Graph Attention Retrospective

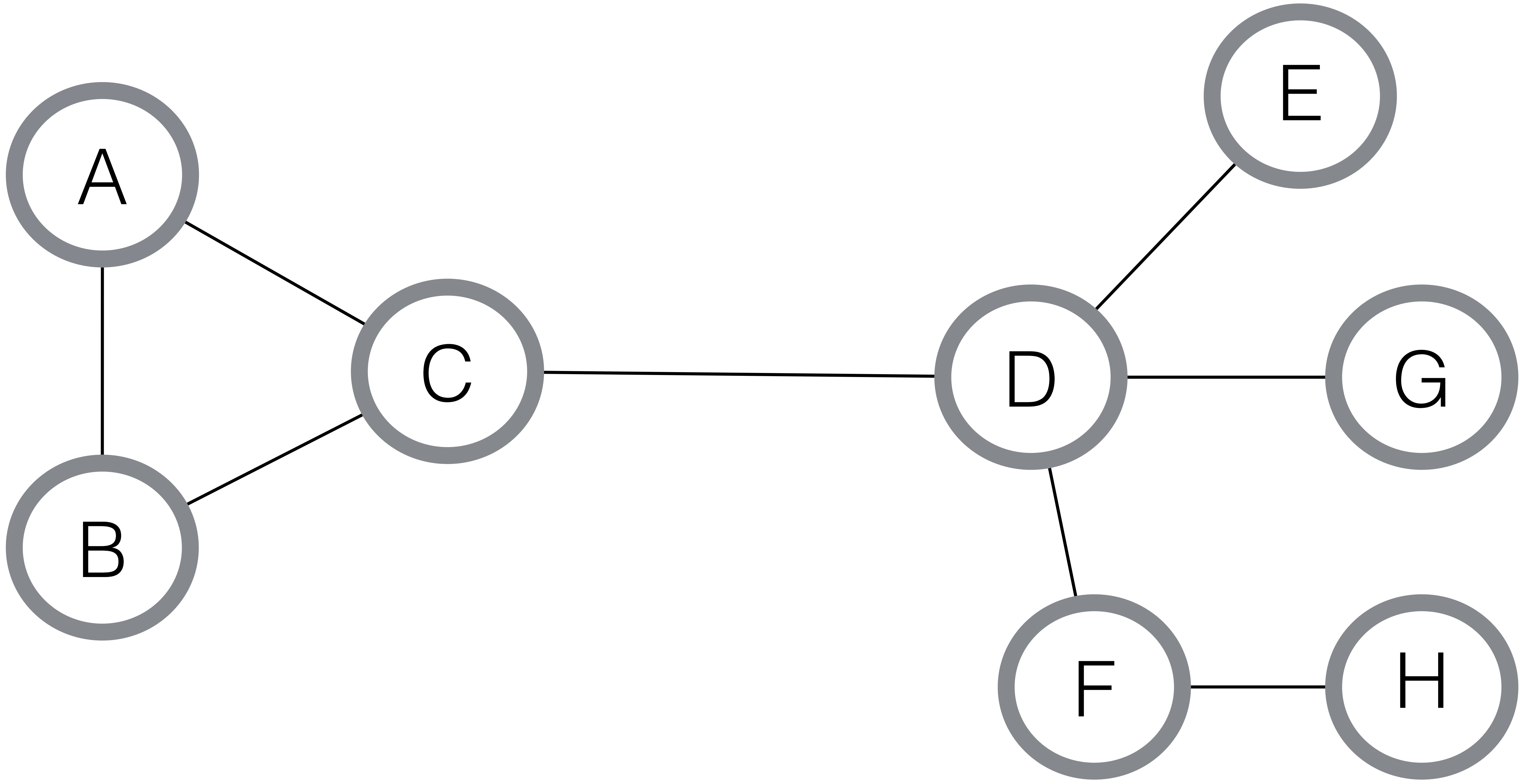
Kimon Fountoulakis, Amit Levi, Shenghao Yang, Aseem Baranwal,
Aukosh Jagannath

