y}}(f(u), f(v)) \leq L \cdot d_{\mathcal{X}}(u, v).$$

**Fact A.4** ([Ver18]). *LeakyRelu is  $L$ -Lipschitz with  $L \leq 1$ .*

Next, for completeness, we state two forms (additive and multiplicative) of Chernoff bounds used in this work.

**Lemma A.5** ([AS04, Han14, Ver18]). *Let  $\chi_1, \dots, \chi_n$  be identical independent random variables ranging in  $[0, 1]$ , and let  $p = \mathbf{E}[\chi_1]$ . Then for any  $\epsilon \in (0, 1)$ , it holds that*

$$\Pr \left[ \left| \frac{1}{n} \sum_{i \in [n]} \chi_i - p \right| > \epsilon \right] < 2 \exp \left( -\frac{\epsilon^2 n}{4} \right),$$

*and for any  $\gamma \in (0, 2]$ , it holds that*

$$\Pr \left[ \left| \frac{1}{n} \sum_{i \in [n]} \chi_i - p \right| > \gamma p \right] < 2 \exp \left( -\frac{\gamma^2 p n}{4} \right).$$

In order to prove Theorem 1 we will need the following concentration result on LeakyRelu whose constant denoted by  $\beta$ . Fix  $(\mathbf{w}, \mathbf{a}) \in \mathbb{R}^d \times \mathbb{R}^2$  and for  $i, j \in [n]$  let

$$\mathbf{z}_{ij} = \mathbf{a}_1 \mathbf{w}^T \mathbf{X}_i + \mathbf{a}_2 \mathbf{w}^T \mathbf{X}_j \sim \begin{cases} N((\mathbf{a}_1 + \mathbf{a}_2) \mathbf{w}^T \boldsymbol{\mu}, \sigma^2 \|\mathbf{a}\|^2 \|\mathbf{w}\|^2) & \text{if } i, j \in C_1 \\ N((\mathbf{a}_1 - \mathbf{a}_2) \mathbf{w}^T \boldsymbol{\mu}, \sigma^2 \|\mathbf{a}\|^2 \|\mathbf{w}\|^2) & \text{if } i \in C_1, j \in C_0 \\ N(-(\mathbf{a}_1 - \mathbf{a}_2) \mathbf{w}^T \boldsymbol{\mu}, \sigma^2 \|\mathbf{a}\|^2 \|\mathbf{w}\|^2) & \text{if } i \in C_0, j \in C_1 \\ N(-(\mathbf{a}_1 + \mathbf{a}_2) \mathbf{w}^T \boldsymbol{\mu}, \sigma^2 \|\mathbf{a}\|^2 \|\mathbf{w}\|^2) & \text{if } i, j \in C_0 \end{cases}.$$

**Lemma A.6.** *There exists an absolute constant  $C > 0$  such that with probability at least  $1 - o_n(1)$ , we have*

$$\begin{aligned} \text{LeakyRelu}(\mathbf{z}_{ij}) &= \text{LeakyRelu}((\mathbf{a}_1 + \mathbf{a}_2) \mathbf{w}^T \boldsymbol{\mu}) \pm C \sigma \|\mathbf{a}\| \|\mathbf{w}\| \sqrt{2 \log n}, & \text{if } i, j \in C_1, \\ \text{LeakyRelu}(\mathbf{z}_{ij}) &= \text{LeakyRelu}((\mathbf{a}_1 - \mathbf{a}_2) \mathbf{w}^T \boldsymbol{\mu}) \pm C \sigma \|\mathbf{a}\| \|\mathbf{w}\| \sqrt{2 \log n}, & \text{if } i \in C_1, j \in C_0, \\ \text{LeakyRelu}(\mathbf{z}_{ij}) &= \text{LeakyRelu}(-(\mathbf{a}_1 - \mathbf{a}_2) \mathbf{w}^T \boldsymbol{\mu}) \pm C \sigma \|\mathbf{a}\| \|\mathbf{w}\| \sqrt{2 \log n}, & \text{if } i \in C_0, j \in C_1, \\ \text{LeakyRelu}(\mathbf{z}_{ij}) &= \text{LeakyRelu}(-(\mathbf{a}_1 + \mathbf{a}_2) \mathbf{w}^T \boldsymbol{\mu}) \pm C \sigma \|\mathbf{a}\| \|\mathbf{w}\| \sqrt{2 \log n}, & \text{if } i, j \in C_0. \end{aligned}$$

**Proof:** Since for every  $i, j \in [n]^2$  the random variable  $\mathbf{z}_{ij}$  follows a normal distribution, by definition it is sub-Gaussian with parameter  $c \cdot \sqrt{\text{Var}[\mathbf{z}_{ij}]}$  for  $c > 1$  large enough constant (see definition A.1). By Fact A.4, LeakyRelu is  $L$ -Lipschitz function with  $L \leq 1$

$$\begin{aligned} \mathbf{E}_{\mathbf{z}} \left[ \exp \left( \frac{(\text{LeakyRelu}(\mathbf{z}) - \mathbf{E}[\text{LeakyRelu}(\mathbf{z})])^2}{K^2} \right) \right] &= \mathbf{E}_{\mathbf{z}'} \left[ \exp \left( \frac{(\text{LeakyRelu}(\mathbf{z}) - \text{LeakyRelu}(\mathbf{z}'))^2}{K^2} \right) \right] \\ &= \mathbf{E}_{\mathbf{z}} \left[ \exp \left( \frac{L^2 (\mathbf{z} - \mathbf{E}[\mathbf{z}])^2}{K^2} \right) \right]. \end{aligned} \quad (1)$$

Setting  $K = c \cdot \sqrt{\text{Var}[\mathbf{z}]} \cdot L$ , implies that (1) is bounded above by 2, which means that Leaky-Relu is sub-Gaussian with parameter  $c \cdot \sqrt{\text{Var}[\mathbf{z}]} \cdot L$  (see [Ver18]). Therefore for any  $t > 0$ ,

$$\Pr_{\mathbf{z}} [|\text{LeakyRelu}(\mathbf{z}) - \mathbf{E}[\text{LeakyRelu}(\mathbf{z})]| \geq t] \leq 2 \exp \left( -\frac{t^2}{c^2 L^2 \text{Var}[\mathbf{z}]} \right). \quad (2)$$

Setting  $t = 10cL \sqrt{\text{Var}[\mathbf{z}] \log n}$ , and applying a union bound over all  $i, j \in [n]^2$ , we get that with probability at least  $1 - \frac{2}{n^{98}}$ , the complement of (2) holds for all  $i, j \in [n]^2$ .

Next, we estimate  $\mathbf{E}[\text{LeakyRelu}(\mathbf{z})]$ . For any  $t' > 0$  we have

$$\mathbf{E}[\text{LeakyRelu}(\mathbf{z})] = \mathbf{E}[\text{LeakyRelu}(\mathbf{z}) \cdot \mathbf{1}_{\{|\mathbf{z} - \mathbf{E}[\mathbf{z}]| \leq t'\}}] + \mathbf{E}[\text{LeakyRelu}(\mathbf{z}) \cdot \mathbf{1}_{\{|\mathbf{z} - \mathbf{E}[\mathbf{z}]| > t'\}}].$$

We consider both terms separately. First, note

$$\mathbf{E}[\text{LeakyRelu}(\mathbf{z}) \cdot \mathbf{1}_{\{|\mathbf{z} - \mathbf{E}[\mathbf{z}]| \leq t'\}}] = \mathbf{E}[\text{LeakyRelu}(\mathbf{z}) \mid |\mathbf{z} - \mathbf{E}[\mathbf{z}]| \leq t'] \cdot \Pr[|\mathbf{z} - \mathbf{E}[\mathbf{z}]| \leq t'].$$

By writing  $\mathbf{z} = \mathbf{E}[\mathbf{z}] + \sqrt{\text{Var}[\mathbf{z}]} \cdot g$ , for  $g \sim N(0, 1)$  and using Lipschitz continuity of the Leaky-Relu

$$\begin{aligned} \mathbf{E}[\text{LeakyRelu}(\mathbf{z}) \mid |\mathbf{z} - \mathbf{E}[\mathbf{z}]| \leq t'] &= \mathbf{E} \left[ \text{LeakyRelu}(\mathbf{E}[\mathbf{z}] + \sqrt{\text{Var}[\mathbf{z}]} \cdot g) \mid \sqrt{\text{Var}[\mathbf{z}]} |g| \leq t' \right] \\ &\in \left[ \text{LeakyRelu}(\mathbf{E}[\mathbf{z}]) - Lt', \text{LeakyRelu}(\mathbf{E}[\mathbf{z}]) + Lt' \right]. \end{aligned} \quad (3)$$

Hence by using sub-Gaussian concentration,

$$\begin{aligned} \mathbf{E}[\text{LeakyRelu}(\mathbf{z}) \cdot \mathbf{1}_{\{|\mathbf{z} - \mathbf{E}[\mathbf{z}]| \leq t'\}}] &\geq \left( 1 - 2e^{-\frac{t'^2}{2\text{Var}[\mathbf{z}]}} \right) (\text{LeakyRelu}(\mathbf{E}[\mathbf{z}]) - Lt'), \\ \mathbf{E}[\text{LeakyRelu}(\mathbf{z}) \cdot \mathbf{1}_{\{|\mathbf{z} - \mathbf{E}[\mathbf{z}]| \leq t'\}}] &\leq \text{LeakyRelu}(\mathbf{E}[\mathbf{z}]) + Lt'. \end{aligned} \quad (4)$$



For the second summand, we apply Cauchy-Schwartz inequality and Lipschitzness of LeakyRelu

$$\begin{aligned}
|\mathbf{E} [\text{LeakyRelu}(\mathbf{z}) \cdot \mathbf{1}_{\{|\mathbf{z} - \mathbf{E}[\mathbf{z}]| > t'\}}]| &\leq \sqrt{\mathbf{E}[|\text{LeakyRelu}(\mathbf{z})|^2] \cdot \Pr[|\mathbf{z} - \mathbf{E}[\mathbf{z}]| > t']} \\
&\leq \sqrt{2L^2 \mathbf{E}[|\mathbf{z}|^2] \exp\left(-\frac{t'^2}{2\text{Var}[\mathbf{z}]}\right)} \\
&\leq \sqrt{2L^2(\mathbf{E}[\mathbf{z}]^2 + \text{Var}[\mathbf{z}]) \exp\left(-\frac{t'^2}{2\text{Var}[\mathbf{z}]}\right)} \\
&\leq L \mathbf{E}[\mathbf{z}] \sqrt{2 \exp\left(-\frac{t'^2}{2\text{Var}[\mathbf{z}]}\right)} + L \sqrt{2\text{Var}[\mathbf{z}] \exp\left(-\frac{t'^2}{2\text{Var}[\mathbf{z}]}\right)}. \quad (5)
\end{aligned}$$

Setting  $t' = 10\sqrt{2\text{Var}[\mathbf{z}] \log n}$ , and combining Equations (4) and (5) results in

$$\mathbf{E}[\text{LeakyRelu}(\mathbf{z})] \leq \text{LeakyRelu}(\mathbf{E}[\mathbf{z}]) + 10L\sqrt{2\text{Var}[\mathbf{z}] \log n} + \frac{L\sqrt{2} \mathbf{E}[\mathbf{z}]}{n^{50}} + \frac{L\sqrt{2\text{Var}[\mathbf{z}]}}{n^{50}}, \quad (6)$$

$$\mathbf{E}[\text{LeakyRelu}(\mathbf{z})] \geq \left(1 - \frac{2}{n^{100}}\right) \left(\text{LeakyRelu}(\mathbf{E}[\mathbf{z}]) - 10L\sqrt{2\text{Var}[\mathbf{z}] \log n}\right) - \frac{L\sqrt{2}(\mathbf{E}[\mathbf{z}] + \sqrt{\text{Var}[\mathbf{z}]})}{n^{50}}. \quad (7)$$

Combining Equations (3), (6), (7) using the choice of  $t$ , we have that with a probability of at least  $1 - O(1/n^{98})$  for all  $i, j \in [n]$

$$\text{LeakyRelu}(\mathbf{z}_{ij}) \leq \text{LeakyRelu}(\mathbf{E}[\mathbf{z}_{ij}]) + 20L(c+1)\sqrt{2\text{Var}[\mathbf{z}_{ij}] \log n} + \frac{L\sqrt{2}(\mathbf{E}[\mathbf{z}_{ij}] + \sqrt{\text{Var}[\mathbf{z}_{ij}]})}{n^{50}}, \quad (8)$$

$$\begin{aligned}
\text{LeakyRelu}(\mathbf{z}_{ij}) &\geq \left(1 - \frac{2}{n^{100}}\right) \left(\text{LeakyRelu}(\mathbf{E}[\mathbf{z}_{ij}]) - 20(c+1)L\sqrt{2\text{Var}[\mathbf{z}_{ij}] \log n}\right) \\
&\quad - \frac{L\sqrt{2}(\mathbf{E}[\mathbf{z}_{ij}] + \sqrt{\text{Var}[\mathbf{z}_{ij}]})}{n^{50}}. \quad (9)
\end{aligned}$$

We henceforth condition on this event. Recall that we have that

$$\text{LeakyRelu}(\mathbf{E}[\mathbf{z}_{ij}]) = \text{LeakyRelu}((\mathbf{a}_1 + \mathbf{a}_2)\mathbf{w}^T \boldsymbol{\mu}) \quad \text{for } i, j \in C_1 \quad (10)$$

$$\text{LeakyRelu}(\mathbf{E}[\mathbf{z}_{ij}]) = \text{LeakyRelu}((\mathbf{a}_1 - \mathbf{a}_2)\mathbf{w}^T \boldsymbol{\mu}) \quad \text{for } i \in C_1, j \in C_0 \quad (11)$$

$$\text{LeakyRelu}(\mathbf{E}[\mathbf{z}_{ij}]) = \text{LeakyRelu}(-(\mathbf{a}_1 - \mathbf{a}_2)\mathbf{w}^T \boldsymbol{\mu}) \quad \text{for } i \in C_0, j \in C_1 \quad (12)$$

$$\text{LeakyRelu}(\mathbf{E}[\mathbf{z}_{ij}]) = \text{LeakyRelu}(-(\mathbf{a}_1 + \mathbf{a}_2)\mathbf{w}^T \boldsymbol{\mu}) \quad \text{for } i, j \in C_0. \quad (13)$$

Using Equations (8)-(13) we have that for  $i, j \in C_1$

$$\text{LeakyRelu}(\mathbf{z}_{ij}) = \text{LeakyRelu}((\mathbf{a}_1 + \mathbf{a}_2)\mathbf{w}^T \boldsymbol{\mu}) \pm 20L(c+1)\sigma\|\mathbf{a}\|\|\mathbf{w}\|\sqrt{2 \log n}.$$

The results for all other cases of  $i, j$  follow similarly. ■

The following statement considers the optimal Bayes classifier for data generated by the Gaussian mixture model.

**Lemma A.7** (See section 6.4 in [And03]). *Let  $(\mathbf{X}, \mathbf{A}) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ . Then, the optimal Bayes classifier for  $\mathbf{X}$  is realized by the linear classifier.*

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

**Proof:** For a given data point  $\mathbf{x}$  and label  $y \in \{0, 1\}$ , the Bayes classifier is given by

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

Note that since the class membership variables  $\epsilon_1, \dots, \epsilon_n \sim \text{Ber}(1/2)$  are independent, we have  $\Pr[y = 0] = \frac{1}{2}$  and  $\Pr[y = 1] = \frac{1}{2}$ . Therefore, by Bayes rule

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

Assume that  $\mathbf{x} \in C_0$ , we have that  $h(\mathbf{x}) = 0$  if and only if  $\Pr[y = 0 \mid \mathbf{x}] \geq 1/2$ . Therefore, if we consider class  $c = 0$  we need that  $\frac{f_{\mathbf{x}|y}(\mathbf{x}|y=1)}{f_{\mathbf{x}|y}(\mathbf{x}|y=0)} \leq 1$ . That is,

$$\frac{f_{\mathbf{x}|y}(\mathbf{x} \mid y = 1)}{f_{\mathbf{x}|y}(\mathbf{x} \mid y = 0)} = \frac{\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \boldsymbol{\mu}\|^2\right)}{\exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} + \boldsymbol{\mu}\|^2\right)} = \exp\left(-\frac{1}{2\sigma^2}(\|\mathbf{x} - \boldsymbol{\mu}\|^2 - \|\mathbf{x} + \boldsymbol{\mu}\|^2)\right) \leq 1,$$

which implies that  $\mathbf{x}^T \boldsymbol{\mu} \leq 0$ . Similarly, for label  $c = 1$  we get that  $\mathbf{x}^T \boldsymbol{\mu} > 0$ . Hence, the Bayes classifier is given by

$$h(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x}^T \boldsymbol{\mu} \leq 0 \\ 1 & \text{if } \mathbf{x}^T \boldsymbol{\mu} > 0 \end{cases},$$

which is a linear classifier. ■

The next lemma relates the fraction of misclassifications of the Bayes optimal classifier to the norm  $\|\boldsymbol{\mu}\|$  (and thus to the distance between the means).