LoGaG Reading Group
19/04/2022

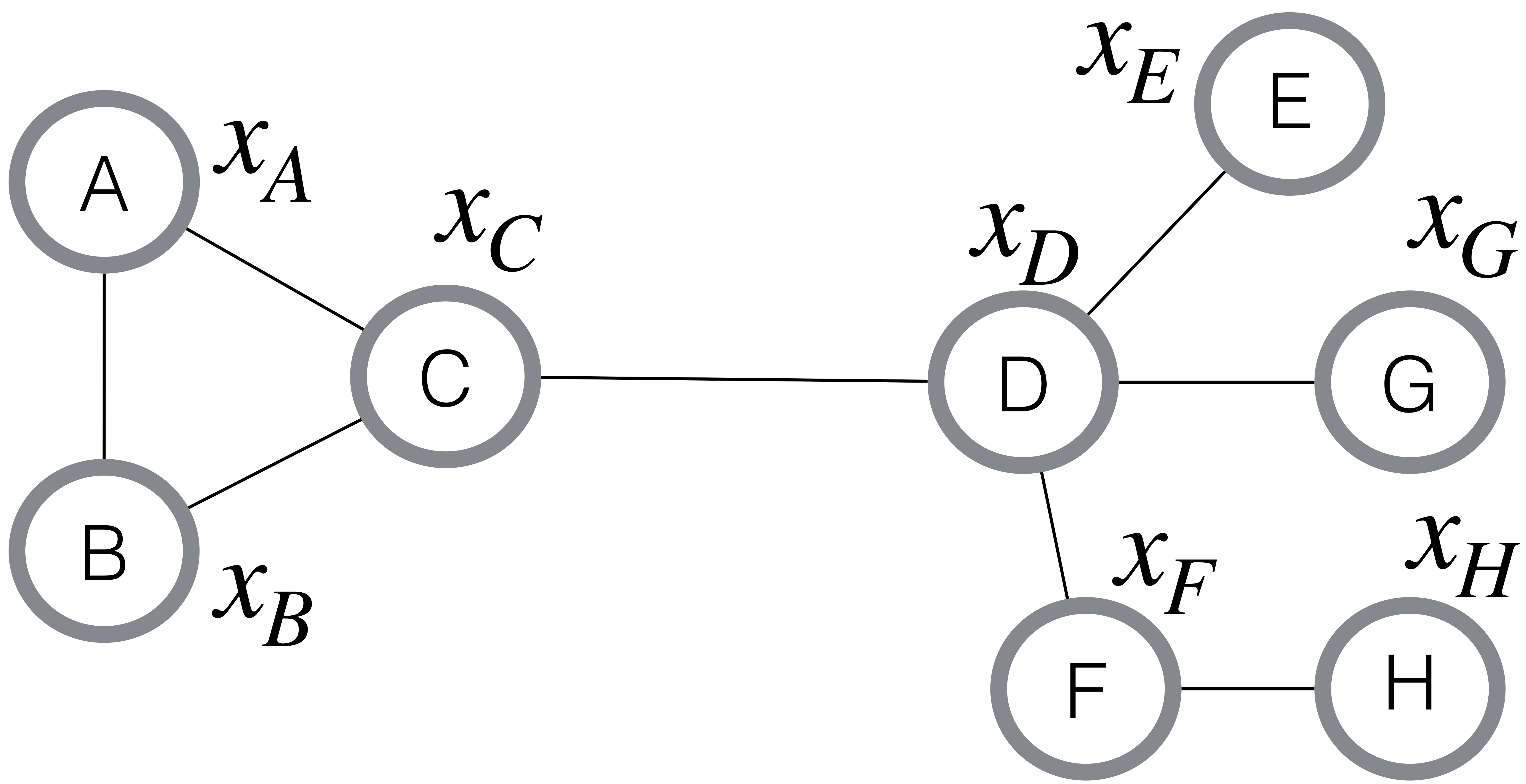
Outline

- First part: informal discussion, intuition and results (10-15 min + questions)
- Second part: details (20-25 min + questions)

Graphs

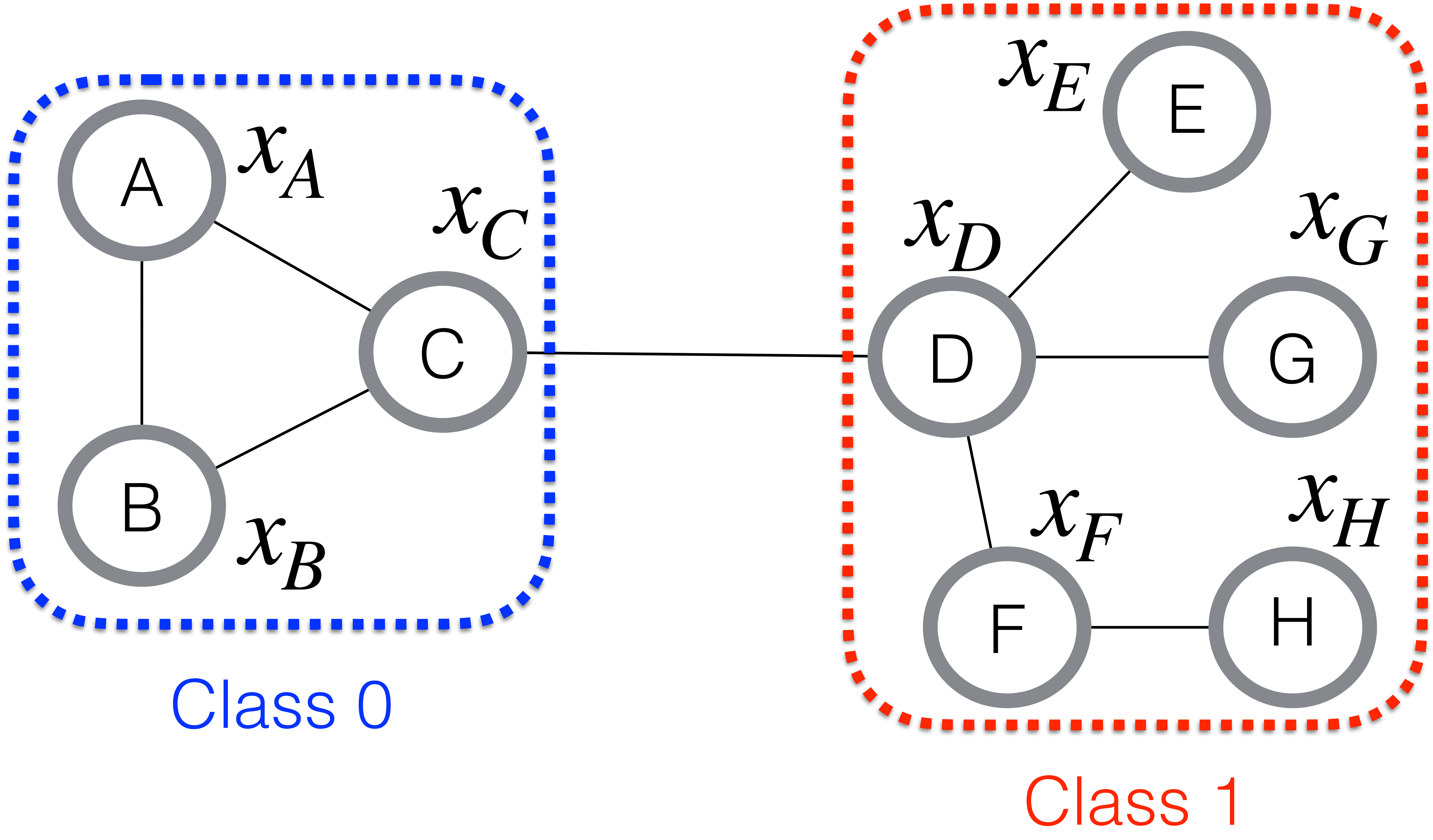


Graphs + features



- x_i is the feature vector for node i

Node classification



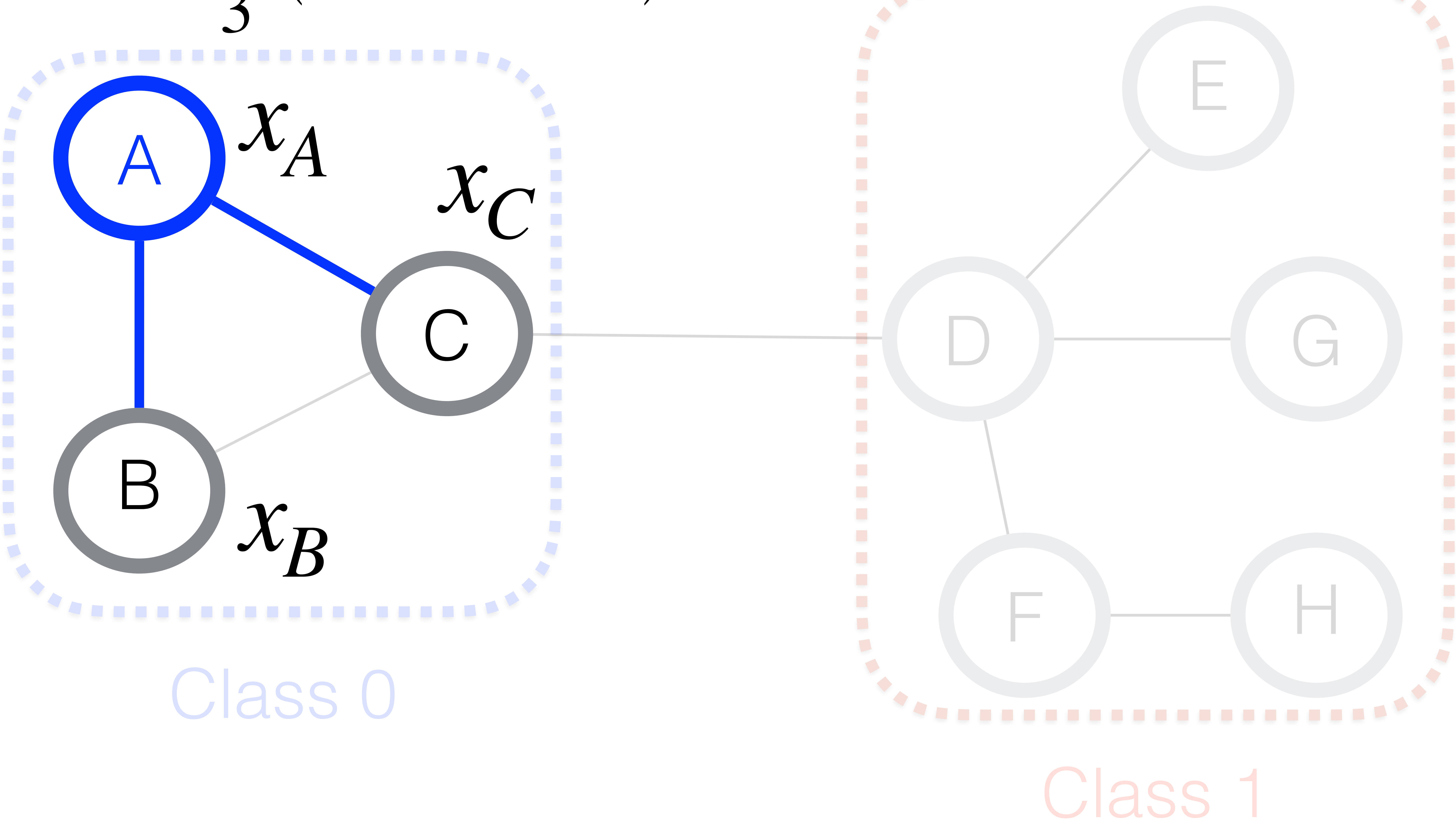
- x_i is the feature vector for node i

Terminology

- Same-class edge = intra-class edge
- Different-class edge = inter-class edge

Graph Convolution Network (GCN)