**Lemma A.8.** *Assuming independence of the underlying data  $\mathbf{X}$ , the following holds for the Bayes classifier.*

1. If  $\|\boldsymbol{\mu}\| \geq \sigma\sqrt{2\log n}$  then with a probability of at least  $1 - o_n(1)$ , the Bayes classifier separates the data.
2. If  $\|\boldsymbol{\mu}\| = K\sigma$  for  $\omega(1) < K < \sigma\sqrt{2\log n}$ , then for any  $\kappa > 1$  with a probability of at least  $1 - O\left(\frac{1}{n^{\kappa\Phi'/4}}\right)$  the number of misclassified nodes is  $\Phi'n \left(1 \pm \sqrt{\frac{4\kappa \log n}{\Phi'n}}\right)$ , where  $\Phi' \stackrel{\text{def}}{=} 1 - \Phi(K)$  and  $\Phi$  denote the CDF of standard Gaussian.
3. If  $\|\boldsymbol{\mu}\| = K\sigma$  for  $K = O(1)$ , then with a probability of at least  $1 - o_n(1)$ , the number of misclassified nodes is at least  $\Phi'n(1 - o_n(1))$  where  $\Phi' \geq \left(\frac{K}{K^2+1}\right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{K^2}{2}\right)$ .

**Proof:** Fix  $i \in [n]$  and write

$$\mathbf{x}_i = (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma\mathbf{g}_i \quad \text{where} \quad \mathbf{g}_i \sim N(0, \mathbf{I}).$$

Assume  $\epsilon_i = 0$  and consider the Bayes classifier from Lemma A.7. Then, the probability of misclassification is

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

where the last equality follows from the fact that  $\frac{\mathbf{g}_i^T \boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} \sim N(0, 1)$ .

Suppose  $\|\boldsymbol{\mu}\| \geq \sigma\sqrt{2\log n}$ . By using standard tail bounds for normal distribution [Ver18],

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

Therefore, 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  $i \in C_1$ , and an application of a union bound on the events that there is  $i \in [n]$  which is misclassified finishes the proof of this case.

Next, consider the case where  $\|\boldsymbol{\mu}\| = K\sigma$  for  $\omega(1) < K < \sigma\sqrt{2\log n}$ . We have that for class  $\epsilon_i = 0$  the misclassification probability is

$$\Phi' \stackrel{\text{def}}{=} 1 - \Phi(K).$$

Therefore, by applying additive Chernoff bound, we have that for any  $\kappa > 1$

$$\Pr \left[ \sum_{i \in C_0} \mathbf{1}_{i \text{ misclassified}} \notin \left( \frac{\Phi' n (1 \pm o(1))}{2} \pm \sqrt{\kappa \Phi' n \log n} \right) \right] \leq \frac{2}{n^{\kappa \Phi'/4}},$$

and similarly for  $\epsilon_i = 1$ . Applying a union bound over the two classes finishes the proof of this case.

Finally, consider the case where  $\|\boldsymbol{\mu}\| = K\sigma$  for some constant  $K > 0$ . For class  $\epsilon_i = 0$ , we have that the misclassification probability is lower-bounded by

$$\Phi' \stackrel{\text{def}}{=} 1 - \Phi(K) \geq \left( \frac{K}{K^2 + 1} \right) \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{K^2}{2}\right) = \Omega(1).$$

Therefore, by applying the Chernoff bound, we have that with a probability of at least  $1 - o(1)$  we have that

$$\Pr \left[ \sum_{i \in C_0} \mathbf{1}_{i \text{ misclassified}} < \frac{\Phi' n}{2} (1 - o(1)) \right] = 1/\text{poly}(n).$$

By a similar argument for  $\epsilon_i = 1$  and a union bound, the result follows. ■

## A.2 Proof of Theorem 1 and its implications

In this subsection, we will show that there exists a choice of attention architecture  $\Psi$  that allows the model to ignore all inter-class edges and keep only intra-class edges. We restate the theorem for convenience

**Theorem.** *Suppose that  $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ . Then, there exists a choice of attention architecture  $\Psi$  such that with a probability of at least  $1 - o_n(1)$  over the data  $(\mathbf{X}, \mathbf{A}) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$  it holds that  $\Psi$  separates intra-edges from inter-edges.*

**Proof:** We consider as an ansatz the following two layer architecture  $\Psi$ .

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

where  $R > 0$  is an arbitrary scaling parameter. The output of the attention model is defined as

$$\mathbf{r} \cdot \text{LeakyRelu} \left( \mathbf{S} \cdot \begin{bmatrix} \tilde{\mathbf{w}}^T \mathbf{X}_i \\ \tilde{\mathbf{w}}^T \mathbf{X}_j \end{bmatrix} \right).$$

Denote the input of  $\text{LeakyRelu}(\cdot)$  by  $\Delta_{ij} \stackrel{\text{def}}{=} \mathbf{S} \cdot \begin{bmatrix} \tilde{\mathbf{w}}^T \mathbf{X}_i \\ \tilde{\mathbf{w}}^T \mathbf{X}_j \end{bmatrix} \in \mathbb{R}^4$ , and note that for  $t \in [4]$ , we have  $(\Delta_{ij})_t = \mathbf{S}_{t,1} \tilde{\mathbf{w}}^T \mathbf{X}_i + \mathbf{S}_{t,2} \tilde{\mathbf{w}}^T \mathbf{X}_j$ . Recall that the random variable  $(\Delta_{ij})_t = \mathbf{S}_{t,1} \tilde{\mathbf{w}}^T \mathbf{X}_i + \mathbf{S}_{t,2} \tilde{\mathbf{w}}^T \mathbf{X}_j$  is distributed as follows:

$$(\Delta_{ij})_t = \mathbf{S}_{t,1} \tilde{\mathbf{w}}^T \mathbf{X}_i + \mathbf{S}_{t,2} \tilde{\mathbf{w}}^T \mathbf{X}_j \sim \begin{cases} N((\mathbf{S}_{t,1} + \mathbf{S}_{t,2}) \tilde{\mathbf{w}}^T \boldsymbol{\mu}, \|\mathbf{S}_t\|^2 \sigma^2) & \text{if } i, j \in C_1 \\ N((\mathbf{S}_{t,1} - \mathbf{S}_{t,2}) \tilde{\mathbf{w}}^T \boldsymbol{\mu}, \|\mathbf{S}_t\|^2 \sigma^2) & \text{if } i \in C_1, j \in C_0 \\ N(-(\mathbf{S}_{t,1} - \mathbf{S}_{t,2}) \tilde{\mathbf{w}}^T \boldsymbol{\mu}, \|\mathbf{S}_t\|^2 \sigma^2) & \text{if } i \in C_0, j \in C_1 \\ N(-(\mathbf{S}_{t,1} + \mathbf{S}_{t,2}) \tilde{\mathbf{w}}^T \boldsymbol{\mu}, \|\mathbf{S}_t\|^2 \sigma^2) & \text{if } i, j \in C_0 \end{cases}.$$

We work on each of the four coordinates separately. Assume  $t = 1$ . In such a case, we have that

$$(\Delta_{ij})_1 \sim \begin{cases} N(2\|\boldsymbol{\mu}\|, 2\sigma^2) & \text{if } i, j \in C_1 \\ N(0, 2\sigma^2) & \text{if } i \in C_1, j \in C_0 \\ N(0, 2\sigma^2) & \text{if } i \in C_0, j \in C_1 \\ N(-2\|\boldsymbol{\mu}\|, 2\sigma^2) & \text{if } i, j \in C_0 \end{cases}.$$

Using our results for the LeakyRelu concentration in Lemma A.6 and our assumption on the norm of  $\boldsymbol{\mu}$ , we have that with a probability of at least  $1 - o_n(1)$ ,

$$\text{LeakyRelu}((\Delta_{ij})_1) = \begin{cases} 2\|\boldsymbol{\mu}\|(1 \pm o(1)) & \text{if } i, j \in C_1 \\ \pm 2C\sigma\sqrt{\log n} & \text{if } i \in C_1, j \in C_0 \\ \pm 2C\sigma\sqrt{\log n} & \text{if } i \in C_0, j \in C_1 \\ -2\beta\|\boldsymbol{\mu}\|(1 \pm o(1)) & \text{if } i, j \in C_0 \end{cases}.$$

Using a similar argument we get

$$\begin{aligned} \text{LeakyRelu}((\Delta_{ij})_2) &= \begin{cases} -2\beta\|\boldsymbol{\mu}\|(1 \pm o(1)) & \text{if } i, j \in C_1 \\ \pm 2C\sigma\sqrt{\log n} & \text{if } i \in C_1, j \in C_0 \\ \pm 2C\sigma\sqrt{\log n} & \text{if } i \in C_0, j \in C_1 \\ 2\|\boldsymbol{\mu}\|(1 \pm o(1)) & \text{if } i, j \in C_0 \end{cases}, \\ \text{LeakyRelu}((\Delta_{ij})_3) &= \begin{cases} \pm 2C\sigma\sqrt{\log n} & \text{if } i, j \in C_1 \\ 2\|\boldsymbol{\mu}\|(1 \pm o(1)) & \text{if } i \in C_1, j \in C_0 \\ -2\beta\|\boldsymbol{\mu}\|(1 \pm o(1)) & \text{if } i \in C_0, j \in C_1 \\ \pm 2C\sigma\sqrt{\log n} & \text{if } i, j \in C_0 \end{cases}, \\ \text{LeakyRelu}((\Delta_{ij})_4) &= \begin{cases} \pm 2C\sigma\sqrt{\log n} & \text{if } i, j \in C_1 \\ -2\beta\|\boldsymbol{\mu}\|(1 \pm o(1)) & \text{if } i \in C_1, j \in C_0 \\ 2\|\boldsymbol{\mu}\|(1 \pm o(1)) & \text{if } i \in C_0, j \in C_1 \\ \pm 2C\sigma\sqrt{\log n} & \text{if } i, j \in C_0 \end{cases}. \end{aligned}$$

Applying a union bound over the four coordinates of the vector  $\Delta_{ij}$ , we get that the above event holds with probability at least  $1 - o_n(1)$  for all  $t$ .

Next, we examine the second layer of the architecture. Suppose  $i, j \in C_1$  so that

$$\text{LeakyRelu}(\Delta_{ij}) = [2\|\boldsymbol{\mu}\|(1 \pm o(1)) \quad -2\beta\|\boldsymbol{\mu}\|(1 \pm o(1)) \quad \pm 2C\sigma\sqrt{\log n} \quad \pm 2C\sigma\sqrt{\log n}].$$

Then,

$$\mathbf{r} \cdot \text{LeakyRelu}(\Delta_{ij}) = 2R\|\boldsymbol{\mu}\|(1 - \beta)(1 \pm o(1)) \pm 4RC\sigma\sqrt{\log n} = 2R\|\boldsymbol{\mu}\|(1 - \beta)(1 \pm o(1)).$$

By applying a similar reasoning to the over pairs

$$\mathbf{r} \cdot \text{LeakyRelu}(\Delta_{ij}) = \begin{cases} 2R\|\boldsymbol{\mu}\|(1-\beta)(1 \pm o(1)) & \text{if } i, j \in C_1 \\ 2R\|\boldsymbol{\mu}\|(1-\beta)(1 \pm o(1)) & \text{if } i, j \in C_0 \\ -2R\|\boldsymbol{\mu}\|(1-\beta)(1 \pm o(1)) & \text{if } i \in C_1, j \in C_0 \\ -2R\|\boldsymbol{\mu}\|(1-\beta)(1 \pm o(1)) & \text{if } i \in C_0, j \in C_1 \end{cases},$$

and the proof is complete. ■

Next, we define a high probability event under which we can obtain some interesting corollaries.

**Definition A.9.** Event  $\mathcal{E}^*$  is the intersection of the following events over the randomness of  $\mathbf{A}$  and  $\{\epsilon_i\}_i$  and  $\mathbf{X}_i$ ,

1.  $\mathcal{E}_1$  is the event that  $|C_0| = \frac{n}{2} \pm O(\sqrt{n \log n})$  and  $|C_1| = \frac{n}{2} \pm O(\sqrt{n \log n})$ .
2.  $\mathcal{E}_2$  is the event that for each  $i \in [n]$ ,  $\mathbf{D}_{ii} = \frac{n(p+q)}{2} \left(1 \pm \frac{10}{\sqrt{\log n}}\right)$ .
3.  $\mathcal{E}_3$  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  $|C_1 \cap N_i| = \mathbf{D}_{ii} \cdot \frac{(1-\epsilon_i)q + \epsilon_i p}{p+q} \left(1 \pm \frac{10}{\sqrt{\log n}}\right)$ .
4.  $\mathcal{E}_4$  is the event that for each  $i \in [n]$ ,  $|\tilde{\mathbf{w}}^T \mathbf{X}_i - \mathbf{E}[\tilde{\mathbf{w}}^T \mathbf{X}_i]| \leq 10\sigma\sqrt{\log n}$ .

The next lemma is a straightforward application of Chernoff bound and a union bound (originally proved in [BFJ21])

**Lemma A.10.** With probability at least  $1 - 1/\text{poly}(n)$  event  $\mathcal{E}^*$  holds.

We present two corollaries of Theorem A.2. The first is regarding the values of  $\gamma_{ij}$ .

**Corollary 8.** Suppose that  $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ . Then, there exists a choice of attention architecture  $\Psi$  such that with probability at least  $1 - o_n(1)$  over the data  $(\mathbf{X}, \mathbf{A}) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$  it holds that

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

**Proof:** The proof is straightforward. First, we use Theorem A.2 (pick any  $R$  such that  $R\|\boldsymbol{\mu}\| = \omega(1)$ ), Lemma A.10 and apply a union bound. Next, we use the definition of the attention coefficients to conclude the result. ■

Next, we show that the model separates the nodes for any choice of  $p, q$  satisfying Assumption 1.

**Corollary 9.** Suppose that  $\|\boldsymbol{\mu}\| = \omega(\sigma\sqrt{\log n})$ . Then, there exists a choice of attention architecture  $\Psi$  such that with probability at least  $1 - o_n(1)$  over the data  $(\mathbf{X}, \mathbf{A}) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ , GAT separates the data for any  $p, q$  satisfying Assumption 1.

**Proof:** Consider the same ansatz described in Theorem A.2 (pick any  $R$  such that  $R\|\boldsymbol{\mu}\| = \omega(1)$ ). Assume that  $i \in C_1$  (the case for  $i \in C_0$  follows similarly), and let

$$\hat{\mathbf{x}}_i \stackrel{\text{def}}{=} \sum_{j \in N_i} \gamma_{ij} \tilde{\mathbf{w}}^T \mathbf{X}_j.$$

We would like to compute the conditional mean and variance of  $\hat{\mathbf{x}}_i$  given  $\mathcal{E}^*$ . By using Corollary 8 we have

$$\begin{aligned}
\mathbf{E} \left[ \sum_{j \in N_i} \gamma_{ij} \tilde{\mathbf{w}}^T \mathbf{X}_j \middle| \mathcal{E}^* \right] &= \mathbf{E} \left[ \sum_{j \in C_0 \cap N_i} \gamma_{ij} \tilde{\mathbf{w}}^T \mathbf{X}_j + \sum_{j \in C_1 \cap N_i} \gamma_{ij} \tilde{\mathbf{w}}^T \mathbf{X}_j \middle| \mathcal{E}^* \right] \\
&\leq |C_1 \cap N_i| \left( \frac{2}{np} (1 \pm o(1)) (\|\boldsymbol{\mu}\|_2 + 10\sigma\sqrt{\log n}) \right) + |C_0 \cap N_i| \left( o\left(\frac{1}{n(p+q)}\right) (-\|\boldsymbol{\mu}\|_2 + 10\sigma\sqrt{\log n}) \right) \\
&= \|\boldsymbol{\mu}\|_2 (1 \pm o(1)) + 10\sigma\sqrt{\log n} - \frac{nq(1 \pm o(1))}{2 \cdot \omega(n(p+q))} (\|\boldsymbol{\mu}\|_2 - 10\sigma\sqrt{\log n}) \\
&= \|\boldsymbol{\mu}\|_2 (1 \pm o(1)).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbf{E} \left[ \sum_{j \in N_i} \gamma_{ij} \tilde{\mathbf{w}}^T \mathbf{X}_j \middle| \mathcal{E}^* \right] &\geq \|\boldsymbol{\mu}\|_2 (1 \pm o(1)) - 10\sigma\sqrt{\log n} - \frac{nq(1 \pm o(1))}{2 \cdot \omega(n(p+q))} (\|\boldsymbol{\mu}\|_2 + 10\sigma\sqrt{\log n}) \\
&= \|\boldsymbol{\mu}\|_2 (1 \pm o(1)).
\end{aligned}$$

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

Next, we claim that for each  $i \in [n]$  the random variable  $\hat{\mathbf{x}}_i$  given  $\mathcal{E}^*$  is sub-Gaussian with a small sub-Gaussian constant compared to the above expectation.