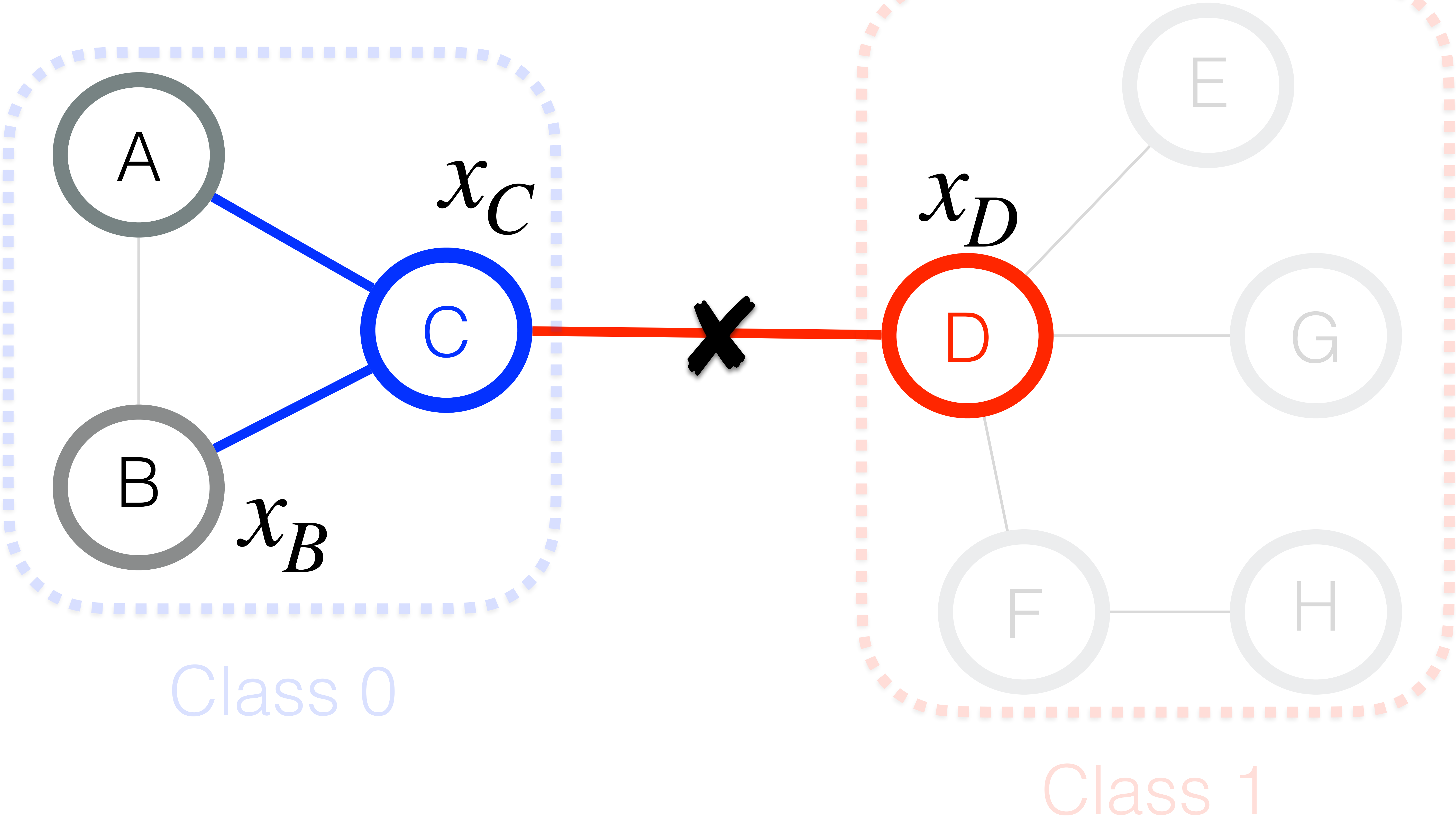
$$x'_A = \frac{1}{3}\left(x_A + x_B + x_C\right)$$



T. N. Kipf and M. Welling. Semi-Supervised Classification with Graph Convolutional Networks, ICLR 2017

Graph Convolution Network (GCN)

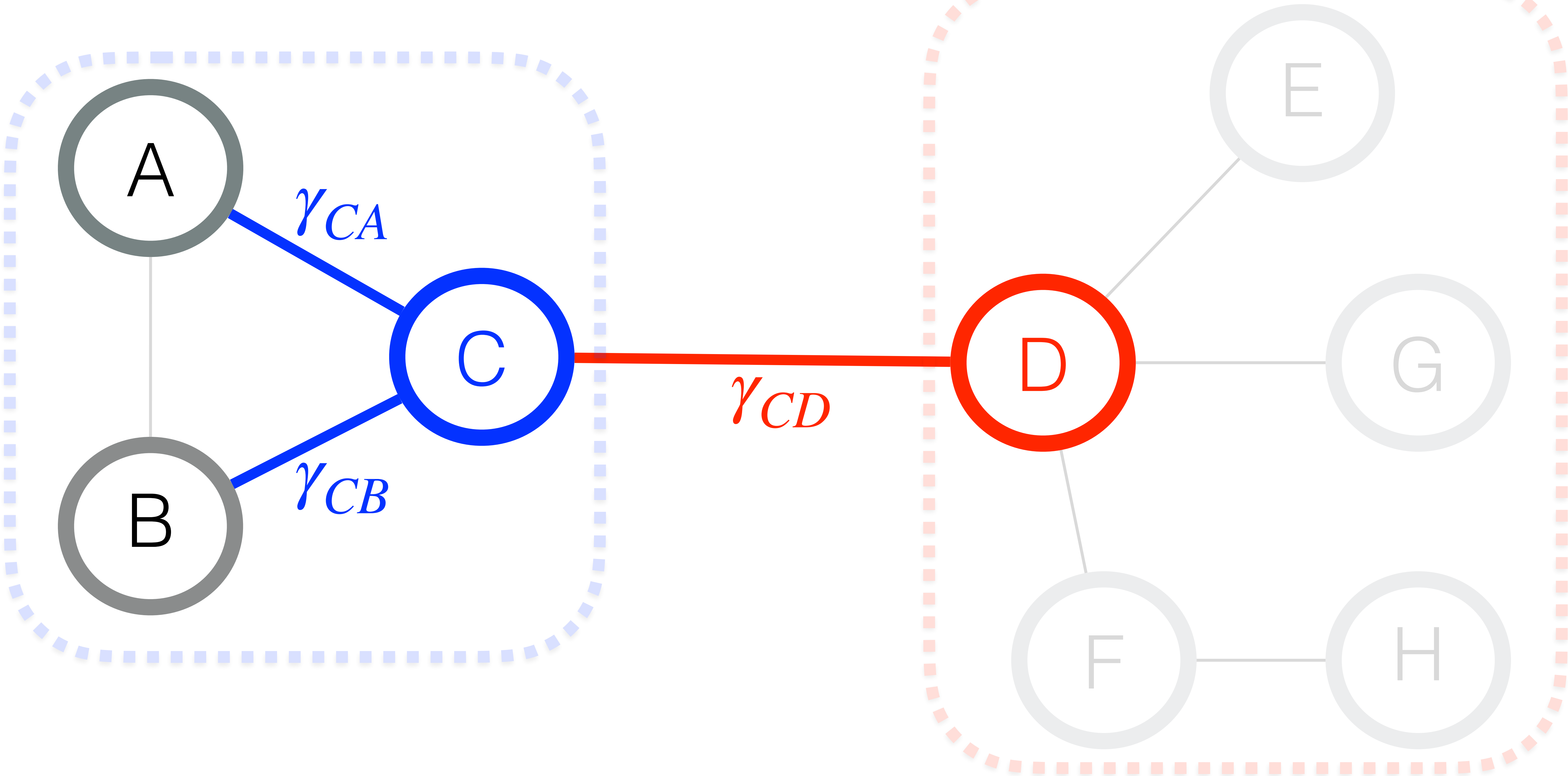
$$x'_C = \frac{1}{3}\left(x_C + x_A + x_B + \cancel{x_D}\right)$$



T. N. Kipf and M. Welling. Semi-Supervised Classification with Graph Convolutional Networks, ICLR 2017

Graph Attention Network (GAT)

$$x'_C = \gamma_{CC} x_C + \gamma_{CA} x_A + \gamma_{CB} x_B + \gamma_{CD} x_D$$



P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y. Bengio. Graph Attention Networks, ICLR 2018

We ask:
How successfully can graph attention discriminate neighbours?

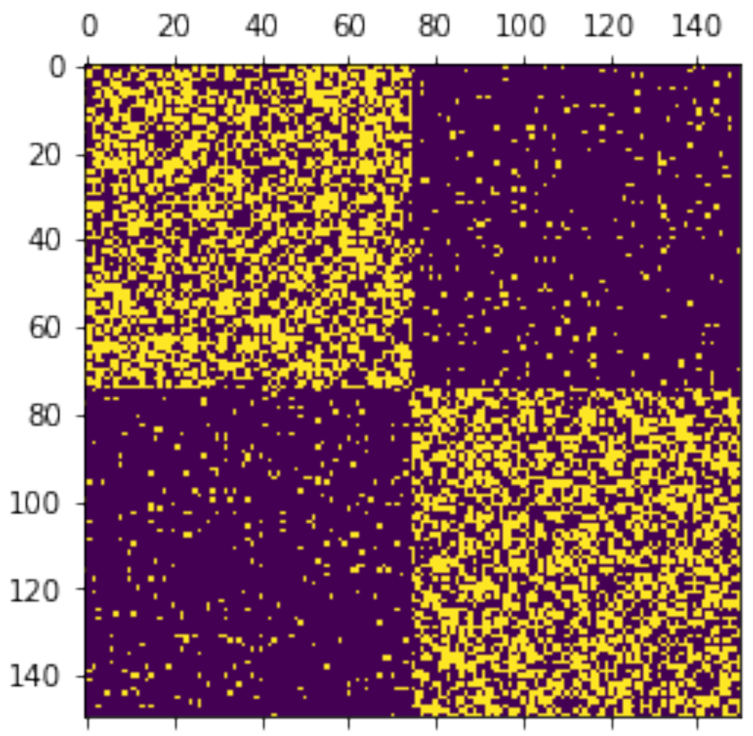
Data model: contextual stochastic block model

- Two-component balanced Gaussian Mixture Model (GMM) coupled with a Stochastic Block Model (SBM)

$$A \sim SBM(p, q)$$

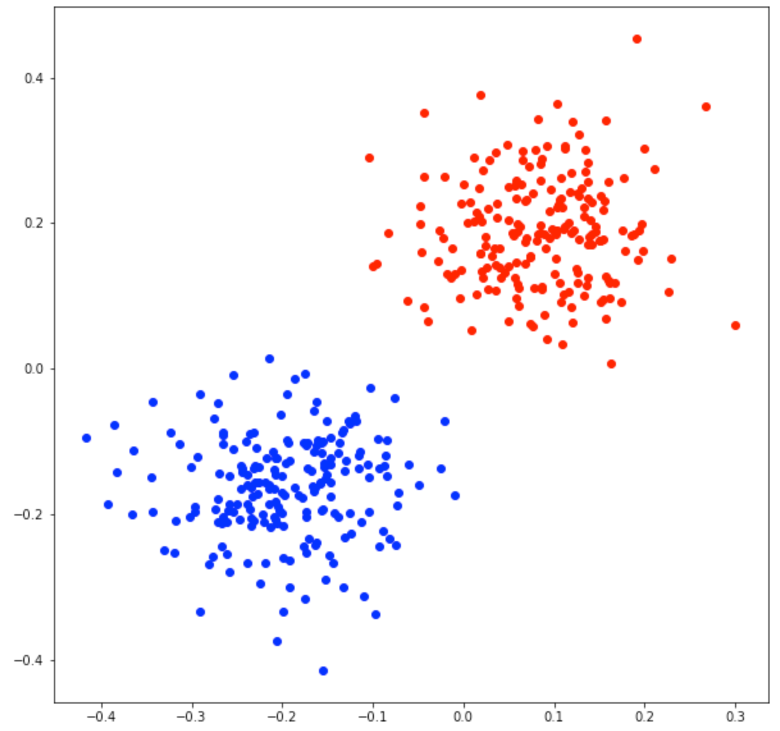
$$\mathbb{P}(A_{ij} = 1) = \begin{cases} p & \text{if } i, j \text{ are in the same class} \\ q & \text{otherwise} \end{cases}$$

$A =$



$X_i \sim \mathcal{N}(\mu, \sigma^2 I)$ if $i \in C_0$

$X_i \sim \mathcal{N}(-\mu, \sigma^2 I)$ if $i \in C_1$



Results (informal)

Hard regime
$\|\mu\| \leq K\sigma$

K const.