**Lemma A.11.** *Conditioned on  $\mathcal{E}^*$ , the random variables  $\{\hat{\mathbf{x}}_i\}_i$  are sub-Gaussian with parameter  $\tilde{\sigma}^2 = O\left(\frac{\sigma^2}{np}\right)$ .*

**Proof:** For  $i \in [n]$ , write  $\mathbf{X}_i = (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma\mathbf{g}_i$  where  $\mathbf{g}_i \sim N(0, \mathbf{I}_d)$ ,  $\epsilon_i = 0$  if  $i \in C_0$  and  $\epsilon_i = 1$  if  $i \in C_1$ . Let us consider  $\hat{\mathbf{x}}_i$  as a function of  $\mathbf{g} = [\mathbf{g}_1 \| \mathbf{g}_2 \| \cdots \| \mathbf{g}_n] \in \mathbb{R}^{nd}$ , where  $\|$  denotes vertical concatenation. That is, let us consider the function

$$\hat{\mathbf{x}}_i = f_i(\mathbf{g}) \stackrel{\text{def}}{=} \sum_{j \in N_i} \gamma_{ij}(\mathbf{g}) \tilde{\mathbf{w}}^T ((2\epsilon_j - 1)\boldsymbol{\mu} + \sigma\mathbf{g}_j), \quad i \in [n].$$

Because  $\mathbf{g} \sim N(0, \mathbf{I}_{nd})$ , in order to see that  $\hat{\mathbf{x}}_i$  given  $\mathcal{E}^*$  is sub-Gaussian for each  $i \in [n]$ , it suffices to show that the function  $f_i : \mathbb{R}^{nd} \rightarrow \mathbb{R}$  is Lipschitz over the subset  $E \subseteq \mathbb{R}^{nd}$  identified by  $\mathcal{E}^*$  and the relation  $\mathbf{X}_i = (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma\mathbf{g}_i$ , that is,  $E \stackrel{\text{def}}{=} \{\mathbf{g} \in \mathbb{R}^{nd} \mid |\tilde{\mathbf{w}}^T \mathbf{g}_i| \leq 10\sigma\sqrt{\log n}, \forall i \in [n]\}$ . In particular, we will show that, conditioning on the event  $\mathcal{E}^*$  (which imposes the restriction that  $\mathbf{g} \in E$ ), the Lipschitz constant  $L_{f_i}$  of  $f_i$  satisfies  $L_{f_i} = O\left(\frac{\sigma}{\sqrt{np}}\right)$  for all  $i \in [n]$ , and hence proving the claim.

To compute the Lipschitz constant of  $f_i(\mathbf{g})$  for  $i \in [n]$ , let us denote  $\mathbf{X} = [\mathbf{X}_1 \| \mathbf{X}_2 \| \cdots \| \mathbf{X}_n]$  and consider the function

$$\tilde{f}_i(\mathbf{X}) \stackrel{\text{def}}{=} \sum_{j \in N_i} \gamma_{ij}(\mathbf{X}) \tilde{\mathbf{w}}^T \mathbf{X}_j, \quad i \in [n]$$

Let us assume without loss of generality that  $i \in C_0$  (the case for  $i \in C_1$  yields the same result and is obtained identically). Conditioning on the event  $\mathcal{E}^*$ , which imposes the restriction that  $\mathbf{X} \in \tilde{E}$  where  $\tilde{E} \stackrel{\text{def}}{=} \{\mathbf{X} \in \mathbb{R}^{nd} \mid \|\mathbf{X}_i - (2\epsilon_i - 1)\boldsymbol{\mu}\| \leq 10\sigma\sqrt{\log n}, \forall i \in [n]\}$ , we know by Corollary 8 that  $\gamma_{ij}(\mathbf{X}) = \frac{2}{np} (1 \pm o(1))$  if  $j \in C_0$  and  $\gamma_{ij}(\mathbf{X}) = \frac{2}{np} \exp(-\Theta(R\|\boldsymbol{\mu}\|)) (1 \pm o(1))$  if  $j \in C_1$ . Conditioning on  $\mathcal{E}^*$  (which restricts  $\mathbf{X}, \mathbf{X}' \in \tilde{E}$ ),



and recalling that  $R$  satisfies  $R\|\boldsymbol{\mu}\| = \omega(1)$ , we get

$$\begin{aligned}
\left| \tilde{f}_i(\mathbf{X}) - \tilde{f}_i(\mathbf{X}') \right| &= \left| \sum_{j \in N_i \cap C_0} \frac{2(1 \pm o(1))}{np} \tilde{\mathbf{w}}^T(\mathbf{X}_j - \mathbf{X}'_j) + \sum_{j \in N_i \cap C_1} \frac{2(1 \pm o(1))}{np} \cdot e^{-\Theta(\|\boldsymbol{\mu}\|)} \tilde{\mathbf{w}}^T(\mathbf{X}_j - \mathbf{X}'_j) \right| \\
&= \left| \begin{bmatrix} \frac{2}{np}(1 \pm o(1))\tilde{\mathbf{w}} & \text{if } j \in N_i \cap C_0 \\ \frac{2}{np}\exp(-\Theta(R\|\boldsymbol{\mu}\|))(1 \pm o(1))\tilde{\mathbf{w}} & \text{if } j \in N_i \cap C_1 \\ 0 & \text{if } j \notin N_i \end{bmatrix}_{j \in [n]}^T (\mathbf{X} - \mathbf{X}') \right| \\
&\leq \left\| \begin{bmatrix} \frac{2}{np}(1 \pm o(1))\tilde{\mathbf{w}} & \text{if } j \in N_i \cap C_0 \\ \frac{2}{np}\exp(-\Theta(R\|\boldsymbol{\mu}\|))(1 \pm o(1))\tilde{\mathbf{w}} & \text{if } j \in N_i \cap C_1 \\ 0 & \text{if } j \notin N_i \end{bmatrix}_{j \in [n]} \right\|_2 \|\mathbf{X} - \mathbf{X}'\|_2 \\
&\leq \sqrt{\frac{2}{np}}(1 + o(1))\|\tilde{\mathbf{w}}\|_2 \|\mathbf{X} - \mathbf{X}'\|_2 \\
&= \sqrt{\frac{2}{np}}(1 + o(1)) \|\mathbf{X} - \mathbf{X}'\|_2.
\end{aligned}$$

This shows the Lipschitz constant of  $\tilde{f}_i(\mathbf{X})$  over  $\tilde{E}$  satisfies  $L_{\tilde{f}_i} = O\left(\frac{1}{\sqrt{np}}\right)$ . On the other hand, by viewing  $\mathbf{X}$  as a function of  $\mathbf{g}$ , it is straightforward to see that the function  $h(\mathbf{g}) : \mathbb{R}^{nd} \rightarrow \mathbb{R}^{nd}$  defined by  $h(\mathbf{g}) \stackrel{\text{def}}{=} \mathbf{X}(\mathbf{g})$  has Lipschitz constant  $L_h = \sigma$ , as

$$\|h(\mathbf{g}) - h(\mathbf{g}')\|_2 = \left\| \begin{bmatrix} \vdots \\ (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma\mathbf{g}_i \\ \vdots \end{bmatrix}_{i \in [n]} - \begin{bmatrix} \vdots \\ (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma\mathbf{g}'_i \\ \vdots \end{bmatrix}_{i \in [n]} \right\|_2 = \sigma \|\mathbf{g} - \mathbf{g}'\|_2.$$

Therefore, since  $f_i(\mathbf{g}) = \tilde{f}_i(h(\mathbf{g}))$  and  $\mathbf{g} \in E$  if and only if  $\mathbf{X} \in \tilde{E}$ , we have that, conditioning on  $\mathcal{E}^*$ , the function  $\hat{\mathbf{x}}_i = f_i(\mathbf{g})$  is Lipschitz continuous with Lipschitz constant  $L_{f_i} = L_{\tilde{f}_i} L_h = O\left(\frac{\sigma}{\sqrt{np}}\right)$ . Now, because  $\mathbf{g} \sim N(0, \mathbf{I}_{nd})$ , we know that  $\hat{\mathbf{x}}_i$  is sub-Gaussian with sub-Gaussian constant  $\tilde{\sigma}^2 = L_{f_i}^2 = O\left(\frac{\sigma^2}{np}\right)$ . Since our choice of  $i$  was arbitrary, this proves the claim.  $\blacksquare$

Now, we have all the tools to finish the proof of the theorem. We bound the probability of misclassification

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

for  $t < |\mathbf{E}[\hat{\mathbf{x}}_i]| = \|\boldsymbol{\mu}\|(1 \pm o(1))$ . By Lemma A.11, picking  $t = \Theta\left(\sigma\sqrt{\log|C_0|}\right)$  and applying Lemma A.2 implies that the above probability is  $1/\text{poly}(n)$ .

Similarly for class  $C_1$  we have that the misclassification probability is

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

for  $t < \mathbf{E}[\hat{\mathbf{x}}_i]$ . Picking  $t = \Theta\left(\sigma\sqrt{\log|C_1|}\right)$  and applying Lemma A.2 and a union bound over the misclassification probabilities of both classes conclude the proof of the corollary.  $\blacksquare$

### A.3 Proof of Proposition 4

We start by restating the proposition for convenience.

**Proposition.** Suppose  $\|\boldsymbol{\mu}\| = \Omega(\sigma\sqrt{\log n})$ , and let  $(\mathbf{X}, \cdot) \sim \text{CSBM}(n, \cdot, \cdot, \boldsymbol{\mu}, \sigma^2)$ . Then, the Bayes optimal classifier is realized by a linear classifier which achieves perfect node separability with probability at least  $1 - o_n(1)$  over  $\mathbf{X}$ .

**Proof:** The proof follows by first applying Lemma A.7 to deduce that a linear classifier obtains the optimal performance and then applying Case (1) of Lemma A.8. ■

#### A.4 Proof of Theorem 5

The goal of the attention mechanism is to separate the pairs of nodes  $(\mathbf{X}_i, \mathbf{X}_j)$  based on whether  $i, j$  are in the same class or different classes. We say  $i \sim j$  if both  $i$  and  $j$  are in the same class  $C_0$  or  $C_1$ , and  $i \not\sim j$  otherwise. Let  $\mathbf{X}'_{ij}$  denote the vector obtained as a result of concatenating  $\mathbf{X}_i$  and  $\mathbf{X}_j$ , i.e.,  $\mathbf{X}'_{ij} = \begin{pmatrix} \mathbf{X}_i \\ \mathbf{X}_j \end{pmatrix}$ . Then we would like to analyze all classifiers with the property

$$y = h(\mathbf{X}'_{ij}) = \begin{cases} 0 & i \not\sim j \\ 1 & i \sim j \end{cases}.$$

To comment on the limitations of all such classifiers, it is sufficient to analyze the Bayes classifier for this data model, since by definition a Bayes classifier is optimal. The following lemma describes the optimal classifier for this classification task.

**Lemma A.12.** The optimal (Bayes) classifier that serves as the attention mechanism for the pairs  $\mathbf{X}'_{ij}$  is realized by the following function.

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

**Proof:** Recall that  $|C_0| = |C_1| = \frac{n}{2}$ . Hence,  $\mathbf{X}'_{ij}$  is a uniform mixture of four  $2d$ -dimensional Gaussian distributions. Let  $\boldsymbol{\mu}' = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix}$  and  $\boldsymbol{\nu}' = \begin{pmatrix} \boldsymbol{\mu} \\ -\boldsymbol{\mu} \end{pmatrix}$ . Then the four Gaussian distributions are

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

The optimal classifier is then given by

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

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

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

Suppose that  $\mathbf{x} = \mathbf{X}'_{ij}$  such that  $i \not\sim j$ . Then  $h^*(\mathbf{x}) = 0$  if and only if  $\Pr[y = 0 \mid \mathbf{x}] \geq \frac{1}{2}$ . Hence, for  $c = 0$  we require  $\frac{p \cdot f_{\mathbf{x}|y}(\mathbf{x} \mid y=1)}{q \cdot f_{\mathbf{x}|y}(\mathbf{x} \mid y=0)} \leq 1$ , which implies that  $\frac{p}{q} \frac{\cosh\left(\frac{1}{\sigma^2} \mathbf{x}^T \boldsymbol{\mu}'\right)}{\cosh\left(\frac{1}{\sigma^2} \mathbf{x}^T \boldsymbol{\nu}'\right)} \leq 1$ . Similarly, we obtain the reverse condition for  $h^*(\mathbf{x}) = 1$ . ■

The next theorem uses [A.12](#) to precisely quantify the misclassification of node pairs that the attention mechanism exhibits. In particular, part 1 of the theorem implies that if  $\|\boldsymbol{\mu}\|_2$  is linear in the standard deviation,  $\sigma$ , then with overwhelming probability the attention mechanism fails to distinguish a constant fraction of inter-edge pairs from the intra-edge pairs.

Furthermore, part 2 of the theorem characterizes a regime for the inter-edge probability  $q$  where the attention mechanism fails to distinguish at least one inter-edge node pair, by providing a lower bound on  $q$  in terms of the scale at which the distance between the means grows compared to the standard deviation  $\sigma$ . This aligns with the intuition that as we increase the distance between the means, it gets easier for the attention mechanism to correctly distinguish the node pairs. However, if  $q$  is also increased in the right proportion, or in other words, if the noise in the graph is increased, then the attention mechanism will still fail to correctly distinguish at least one of the inter-edge node pairs.

We restate the theorem for the readers' convenience.

**Theorem** (Restatement of [Theorem 5](#)). *Assume that  $q = \Omega\left(\frac{\log^2 n}{n}\right)$  and let  $\Psi$  be any attention mechanism. Let  $\Phi_c(\cdot) \stackrel{\text{def}}{=} 1 - \Phi(\cdot)$ , where  $\Phi$  is the standard normal CDF. Then for any  $K > 0$  if  $\|\boldsymbol{\mu}\|_2 = K\sigma$  then we have:*

1. *For any  $c > 0$ , with probability at least  $1 - O(n^{-c})$ , the attention mechanism  $\Psi$  fails to distinguish at least  $2\Phi_c(K)^2$  fraction of the inter-edge pairs  $(\mathbf{X}_i, \mathbf{X}_j), i \asymp j$  from the intra-edge pairs  $(\mathbf{X}_i, \mathbf{X}_j), i \sim j$ .*
2. *For any  $\kappa > 1$  if  $q > \frac{\kappa \log^2 n}{n \Phi_c(K)^2}$ , then with probability at least  $1 - O\left(\frac{1}{n^{\frac{\kappa}{4} \Phi_c(K)^2 \log n}}\right)$ , at least one inter-edge pair is indistinguishable from the intra-edge pairs under  $\Psi$ .*

**Proof:** From [A.12](#), we observe that for successful classification by the optimal classifier, we need

$$\begin{aligned} p \cosh\left(\frac{\mathbf{x}^T \boldsymbol{\mu}'}{\sigma^2}\right) &\leq q \cosh\left(\frac{\mathbf{x}^T \boldsymbol{\nu}'}{\sigma^2}\right) && \text{for } i \asymp j, \\ p \cosh\left(\frac{\mathbf{x}^T \boldsymbol{\mu}'}{\sigma^2}\right) &> q \cosh\left(\frac{\mathbf{x}^T \boldsymbol{\nu}'}{\sigma^2}\right) && \text{for } i \sim j. \end{aligned}$$

Let us focus on the case where  $i \asymp j$ . We want

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

In the first implication, we used that  $q \leq p$ , while the second implication follows from the fact that  $\cosh(a) \leq \cosh(b) \implies |a| \leq |b|$  for all  $a, b \in \mathbb{R}$ . Therefore, we can upper bound the probability that  $\mathbf{X}'_{ij}$  is correctly classified by the probability of the event  $|\mathbf{X}'_{ij}{}^T \boldsymbol{\mu}'| \leq |\mathbf{X}'_{ij}{}^T \boldsymbol{\nu}'|$ . Writing  $\mathbf{X}_i = \boldsymbol{\mu} + \sigma \mathbf{g}_i$  and  $\mathbf{X}_j = -\boldsymbol{\mu} + \sigma \mathbf{g}_j$ , we have that for  $i \in C_1$  and  $j \in C_0$ ,

$$\begin{aligned} \Pr[h^*(\mathbf{X}'_{ij}) = 0] &\leq \Pr\left[|\mathbf{X}'_{ij}{}^T \boldsymbol{\mu}'| \leq |\mathbf{X}'_{ij}{}^T \boldsymbol{\nu}'|\right] \\ &= \Pr\left[|\mathbf{X}_i^T \boldsymbol{\mu} + \mathbf{X}_j^T \boldsymbol{\mu}| \leq |\mathbf{X}_i^T \boldsymbol{\mu} - \mathbf{X}_j^T \boldsymbol{\mu}|\right] \\ &= \Pr\left[\sigma |\mathbf{g}_i^T \boldsymbol{\mu} + \mathbf{g}_j^T \boldsymbol{\mu}| \leq |\pm 2\|\boldsymbol{\mu}\|_2^2 + \sigma \mathbf{g}_i^T \boldsymbol{\mu} - \sigma \mathbf{g}_j^T \boldsymbol{\mu}|\right] \\ &\leq \Pr\left[|\mathbf{g}_i^T \hat{\boldsymbol{\mu}} + \mathbf{g}_j^T \hat{\boldsymbol{\mu}}| - |\mathbf{g}_i^T \hat{\boldsymbol{\mu}} - \mathbf{g}_j^T \hat{\boldsymbol{\mu}}| \leq \frac{2\|\boldsymbol{\mu}\|_2}{\sigma}\right] \\ &= \Pr\left[|\mathbf{g}_i^T \hat{\boldsymbol{\mu}} + \mathbf{g}_j^T \hat{\boldsymbol{\mu}}| - |\mathbf{g}_i^T \hat{\boldsymbol{\mu}} - \mathbf{g}_j^T \hat{\boldsymbol{\mu}}| \leq 2K\right], \end{aligned}$$

where  $\hat{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}$ . In the second to last step above, we used triangle inequality to pull  $2\|\boldsymbol{\mu}\|_2^2$  outside the absolute value, while in the last equation we use  $\|\boldsymbol{\mu}\|_2 = K\sigma$ .

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

$$\begin{aligned} \Pr[h^*(\mathbf{X}'_{ij}) = 0] &\leq \Pr[|z_i + z_j| - |z_i - z_j| \leq 2K] \\ &= \Pr[|z_i + z_j| - |z_i - z_j| \leq 2K, |z_i| \leq K] + \Pr[|z_i + z_j| - |z_i - z_j| \leq 2K, |z_i| > K] \\ &= \Pr[|z_i| \leq K] + \Phi(K) \Pr[|z_i| > K]. \end{aligned} \quad (14)$$

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

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

hence,  $\Pr[|z_i + z_j| - |z_i - z_j| \leq 2K, |z_i| \leq K] = \Pr[|z_i| \leq K]$ . On the other hand, for  $|z_i| > K$ , we look at each case, conditioned on the events  $z_i > K$  and  $z_i < -K$  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  $|z_i + z_j| - |z_i - z_j| \leq 2K$ , and analyze the cases in detail.

Conditioned on  $z_i < -K$ :

$$\begin{aligned} \Pr[E, z_i + z_j \geq 0, z_i - z_j \geq 0 \mid z_i < -K] &= \Pr[z_j \leq z_i, z_j \geq -z_i \mid z_i < -K] = 0, \\ \Pr[E, z_i + z_j \geq 0, z_i - z_j < 0 \mid z_i < -K] &= \Pr[z_j > |z_i|, z_i \leq K \mid z_i < -K] = \Phi(z_i), \\ \Pr[E, z_i + z_j < 0, z_i - z_j \geq 0 \mid z_i < -K] &= \Pr[z_j < -|z_i|, z_i \geq -K \mid z_i < -K] = 0, \\ \Pr[E, z_i + z_j < 0, z_i - z_j < 0 \mid z_i < -K] &= \Pr[z_i < z_j < -z_i, z_j > -K \mid z_i < -K] = \Phi(K) - \Phi(z_i). \end{aligned}$$

The sum of the four probabilities in the above display is  $\Pr[E \mid z_i < -K] = \Phi(K)$ . Similarly, we analyze the other case.

Conditioned on  $z_i > K$ :

$$\begin{aligned} \Pr[E, z_i + z_j \geq 0, z_i - z_j \geq 0 \mid z_i > K] &= \Pr[-z_i \leq z_j \leq z_i, z_j \leq K \mid z_i > K] = \Phi(K) - \Phi_c(z_i), \\ \Pr[E, z_i + z_j \geq 0, z_i - z_j < 0 \mid z_i > K] &= \Pr[z_j > |z_i|, z_i \leq K \mid z_i > K] = 0, \\ \Pr[E, z_i + z_j < 0, z_i - z_j \geq 0 \mid z_i > K] &= \Pr[z_j < -|z_i|, z_i \geq -K \mid z_i > K] = \Phi_c(z_i), \\ \Pr[E, z_i + z_j < 0, z_i - z_j < 0 \mid z_i > K] &= \Pr[z_j < -z_i, z_j > z_i \mid z_i > K] = 0. \end{aligned}$$

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

$$\Pr[|z_i + z_j| - |z_i - z_j| \leq 2K \mid |z_i| > K] = \Phi(K),$$

which justifies 14.

Next, note that  $\Pr[|z_i| \leq K] = \Phi(K) - \Phi_c(K)$  and  $\Pr[|z_i| > K] = 2\Phi_c(K)$ , so we have from 14 that

$$\Pr[h^*(\mathbf{X}'_{ij}) = 0] \leq \Phi(K) - \Phi_c(K) + 2\Phi_c(K)\Phi(K) = 1 - 2\Phi_c(K) + 2\Phi_c(K)\Phi(K) = 1 - 2\Phi_c(K)^2.$$

Thus,  $\mathbf{X}'_{ij}$  is misclassified with probability at least  $2\Phi_c(K)^2$ .

We will now construct sets of pairs with mutually independent elements, such that the union of those sets covers all inter-edge pairs. This will enable us to use a concentration argument that computes the fraction of the inter-edge pairs that are misclassified. Since the graph operations are permutation invariant, let us

assume for simplicity that  $C_0 = \{1, \dots, \frac{n}{2}\}$  and  $C_1 = \{\frac{n}{2} + 1, \dots, n\}$  for an even number of nodes  $n$ . Also define the function

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

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

$$S_l = \{(X_{m(i,l)}, X_{i+\frac{n}{2}}) \text{ for all } i \in C_0 \cap N_{m(i,l)}\}.$$

We now fix  $l \in \{0, \dots, \frac{n}{2} - 1\}$  and observe that the pairs in the set  $S_l$  are mutually independent. Define a Bernoulli random variable,  $\beta_i$ , to be the indicator that  $(X_{m(i,l)}, X_{i+\frac{n}{2}})$  is misclassified. We have that  $\mathbf{E}[\beta_i] \geq 2\Phi_c(K)^2$ . Note that the fraction of pairs in the set  $S_l$  that are misclassified is  $\frac{1}{|S_l|} \sum_{i \in C_0 \cap N_{m(i,l)}} \beta_i$ , which is a sum of independent Bernoulli random variables. Hence, by Hoeffding's inequality, we obtain

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

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

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

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

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

Thus, out of all the pairs  $\mathbf{X}'_{ij}$  with  $j \approx i$ , with probability at least  $1 - O(n^{-c})$  for any  $c > 0$ , we have that at least a fraction  $2\Phi_c(K)^2$  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 [A.5](#) we have for any  $t \in (0, 1)$ ,

$$\Pr \left[ \sum_{i \in C_0 \cap N_{m(i,l)}} \beta_i \geq 2|S_l|\Phi_c(K)^2 - |S_l|t \right] \geq 1 - \exp(-|S_l|t^2/4).$$

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

$$\Pr \left[ \sum_{i \in C_0 \cap N_{m(i,l)}} \beta_i \geq nq\Phi_c(K)^2(1 \pm o_n(1)) - \sqrt{\kappa nq\Phi_c(K)^2 \log^2 n} \right] \geq 1 - O(n^{-\frac{\kappa}{4}\Phi_c(K)^2 \log n}).$$

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

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

Hence, in this regime of  $q$ ,

$$\Pr \left[ \sum_{i \in C_0 \cap N_{m(i,t)}} \beta_i > 0 \right] \geq 1 - O(n^{-\frac{\kappa}{4} \Phi_c(K)^2 \log n}),$$

and the proof is complete. ■

## A.5 Proof of Theorem 6

We first state the formal version of the theorem.

**Theorem 10** (Formal restatement of Theorem 6). *Assume that  $\|\boldsymbol{\mu}\|_2 \leq K\sigma$  and  $\sigma \leq K'$  for some constants  $K$  and  $K'$ . Moreover, assume that the parameters  $(\mathbf{w}, \mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}^2 \times \mathbb{R}$  are bounded by a constant. Then, with probability at least  $1 - o_n(1)$  over the data  $(\mathbf{X}, \mathbf{A}) \sim \text{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ , there exists a subset  $\mathcal{A} \subseteq [n]$  with cardinality at least  $n - o(\sqrt{n})$  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  $\mathbf{g}_i$  be independent Gaussian random variables with mean 0 and variance 1, so we have  $\mathbf{X}_i = -\boldsymbol{\mu} + \sigma \mathbf{g}_i$  for  $i \in C_0$  and  $\mathbf{X}_i = \boldsymbol{\mu} + \sigma \mathbf{g}_i$  for  $i \in C_1$ . Moreover, since the parameters  $(\mathbf{w}, \mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}^2 \times \mathbb{R}$  are bounded, we can write  $\mathbf{w} = R\hat{\mathbf{w}}$  and  $\mathbf{a} = R'\hat{\mathbf{a}}$  where  $\hat{\mathbf{w}}$  and  $\hat{\mathbf{a}}$  are unit vectors and  $R$  and  $R'$  are some constants. We define the following sets which will become useful later in our computation of  $\gamma_{ij}$ 's.

Define

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

For  $i \in [n]$  define

$$\begin{aligned} J_{i,0} &\stackrel{\text{def}}{=} \left\{ j \in N_i \cap C_0 \mid |\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j| \leq \sqrt{10} \right\}, \\ J_{i,1} &\stackrel{\text{def}}{=} \left\{ j \in N_i \cap C_1 \mid |\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j| \leq \sqrt{10} \right\}, \\ B_{i,0}^t &\stackrel{\text{def}}{=} \left\{ j \in N_i \cap C_0 \mid 2^{t-1} \leq \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j \leq 2^t \right\}, \quad t = 1, 2, \dots, T, \\ B_{i,1}^t &\stackrel{\text{def}}{=} \left\{ j \in N_i \cap C_1 \mid 2^{t-1} \leq \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j \leq 2^t \right\}, \quad t = 1, 2, \dots, T, \end{aligned}$$

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 A.13.** *With probability at least  $1 - o(1)$ , we have that  $|\mathcal{A}| \geq n - o(\sqrt{n})$ .*

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

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



We give a lower bound for  $|\mathcal{A}'|$ . Denote  $b \stackrel{\text{def}}{=} \Pr\left(|\hat{\mathbf{w}}^T \mathbf{g}_i| \geq 10\sqrt{\log(n(p+q))}\right)$ . Then

$$\frac{1}{2} \frac{1}{10\sqrt{\log(n(p+q))}} e^{-50\log(n(p+q))} \leq b \leq \frac{1}{10\sqrt{\log(n(p+q))}} e^{-50\log(n(p+q))}$$

Apply the multiplicative Chernoff bound

$$\Pr\left[\sum_{i \in [n]} \mathbf{1}_{\{|\hat{\mathbf{w}}^T \mathbf{g}_i| \geq 10\sqrt{\log(n(p+q))}\}} \geq nb(1+\delta)\right] \leq e^{-\frac{1}{3}nb\delta^2}$$

and set  $\delta = \frac{1}{\sqrt{n}}(10\sqrt{\log(n(p+q))})(n(p+q))^{25}$ , we see that with probability at least  $1 - o(1)$ ,

$$\left|\left\{i \in [n] \mid |\hat{\mathbf{w}}^T \mathbf{g}_i| \geq 10\sqrt{\log(n(p+q))}\right\}\right| \leq \frac{\sqrt{n}}{(n(p+q))^{25}}.$$

This means that

$$\begin{aligned} & \left|\left\{i \in [n] \mid |\hat{\mathbf{w}}^T \mathbf{g}_i| \geq 10\sqrt{\log(n(p+q))} \text{ or } \exists j \in N_i \text{ such that } |\hat{\mathbf{w}}^T \mathbf{g}_j| \geq 10\sqrt{\log(n(p+q))}\right\}\right| \\ & \leq \frac{\sqrt{n}}{(n(p+q))^{25}} \cdot \frac{n}{2}(p+q)(1 \pm o(1)) = \frac{\sqrt{n}}{2(n(p+q))^{24}}(1 \pm o(1)) = o(\sqrt{n}). \end{aligned}$$

Therefore we have

$$|\mathcal{A}'| \geq n - o(\sqrt{n}).$$

■

**Claim A.14.** *With probability at least  $1 - o(1)$ , we have that for all  $i \in [n]$ ,*

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

**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

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

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

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

Apply Chernoff's inequality we have

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

Apply the union bound we get

$$\Pr \left[ |J_{i,0}| \geq \frac{9}{10} |C_0 \cap N_i|, \forall i \in [n] \right] \geq 1 - \sum_{i \in [n]} e^{-\frac{4}{25} |N_i \cap C_0|} \geq (1 - o(1)) \left( 1 - ne^{-\frac{2nq(1 \pm o(1))}{25}} \right) = 1 - o(1).$$

The second last inequality follows because  $|N_i \cap C_0| \geq \frac{n}{2} \min\{p, q\}(1 \pm o(1)) = \frac{nq}{2}(1 \pm o(1))$  under degree concentration for all  $i \in [n]$ . Moreover, since we have used the degree concentration, this introduces the additional multiplicative  $(1 - o(1))$  term in the probability lower bound. The last equality is due to our assumption that  $q = \Omega(\frac{\log^2 n}{n})$ .  $\blacksquare$

**Claim A.15.** *With probability at least  $1 - o(1)$ , we have that for all  $i \in [n]$  and for all  $t \in [T]$ ,*

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

**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

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

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

$$\begin{aligned} & \Pr \left[ \bigcup_{i \in [n]} \bigcup_{t \in [T]} \left\{ |B_{i,0}^t| \geq \mathbf{E}[|B_{i,0}^t|] + \sqrt{T} |N_i \cap C_0|^{\frac{4}{5}} \right\} \right] \\ & \leq nT \exp \left( -2T \left( \frac{n}{2} \min\{p, q\}(1 \pm o(1)) \right)^{3/5} \right) + o(1) = o(1), \end{aligned}$$

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

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

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

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

**Proof of Theorem 10:** 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  $\mathbf{X}_i = (2\epsilon_i - 1)\boldsymbol{\mu} + \sigma \mathbf{g}_i$ ,

- $\mathcal{E}_0$  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  $|C_1 \cap N_i| = \frac{n}{2}((1 - \epsilon_i)q + \epsilon_i p)(1 \pm o(1))$ .
- $\mathcal{E}_1$  is the event that  $|\mathcal{A}| \geq n - o(\sqrt{n})$ .
- $\mathcal{E}_2$  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  $i \in [n]$ .
- $\mathcal{E}_3$  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 A.13, A.14, A.15, we get that with probability at least  $1 - o(1)$ , the event  $\mathcal{E}^* = \bigcap_{i=0}^3 \mathcal{E}_i$  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

$$\begin{aligned} \gamma_{ij} &= \frac{\exp(\text{LeakyRelu}(\mathbf{a}_1 \mathbf{w}^T \mathbf{X}_i + \mathbf{a}_2 \mathbf{w}^T \mathbf{X}_j + b))}{\sum_{k \in N_i} \exp(\text{LeakyRelu}(\mathbf{a}_1 \mathbf{w}^T \mathbf{X}_i + \mathbf{a}_2 \mathbf{w}^T \mathbf{X}_k + b))} \\ &= \frac{\exp(\sigma R R' \text{LeakyRelu}(\kappa_{ij} + \hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i + \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j + b'))}{\sum_{k \in N_i} \exp(\sigma R R' \text{LeakyRelu}(\kappa_{ik} + \hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i + \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k + b'))} \\ &= \frac{1}{\sum_{k \in N_i} \exp(\Delta_{ik} - \Delta_{ij})} \end{aligned}$$

where for  $l \in N_i$ , we denote

$$\begin{aligned} \kappa_{il} &\stackrel{\text{def}}{=} (2\epsilon_i - 1) \hat{\mathbf{w}}^T \boldsymbol{\mu} / \sigma + (2\epsilon_l - 1) \hat{\mathbf{w}}^T \boldsymbol{\mu} / \sigma, \\ \Delta_{il} &\stackrel{\text{def}}{=} \sigma R R' \text{LeakyRelu}(\kappa_{il} + \hat{\mathbf{a}}_1 \mathbf{w}^T \mathbf{g}_i + \hat{\mathbf{a}}_2 \mathbf{w}^T \mathbf{g}_l + b'), \end{aligned}$$

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  $\|\boldsymbol{\mu}\|_2 \leq K\sigma$  for some absolute constant  $K$ , we know that

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

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

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

$$\begin{aligned} \Delta_{ik} - \Delta_{ij} &= \sigma R R' \left( \text{LeakyRelu}(\kappa_{ik} + \hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i + \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k + b') \right. \\ &\quad \left. - \text{LeakyRelu}(\kappa_{ij} + \hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i + \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j + b') \right) \\ &= \sigma R R' \left( \text{LeakyRelu}(\hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i + \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k \pm O(1)) \right. \\ &\quad \left. - \beta(\kappa_{ij} + \hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i + \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j + b') \right) \\ &= \sigma R R' \left( \text{LeakyRelu}(\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k \pm O(1)) \pm O(1) \right) \\ &= \sigma R R' \left( \Theta(\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k) \pm O(1) \right), \end{aligned}$$

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

- We have  $j \in J_{i,0}$  and hence  $|\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j| = O(1)$ ;
- We have  $\kappa_{ij} + \hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i + \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j + b' < 0$ , so  $\hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i < |\kappa_{ij}| + |\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j| + |b'| = O(1)$ , moreover, because  $\hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i \geq 0$ , we get that  $|\hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i| = O(1)$ ;
- We have  $|\kappa_{ij} + \hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i + \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j + b'| \leq |\hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i| + |\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_j| + |\kappa_{ij} + b'| = O(1) + O(1) + O(1) = O(1)$ .

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

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

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  $\kappa_{ik} + \hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i + \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k + b'$ .

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

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

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

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

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

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

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

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

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

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

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

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

$$\Delta_{ik} - \Delta_{ij} = \begin{cases} \Theta(\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k) \pm O(1), & \text{if } k \in J_{i,0} \cup J_{i,1} \\ O(\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k) \pm O(1), & \text{otherwise.} \end{cases} \quad (15)$$

By following a similar argument, one can show that Equation equation 15 holds for any  $i$  such that  $\hat{\mathbf{a}}_1 \hat{\mathbf{w}}^T \mathbf{g}_i < 0$ .