- MLP: constant fraction of misclassified nodes

- GAT: 90% of learned edge weights are approximately uniform $\Theta(1/N_i)$ (**no discrimination**)

K non const.

- MLP: at least one misclassified node

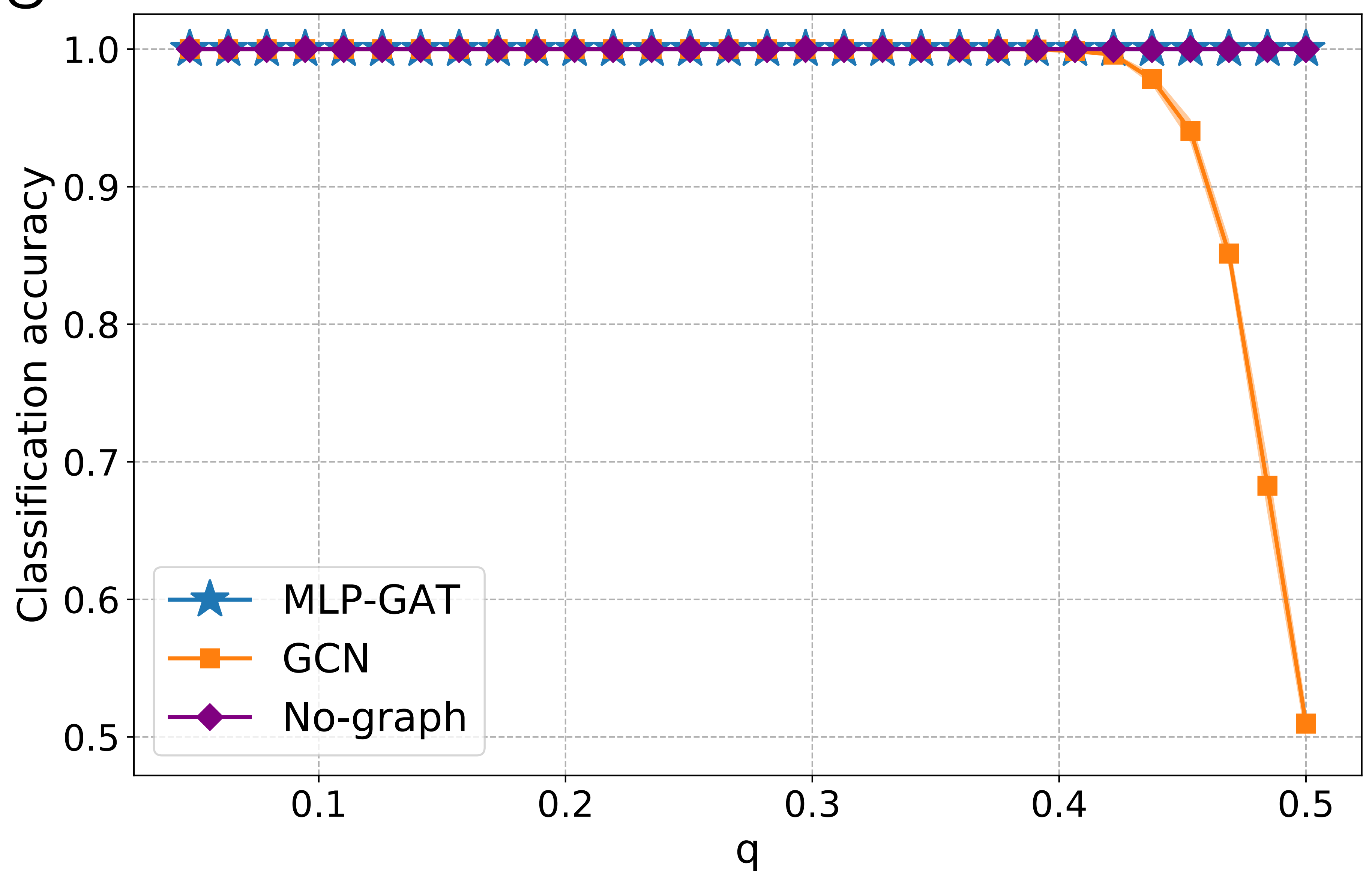
- GAT: at least one inter-edge is not down-weighted

Easy regime
$\|\mu\| \geq \sigma\sqrt{\log n}$

- MLP (no graph) achieves perfect classification

- GAT: significant down-weight of different-class edges

Distance between means
$\|\mu\|$



Results (informal)