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

$$\begin{aligned} \sum_{k \in N_i \cap C_0} \exp(\Delta_{ik} - \Delta_{ij}) &\geq \sum_{k \in J_{i,0}} \exp(\Delta_{ik} - \Delta_{ij}) = \sum_{k \in J_{i,0}} \exp(\Theta(\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k) \pm O(1)) \\ &\geq \sum_{k \in J_{i,0}} e^{c_1} = |J_{i,0}| e^{c_1} = \Omega(|N_i \cap C_0|), \end{aligned} \quad (16)$$

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

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

It is easy to see that

$$\sum_{k \in P_1} \exp(\Delta_{ik} - \Delta_{ij}) \leq \sum_{k \in P_1} \exp(O(\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k) \pm O(1)) \leq \sum_{k \in P_1} e^{c_2} = |P_1| e^{c_2} = O(|N_i \cap C_0|), \quad (17)$$

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

$$\begin{aligned} \sum_{k \in P_2} \exp(\Delta_{ik} - \Delta_{ij}) &= \sum_{t \in [T]} \sum_{k \in B_{i,0}^t} \exp(\Delta_{ik} - \Delta_{ij}) \\ &\leq \sum_{t \in [T]} \sum_{k \in B_{i,0}^t} \exp(O(\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k) \pm O(1)) \\ &\leq \sum_{t \in [T]} |B_{i,0}^t| e^{c_3 2^t}, \end{aligned} \quad (18)$$

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

$$\begin{aligned} m_t &\stackrel{\text{def}}{=} \mathbf{E}[|B_{i,0}^t|] = \sum_{k \in N_i \cap C_0} \Pr(2^{t-1} \leq \hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k \leq 2^t) \leq \sum_{k \in N_i \cap C_0} \Pr[\hat{\mathbf{a}}_2 \hat{\mathbf{w}}^T \mathbf{g}_k \geq 2^{t-1}] \\ &\leq \sum_{k \in N_i \cap C_0} \Pr[\hat{\mathbf{w}}^T \mathbf{g}_k \geq 2^{t-1}] \leq |N_i \cap C_0| e^{-2^{2t-3}}. \end{aligned}$$

It follows that

$$\begin{aligned}
\sum_{t \in [T]} |B_{i,0}^t| e^{c_3 2^t} &\leq \sum_{t \in [T]} \left( |N_i \cap C_0| e^{-2^{2t-3}} + \sqrt{T} |N_i \cap C_0|^{\frac{4}{5}} \right) e^{c_3 2^t} \\
&\leq |N_i \cap C_0| \sum_{t=1}^{\infty} e^{-2^{2t-3}} e^{c_3 2^t} + \sum_{t \in [T]} \sqrt{T} |N_i \cap C_0|^{\frac{4}{5}} e^{c_3 2^t} \\
&\leq c_4 |N_i \cap C_0| + o(|N_i|) \\
&\leq O(|N_i|),
\end{aligned} \tag{19}$$

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

$$\begin{aligned}
\log \left( T^{\frac{3}{2}} e^{c_3 2^T} \right) &= \frac{3}{2} \log \left[ \log_2 \left( 10 \sqrt{\log(n(p+q))} \right) \right] + c_3 2^{\lceil \log_2(10 \sqrt{\log(n(p+q))}) \rceil} \\
&\leq \frac{3}{2} \log \left[ \log_2 \left( 10 \sqrt{\log(n(p+q))} \right) \right] + 20c_3 \sqrt{\log(n(p+q))} \\
&\leq O \left( \frac{1}{c} \log(n(p+q)) \right),
\end{aligned}$$

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 equation 18 and equation 19 we get

$$\sum_{k \in P_2} \exp(\Delta_{ik} - \Delta_{ij}) \leq O(|N_i|), \tag{20}$$

and combining Equations equation 17 and equation 20 we get

$$\sum_{k \in N_i \cap C_0} \exp(\Delta_{ik} - \Delta_{ij}) = \sum_{k \in P_1} \exp(\Delta_{ik} - \Delta_{ij}) + \sum_{k \in P_2} \exp(\Delta_{ik} - \Delta_{ij}) \leq O(|N_i|). \tag{21}$$

Now, by Equations equation 16 and equation 21 we get

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

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

$$\Omega(|N_i \cap C_1|) \leq \sum_{k \in N_i \cap C_1} \exp(\Delta_{ik} - \Delta_{ij}) \leq O(|N_i|). \tag{23}$$

Finally, Equations equation 22 and equation 23 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 A.13 and A.14 about the cardinalities of  $\mathcal{A}$ ,  $J_{i,0}$  and  $J_{i,1}$  we have thus proved Theorem 10.  $\blacksquare$

## B Additional experimental results

### B.1 Ansatz for GAT, MLP-GAT and GCN

For the original GAT architecture we fixed  $\mathbf{w} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$  and defined the first head as  $\mathbf{a}_1 = \frac{1}{\sqrt{2}}(1, 1)$  and  $b_1 = -\frac{1}{\sqrt{2}}\mathbf{w}^T\boldsymbol{\mu}$ ; The second head is defined as  $\mathbf{a}_2 = -\mathbf{a}_1$  and  $b_2 = -b_1$ . We briefly discuss the choice of such ansatz. The parameter  $\mathbf{w}$  is picked based on the optimal Bayes classifier without a graph, and the attention is set such that the first head maintains pairs in  $C_1$  and the second head maintains pairs in  $C_0$ <sup>8</sup>. We will clearly see from the results, this choice of ansatz produces good node classification performance (in the easy regime, where we vary  $q$  we clearly see how those performances degrade since GAT linear attention mechanism is unable to separate inter- from intra-edges). More specifically, one may use the same techniques in the proof of Theorem 1 and Corollaries 2 and 3 to prove the node separability results (in this particular case, the result will *depend on*  $q$  in contrast to the result we get for MLP-GAT, where the no dependence of  $q$  was needed).

For MLP-GAT we use the following choice of ansatz.

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

so that the output of the attention model is defined as

$$\mathbf{r} \cdot \text{LeakyRelu} \left( \mathbf{S} \cdot \begin{bmatrix} \tilde{\mathbf{w}}^T \mathbf{X}_i \\ \tilde{\mathbf{w}}^T \mathbf{X}_j \end{bmatrix} \right).$$

This choice of two layer network allows us to bypass the ‘‘XOR problem’’ [MP69] and separate inter- from intra-edges, which is clearly impossible with linear architecture.

For GCN we used the ansatz from [BFJ21], which is also  $\mathbf{w} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$ .

### B.2 Synthetic data

#### B.2.1 Fixing the distance between the means and varying $q$

In this subsection, we present additional experimental results for the original GAT [VCC+18] mechanism with two heads, and MLP-GAT on synthetic data. Unless stated otherwise, we use the exact parameter setting as in Subsection 4.1.1.

As explained in the introduction, a GAT head (equipped with a linear attention mechanism) is bound to fail to separate the edges. Recall that when considering the pair space  $(\mathbf{w}^T \mathbf{X}_i, \mathbf{w}^T \mathbf{X}_j)$ , we can think of each pair as a two-dimensional Gaussian with means in either one of the four quadrants. For correct edge classification, we need to classify the data originating from the distributions whose means are in the second and fourth quadrant from the data originating from the distributions whose means are in the first and the third quadrants (that is, the problem is a ‘‘XOR problem’’ [MP69]). However, it is readily seen that any linear classifier fails on the above task. In Figure 5a we can clearly see the lack of ability of the heads to ignore inter-edges. For example, head 0 maintains all the intra-edges of class 1 but also maintains the inter-edges between class 1 to class 0. Figure 5b presents the attention coefficients of GAT in the hard regime

<sup>8</sup>Note that in this case, due to the linearity of the attention mechanism, it will be impossible for the model to keep only  $\gamma_{ij}$  which correspond to intra-class edges.

and demonstrates Theorem 6, where most  $\gamma$  concentrate around uniform (GCN) coefficients. In Figure 5c we observe the node classification performance of the GAT model in the easy regime. As opposed to MLP-GAT, we can clearly see that the node classification performance of GAT is affected by increasing  $q$ . Figure 5d demonstrates the node classification performance in the hard regime (which is conjectured in Conjecture 7).

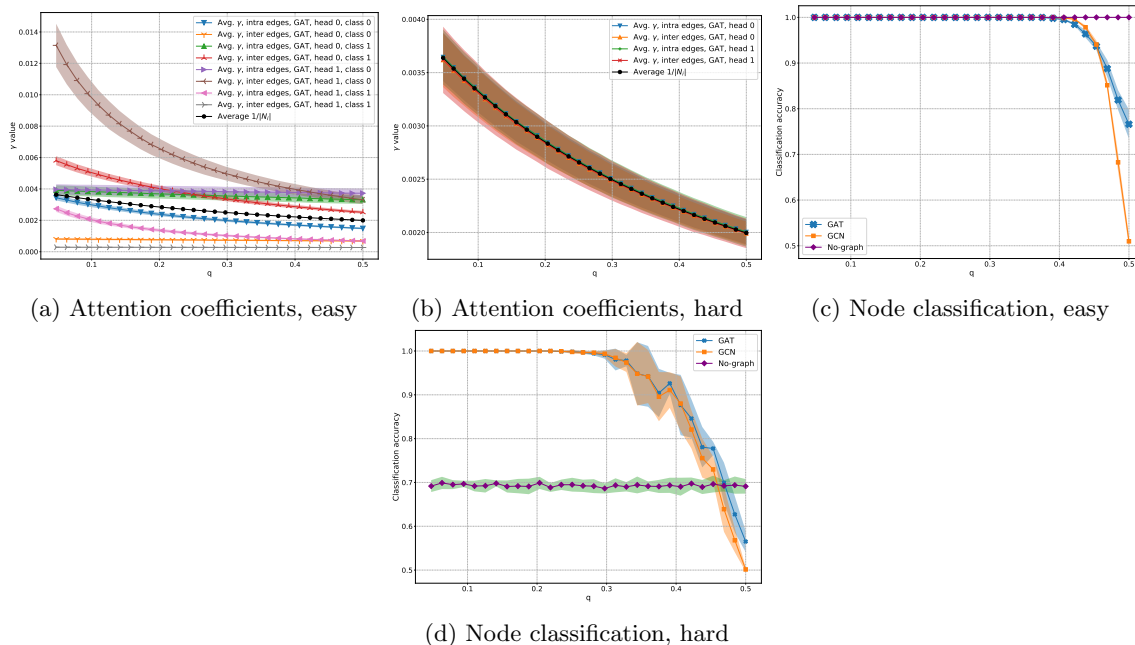


Figure 5: Attention coefficients and node classification accuracy for GAT.

### B.2.2 Fixing $q$ and varying the distance between the means

We consider the case where  $q = 0.4$ . In Figure 6 we show the attention coefficients for both MLP-GAT and two head GAT as a function of the distance between the means, and the node classification performance of GAT as a function of the distance between the means. In Figure 6a we see that in the hard regime MLP-GAT produce attention coefficients that concentrate around uniform (GCN) coefficients, while in the easy regime the model is able to maintain only the  $\gamma$  that correspond to intra-class edges (as stated in Corollary 2) while ignoring all coefficients corresponding to inter-class edges. In Figure 6b we observe that in the hard regime the attention coefficients of GAT concentrate around the uniform coefficients (as proved in Theorem 6), while in the easy regime, even though the attention coefficients concentrate, GAT is not able to distinguish intra from inter edges. As explained before, this is due to the fact that a linear attention mechanism is bound to fail on the “XOR problem”. Lastly, in Figure 6c we show classification results for GAT. Note that in the easy regime GAT achieves perfect separability. However, as the distance between the means decreases, GAT begins to misclassify nodes.



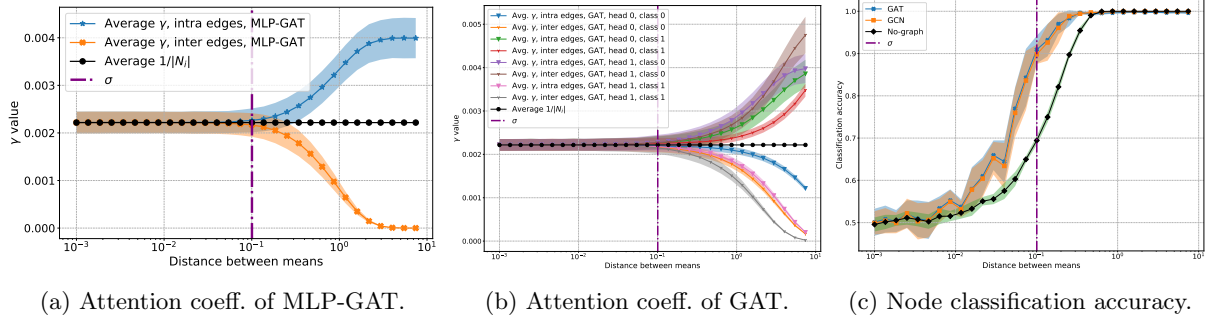


Figure 6: Attention coefficients MLP-GAT/GAT and node classification for GAT.

## B.3 Real-world data

### B.3.1 Ansatz

In Figure 7 we present the results for MLP-GAT on CiteSeer using the chosen ansatz.

In Figure 8 we present the results for MLP-GAT on Cora using the chosen ansatz.

In Figure 9 we present the results for MLP-GAT on PubMed using the chosen ansatz.

In Figure 10 we present the results for GAT on CiteSeer using the chosen ansatz.

In Figure 11 we present the results for GAT on Cora using the chosen ansatz.

In Figure 12 we present the results for GAT on PubMed using the chosen ansatz.

### B.3.2 Training

For experiments on real data, we also used PyTorch Geometric [FL19] to train the models GAT, MLP-GAT, GCN, and the linear classifier. We train the models using the Adam optimizer with a learning rate of  $10^{-3}$ , weight decay  $10^{-3}$ , and 200 epochs. We terminate the process if the average binary cross-entropy loss is less than  $10^{-2}$ . We report the results of the last epoch.

It is important to note that for MLP-GAT the results that we get from the training process are very similar to the ones reported in the main paper. One difference that we observed is that for a very large distance between the means the trained parameters of MLP-GAT resulted in some misclassifications for the edges. This happens because in this easy regime the graph is not needed at all and there must exist many parameter settings that achieve a near-perfect node classification without necessarily distinguishing intra-class from inter-class edges.

For GAT we also observe similar trends for training to the ones that we observed using an ansatz. One difference is the behavior of  $\gamma$ , which exhibit some irregular behavior when the distance between the means is very large. Again, this does not seem to affect node classification accuracy. In the case where the distance between the means is small, we do observe that the average  $\gamma$  is close to uniform, which is also what we have proved in the main paper for the CSBM model.

In Figure 13 we present the results for MLP-GAT on CiteSeer under training.

In Figure 14 we present the results for MLP-GAT on Cora under training.

In Figure 15 we present the results for MLP-GAT on PubMed under training.

In Figure 16 we present the results for GAT on CiteSeer under training.

In Figure 17 we present the results for GAT on Cora under training.

In Figure 18 we present the results for GAT on PubMed under training.

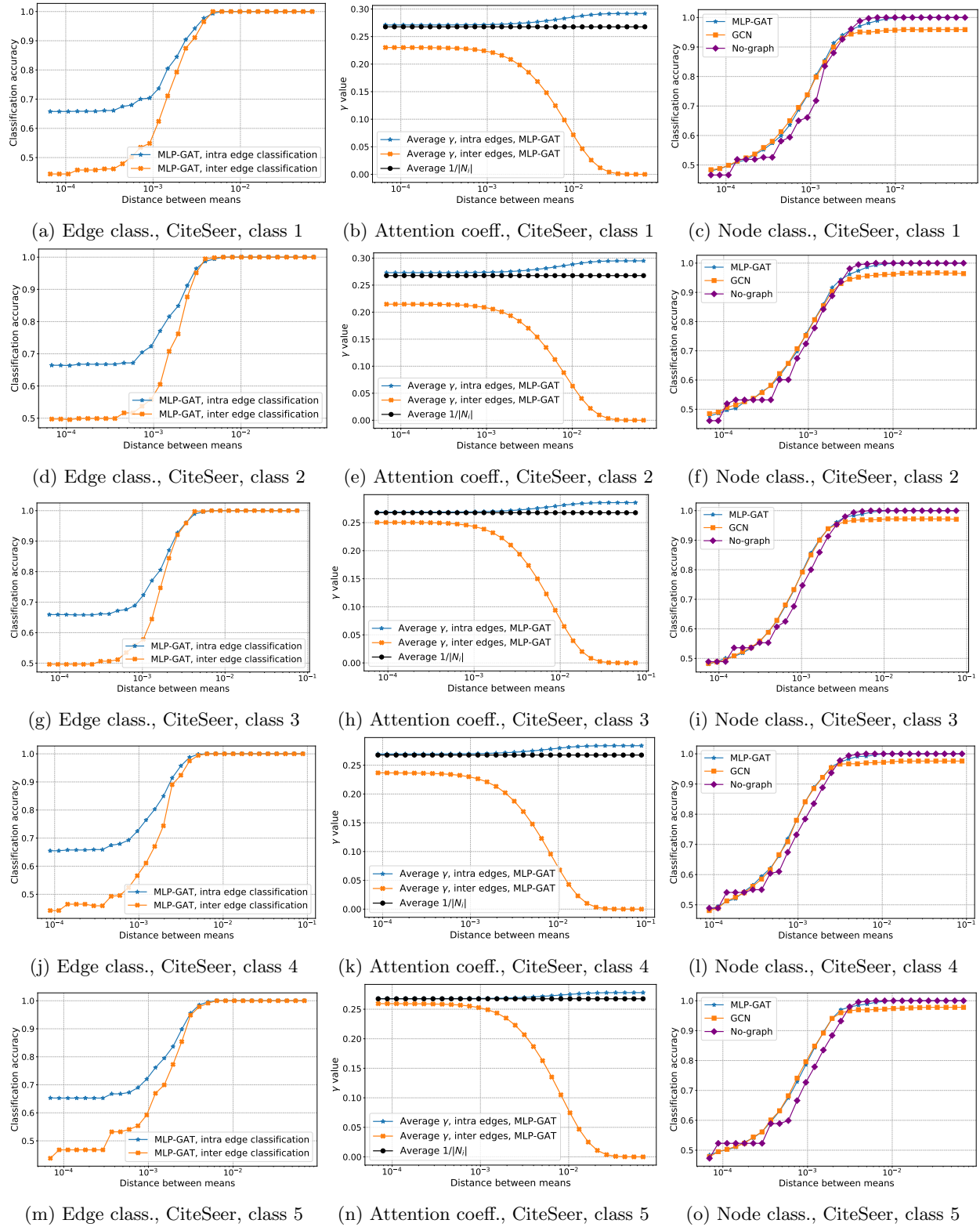
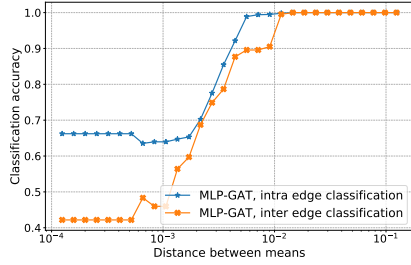
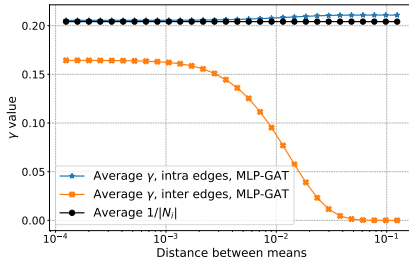


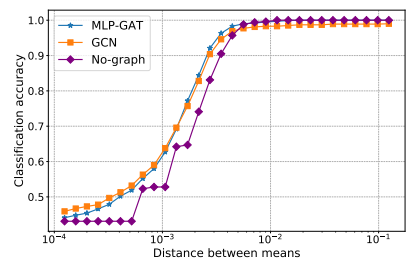
Figure 7: Ansatz MLP-GAT on CiteSeer.



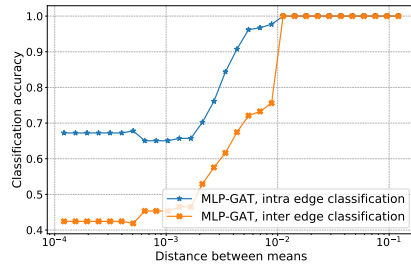
(a) Edge class., Cora, class 1



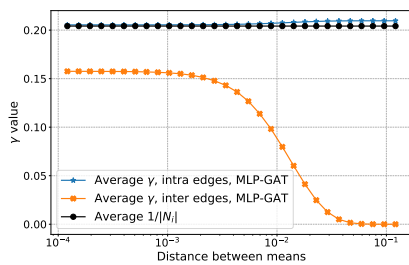
(b) Attention coeff., Cora, class 1



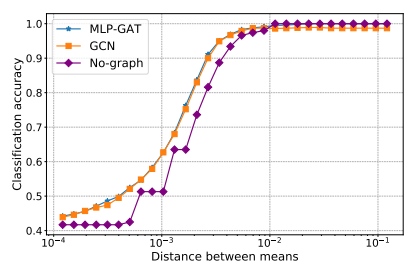
(c) Node class., Cora, class 1



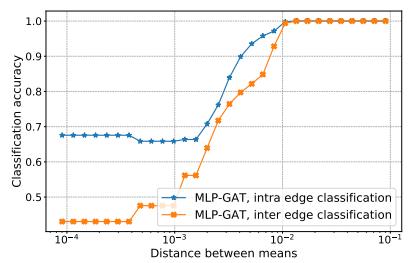
(d) Edge class., Cora, class 2



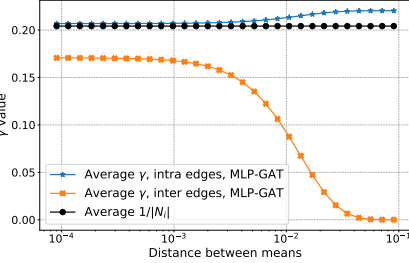
(e) Attention coeff., Cora, class 2



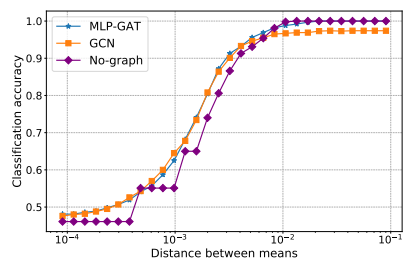
(f) Node class., Cora, class 2



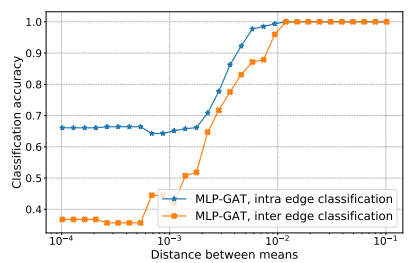
(g) Edge class., Cora, class 3



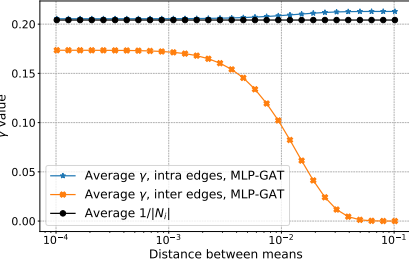
(h) Attention coeff., Cora, class 3



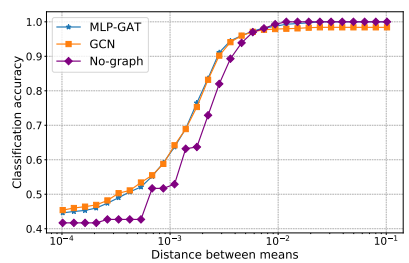
(i) Node class., Cora, class 3



(j) Edge class., Cora, class 4



(k) Attention coeff., Cora, class 4



(l) Node class., Cora, class 4

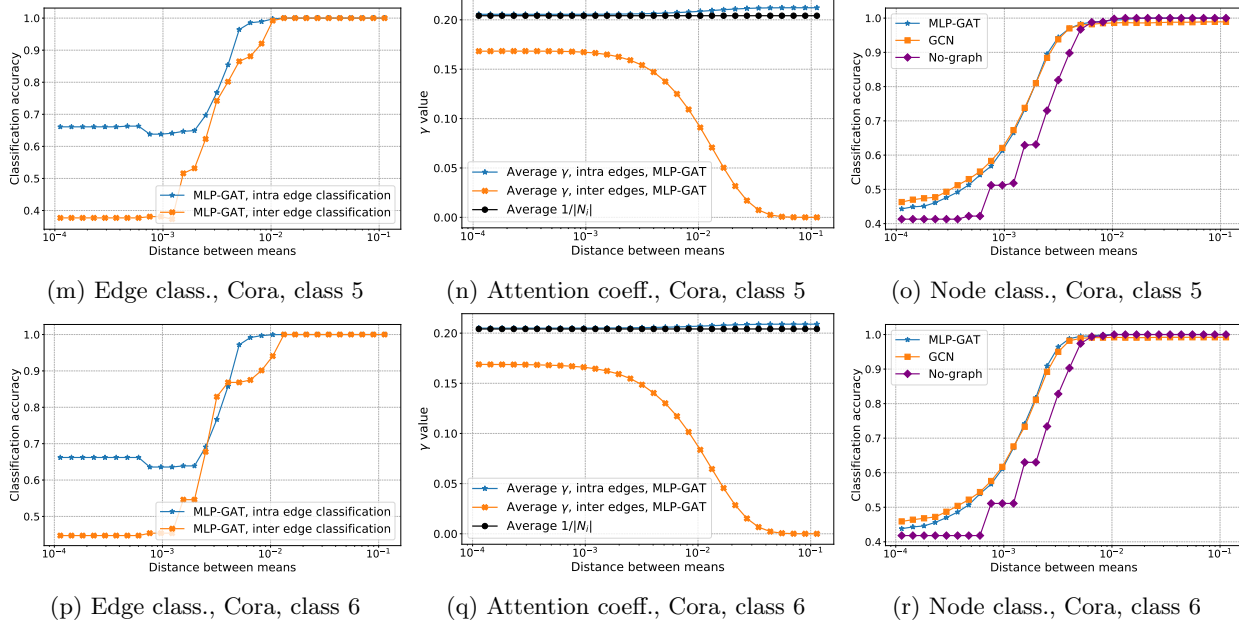


Figure 8: Ansatz MLP-GAT on Cora.

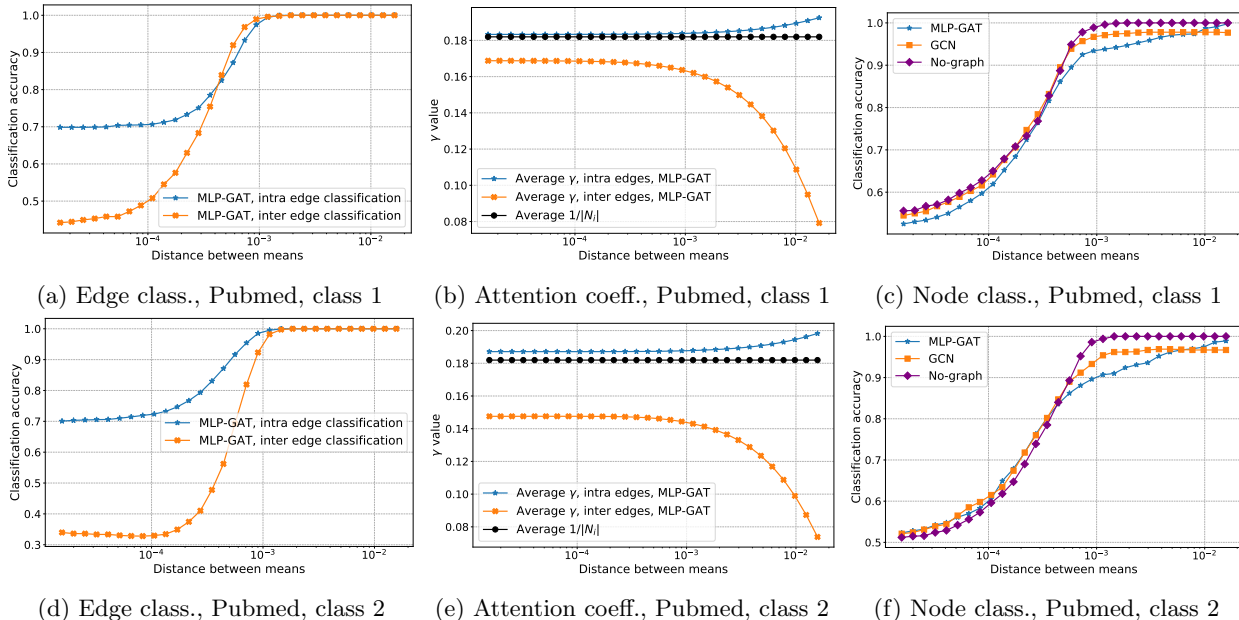


Figure 9: Ansatz MLP-GAT on PubMed.

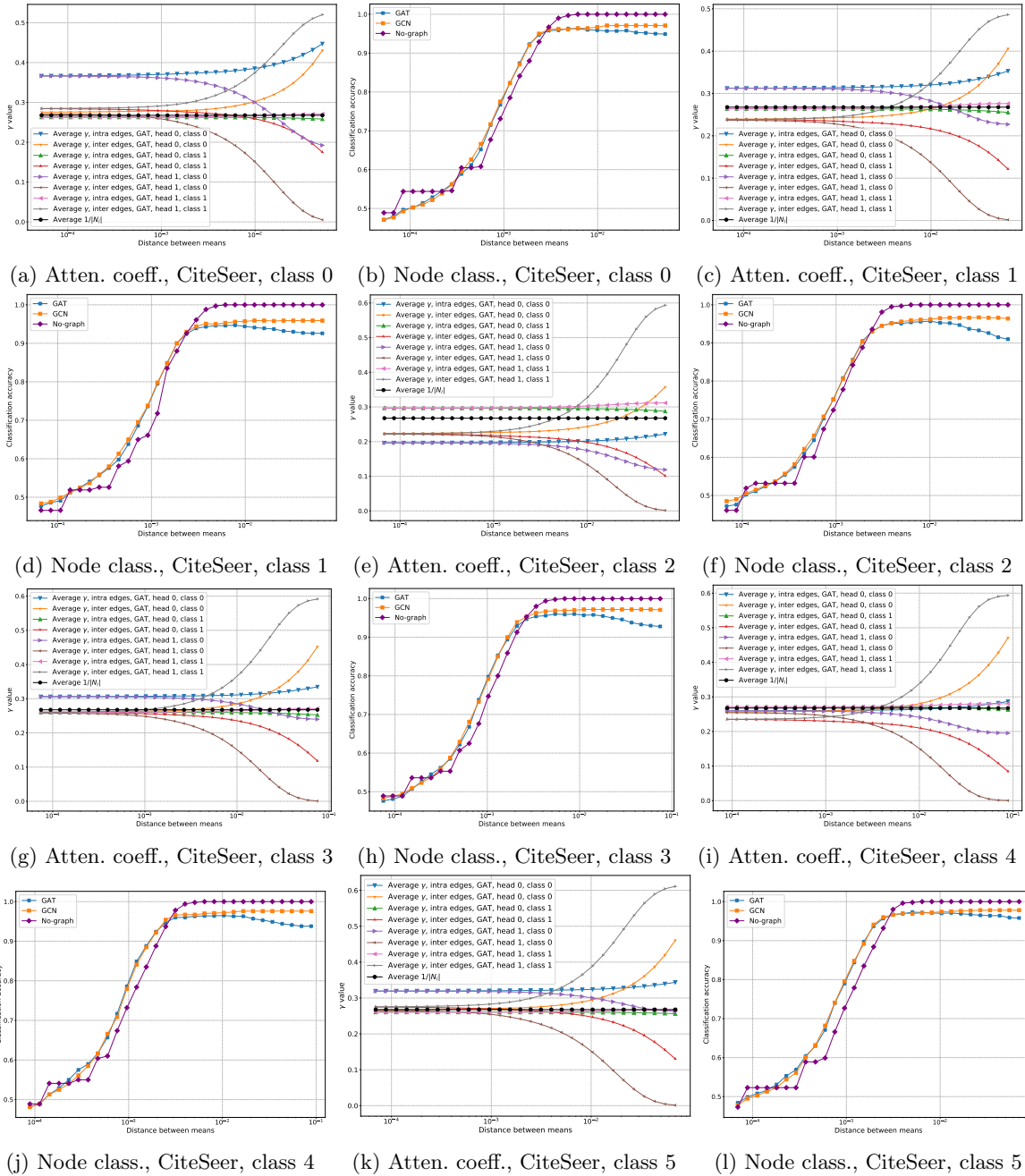
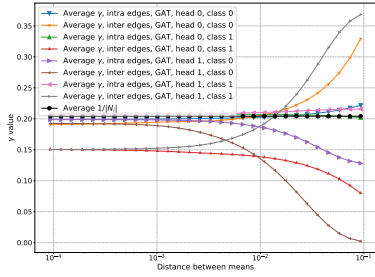
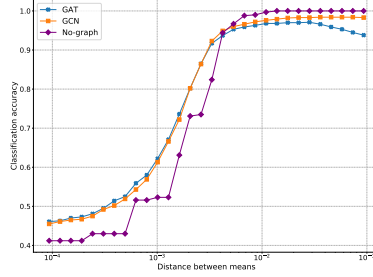


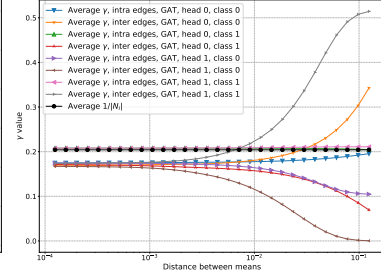
Figure 10: Ansatz GAT on CiteSeer.