Hard regime

$\|\mu\| \leq K\sigma$

Easy regime

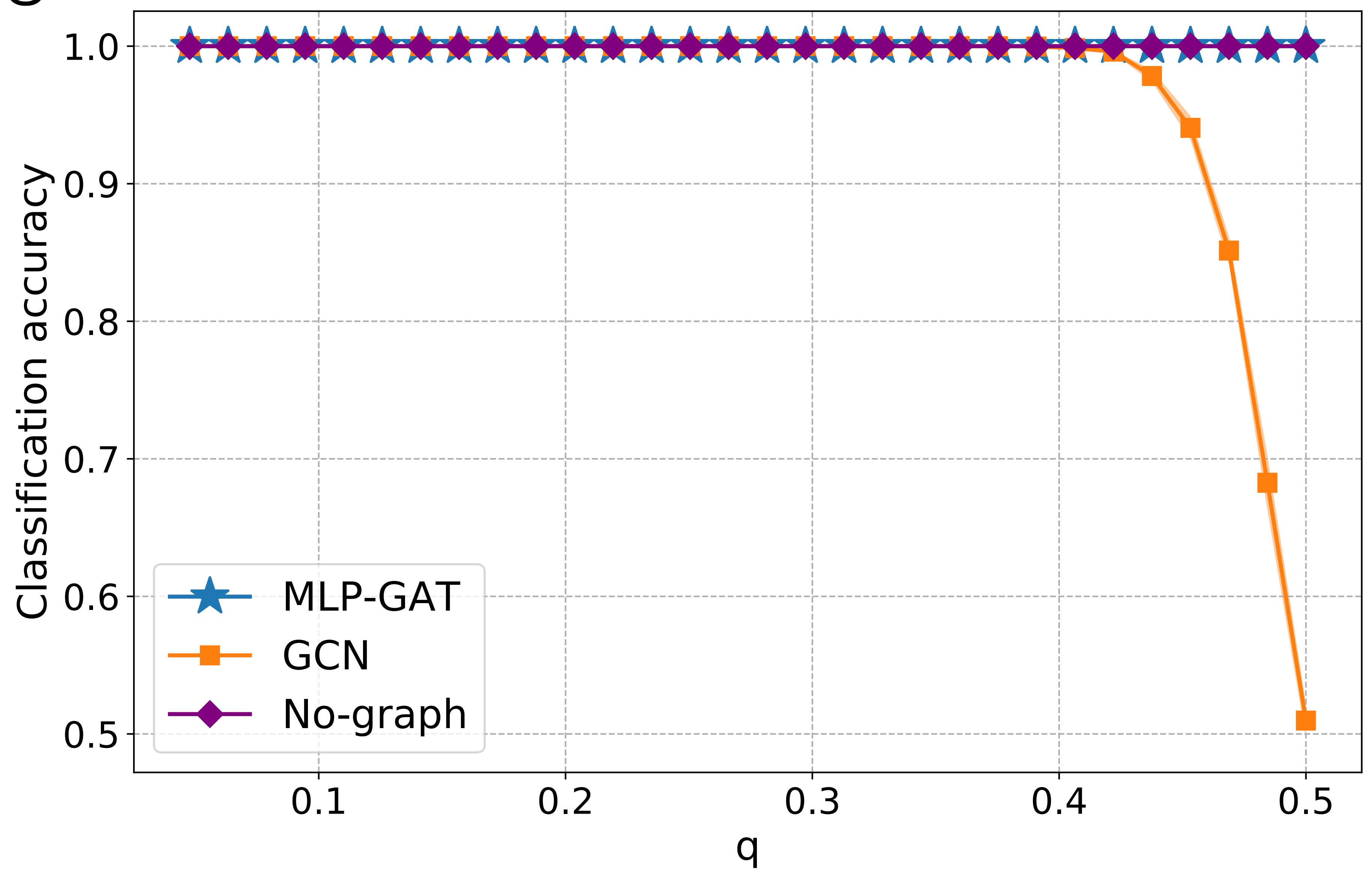
$\|\mu\| \geq \sigma\sqrt{\log n}$

K const. K non const.

- MLP: constant fraction of misclassified nodes
- MLP: at least one misclassified node
- MLP (no graph) achieves perfect classification

Distance between means $\|\mu\|$

- GAT: significant down-weight of different-class edges



- GAT: Node classification is possible, but it depends on q
- Conjecture: dependence on q is similar to GCN. Graph attention isn't better than GCN.

Results (informal)

Hard regime

$\|\mu\| \leq K\sigma$

K const.

- MLP: constant fraction of misclassified nodes

K non const.