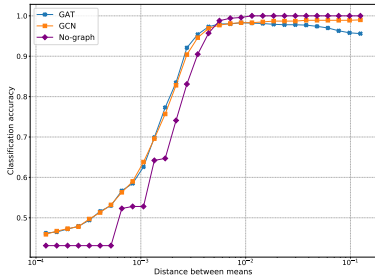
(a) Atten. coeff., Cora, class 0



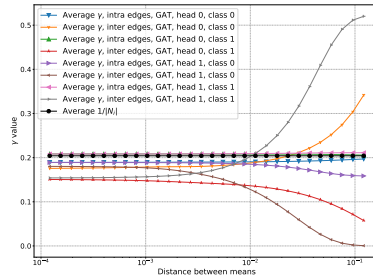
(b) Node class., Cora, class 0



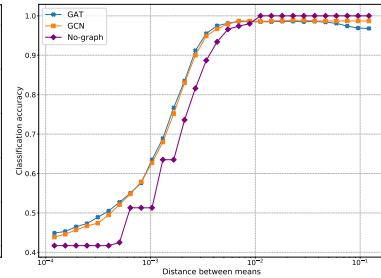
(c) Atten. coeff., Cora, class 1



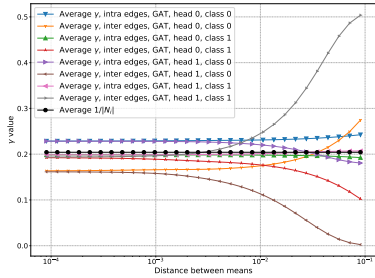
(d) Node class., Cora, class 1



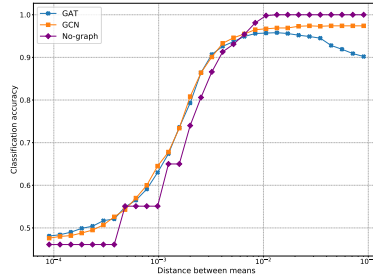
(e) Atten. coeff., Cora, class 2



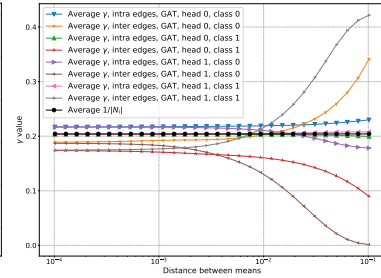
(f) Node class., Cora, class 2



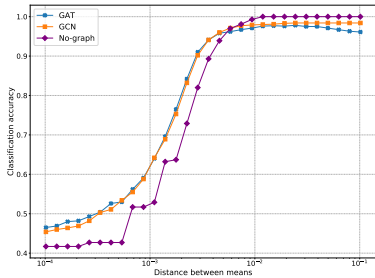
(g) Atten. coeff., Cora, class 3



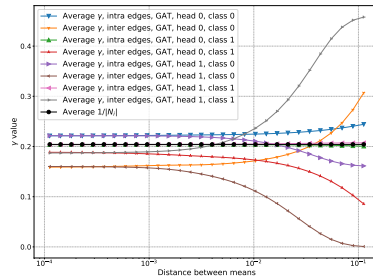
(h) Node class., Cora, class 3



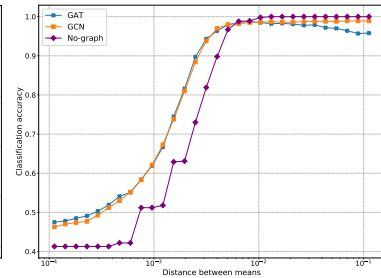
(i) Atten. coeff., Cora, class 4



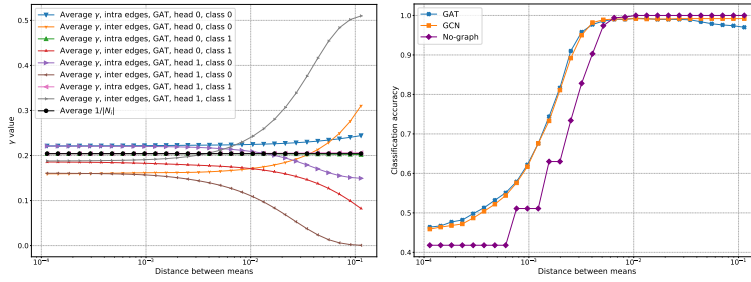
(j) Node class., Cora, class 4



(k) Atten. coeff., Cora, class 5

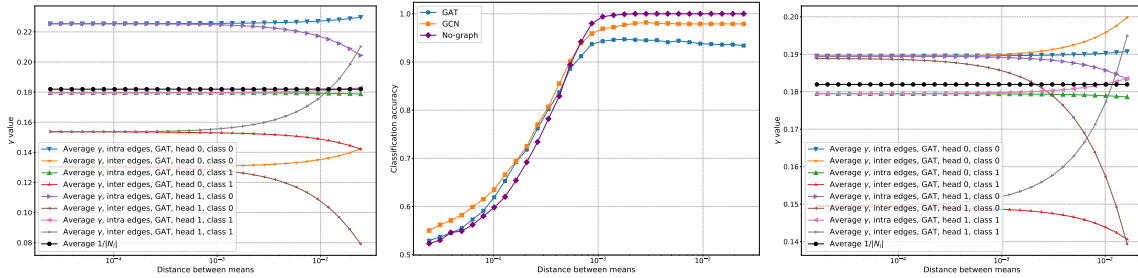


(l) Node class., Cora, class 5

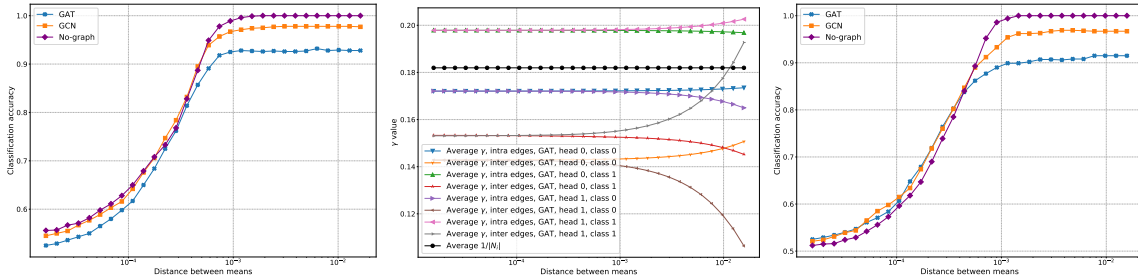


(m) Atten. coeff., Cora, class 6      (n) Node class., Cora, class 6

Figure 11: Ansatz GAT on Cora.



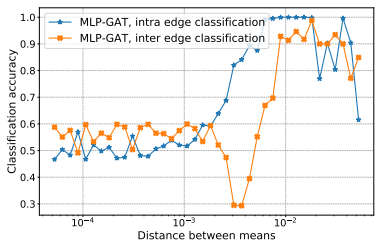
(a) Atten. coeff., Pubmed, class 0      (b) Node class., Pubmed, class 0      (c) Atten. coeff., Pubmed, class 1



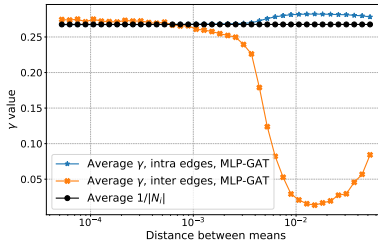
(d) Node class., Pubmed, class 1      (e) Atten. coeff., Pubmed, class 2      (f) Node class., Pubmed, class 2

Figure 12: Ansatz GAT on PubMed.

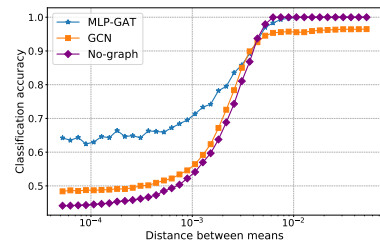




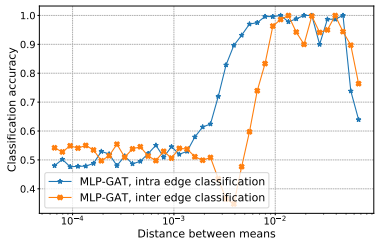
(a) Edge class., CiteSeer, class 0



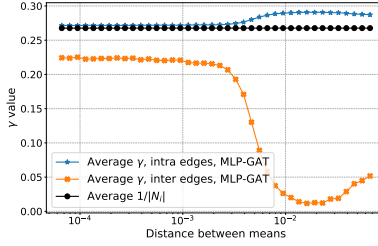
(b) Atten. coeff., CiteSeer, class 0



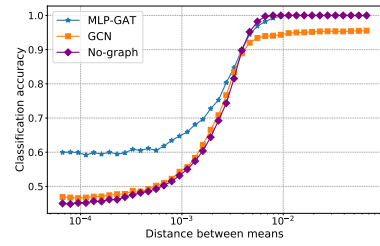
(c) Node class., CiteSeer, class 0



(d) Edge class., CiteSeer, class 1



(e) Atten. coeff., CiteSeer, class 1



(f) Node class., CiteSeer, class 1

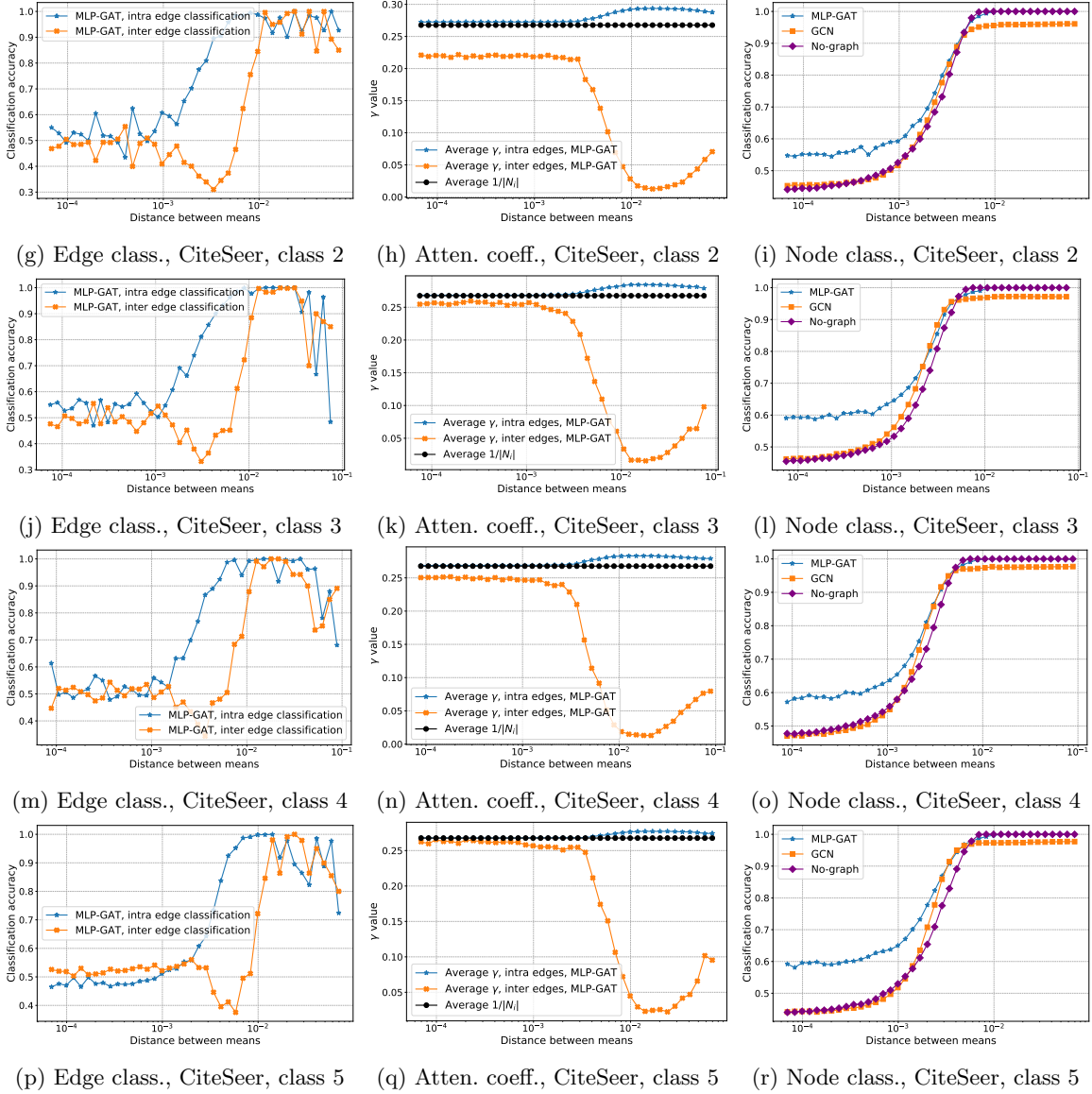
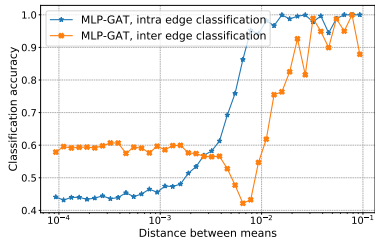
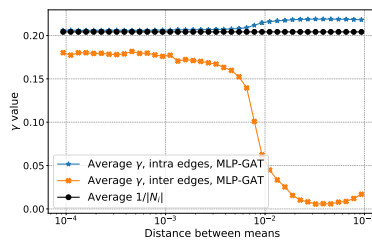


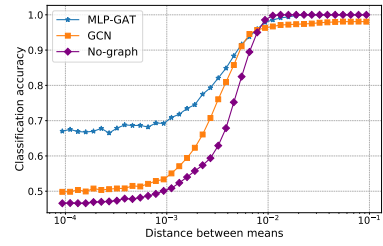
Figure 13: Training MLP-GAT on CiteSeer.



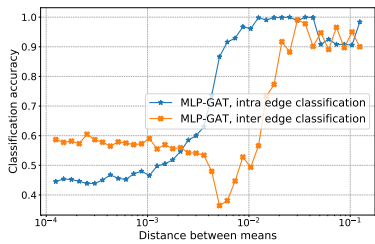
(a) Edge class., Cora, class 0



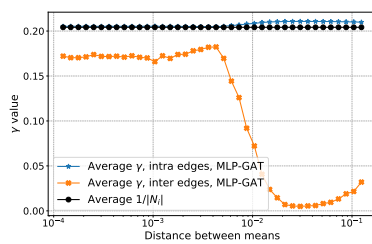
(b) Atten. coeff., Cora, class 0



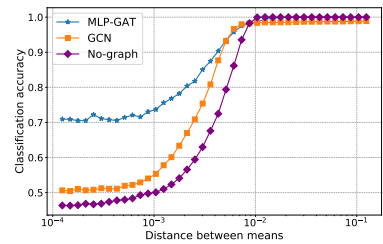
(c) Node class., Cora, class 0



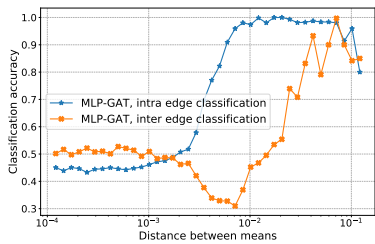
(d) Edge class., Cora, class 1



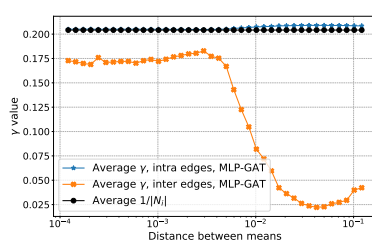
(e) Atten. coeff., Cora, class 1



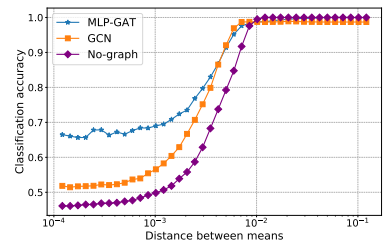
(f) Node class., Cora, class 1



(g) Edge class., Cora, class 2



(h) Atten. coeff., Cora, class 2



(i) Node class., Cora, class 2

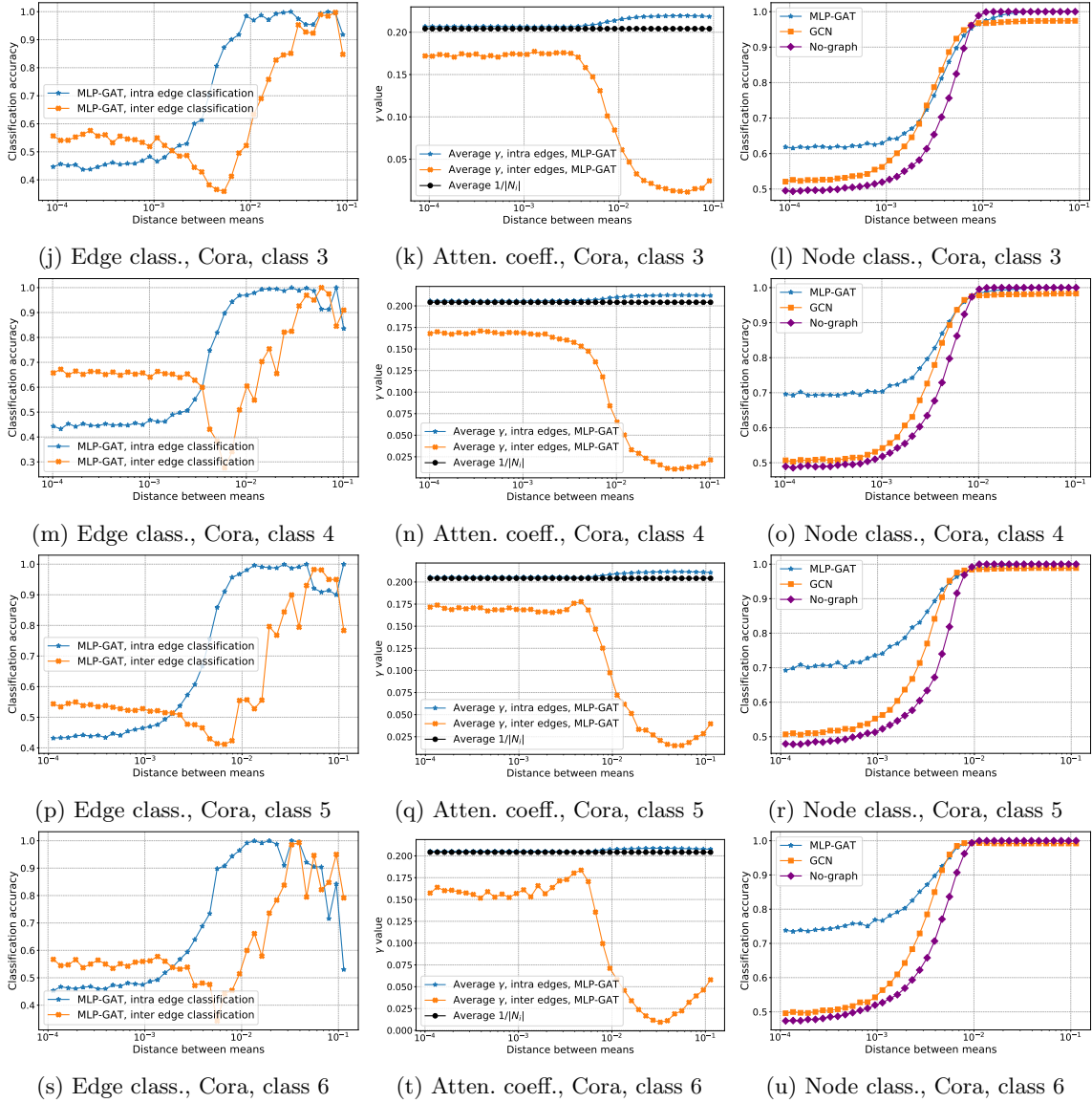


Figure 14: Training MLP-GAT on Cora.

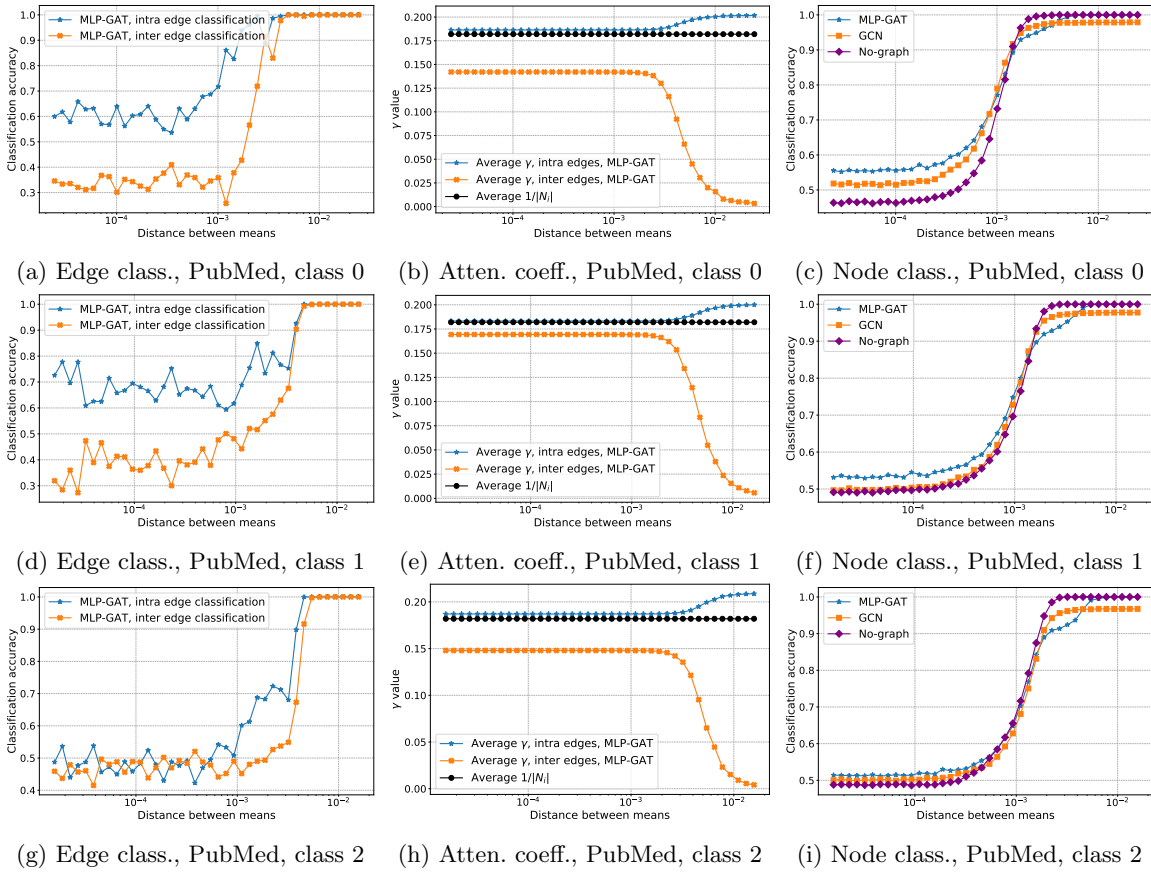


Figure 15: Training MLP-GAT on PubMed.

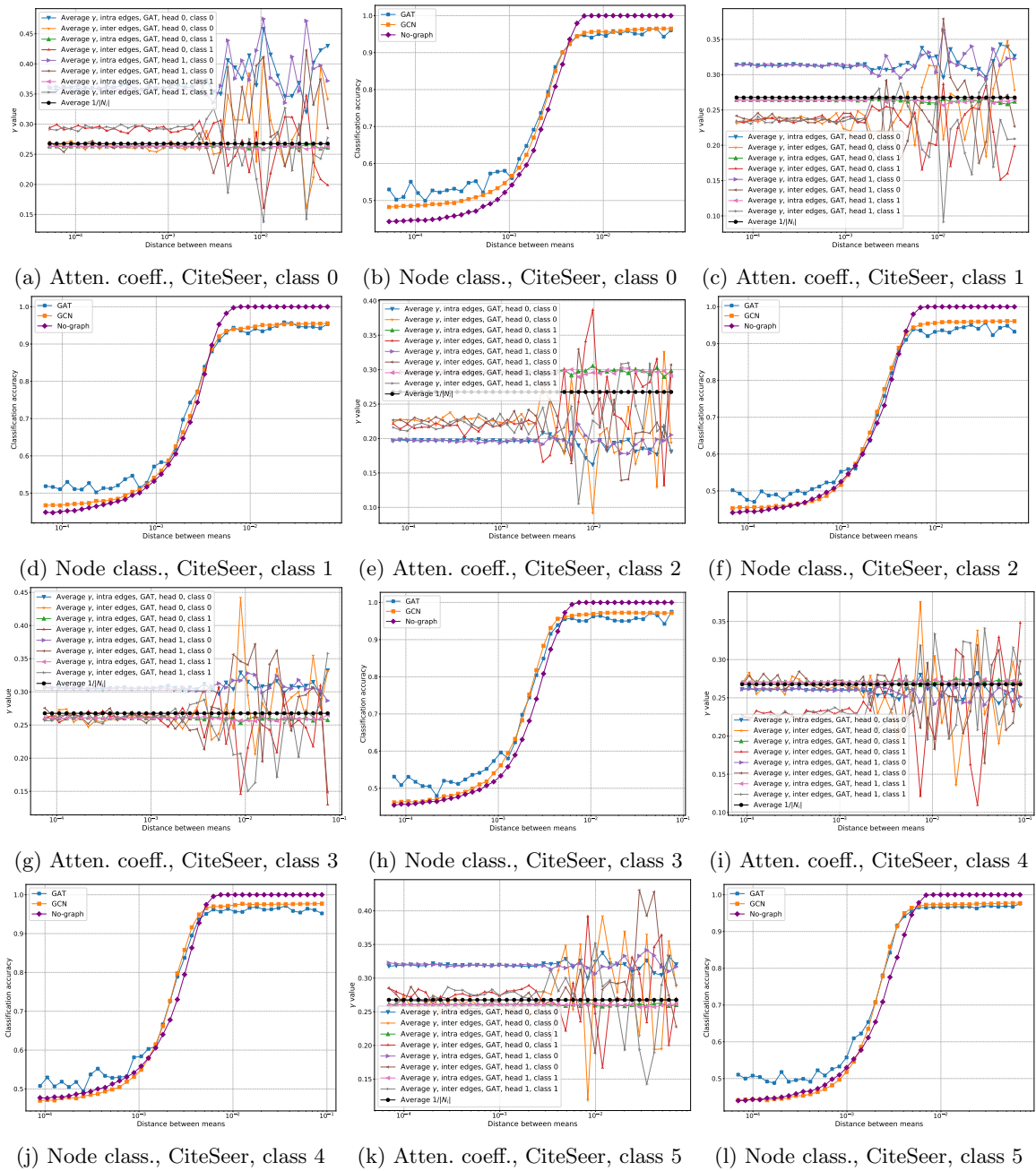
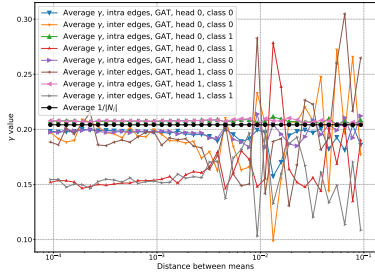
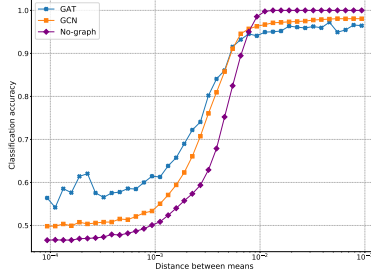


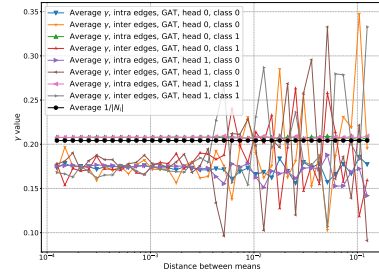
Figure 16: Training GAT on CiteSeer.



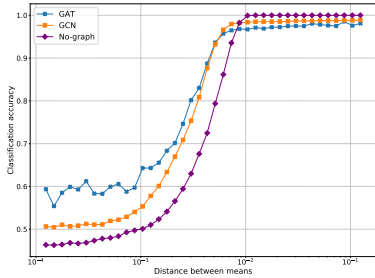
(a) Atten. coeff., Cora, class 0



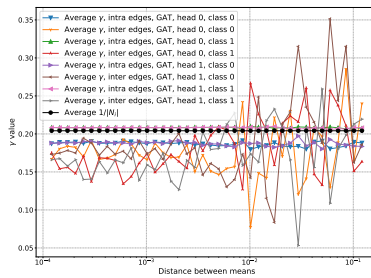
(b) Node class., Cora, class 0



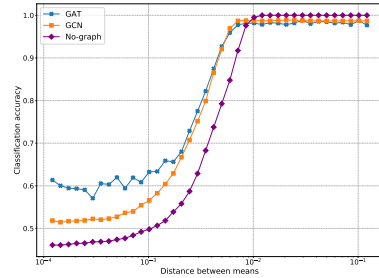
(c) Atten. coeff., Cora, class 1



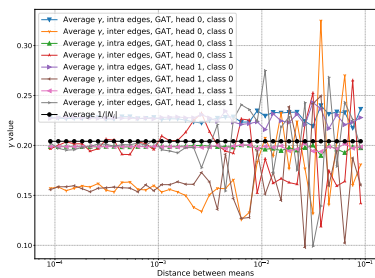
(d) Node class., Cora, class 1



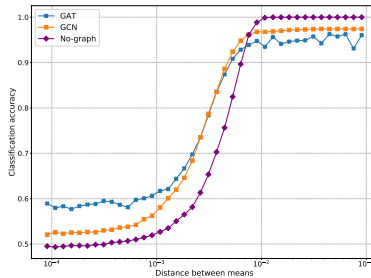
(e) Atten. coeff., Cora, class 2



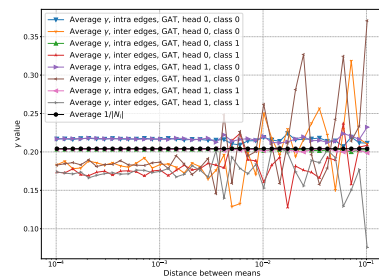
(f) Node class., Cora, class 2



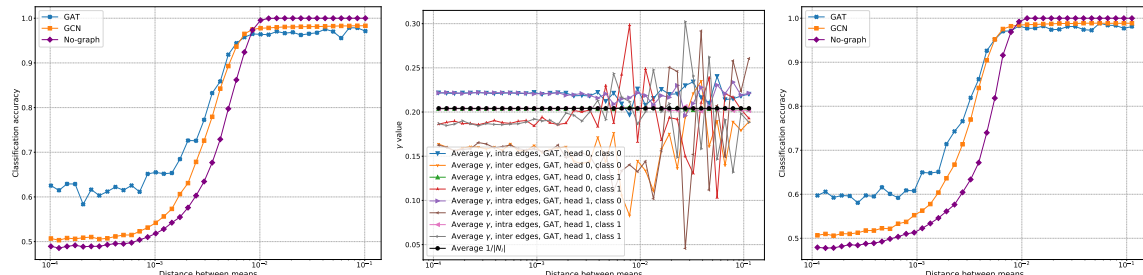
(g) Atten. coeff., Cora, class 3



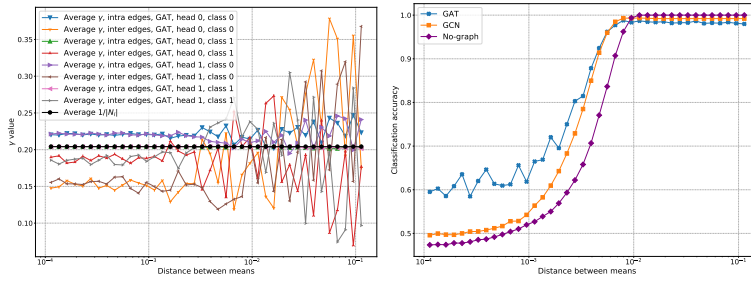
(h) Node class., Cora, class 3



(i) Atten. coeff., Cora, class 4

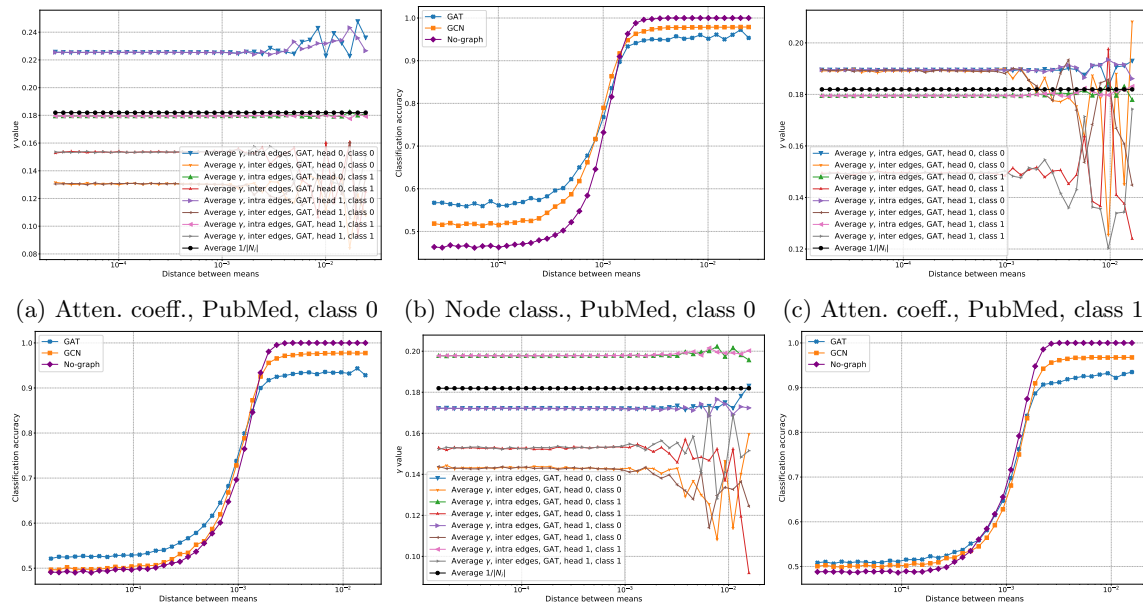


(j) Node class., Cora, class 4      (k) Atten. coeff., Cora, class 5      (l) Node class., Cora, class 5



(m) Atten. coeff., Cora, class 6      (n) Node class., Cora, class 6

Figure 17: Training GAT on Cora.



(a) Atten. coeff., PubMed, class 0      (b) Node class., PubMed, class 0      (c) Atten. coeff., PubMed, class 1  
 (d) Node class., PubMed, class 1      (e) Atten. coeff., PubMed, class 2      (f) Node class., PubMed, class 2

Figure 18: Training GAT on PubMed.