- MLP: at least one misclassified node

Easy regime

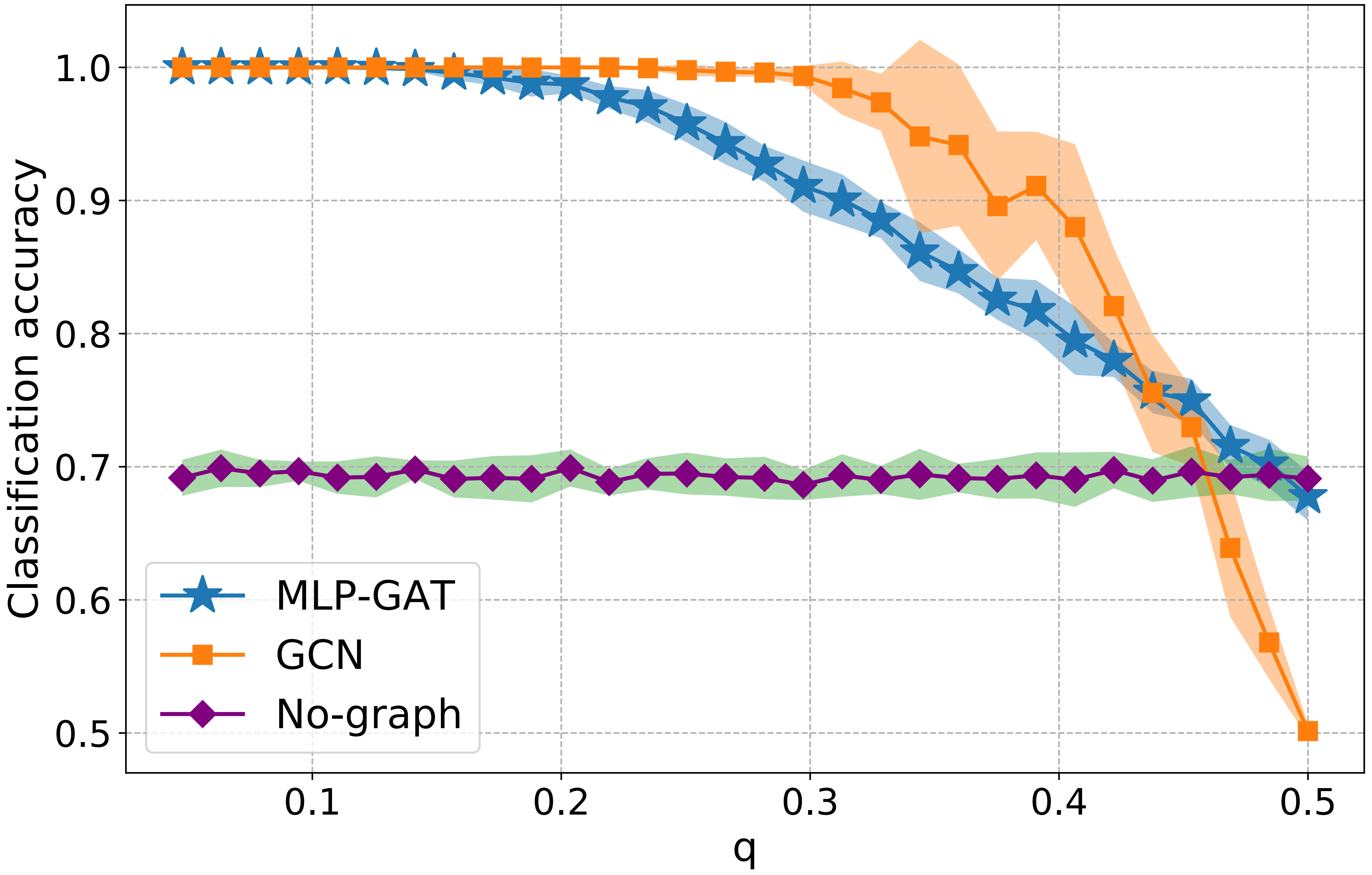
$\|\mu\| \geq \sigma\sqrt{\log n}$

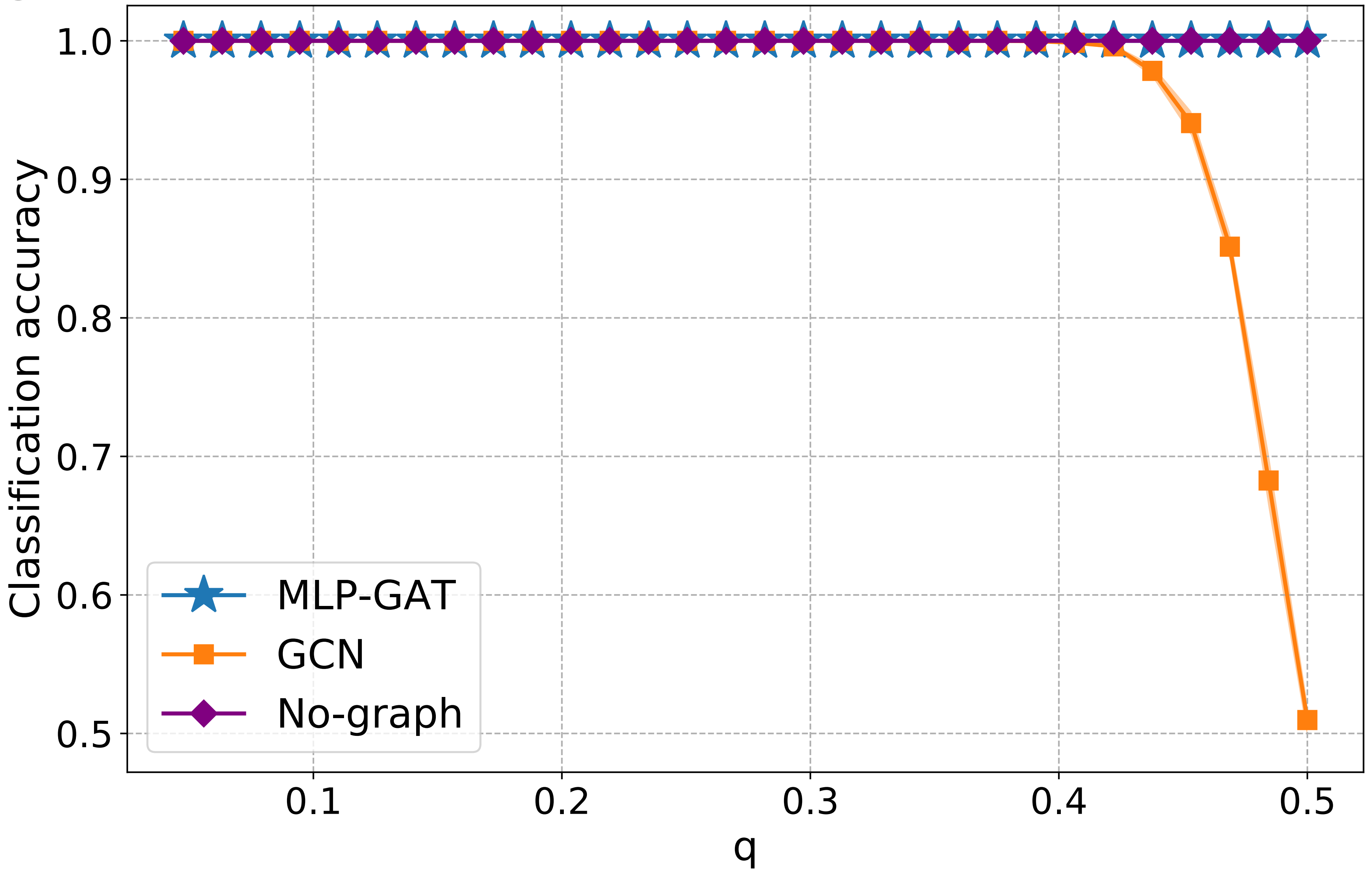
- MLP (no graph) achieves perfect classification

Distance between means

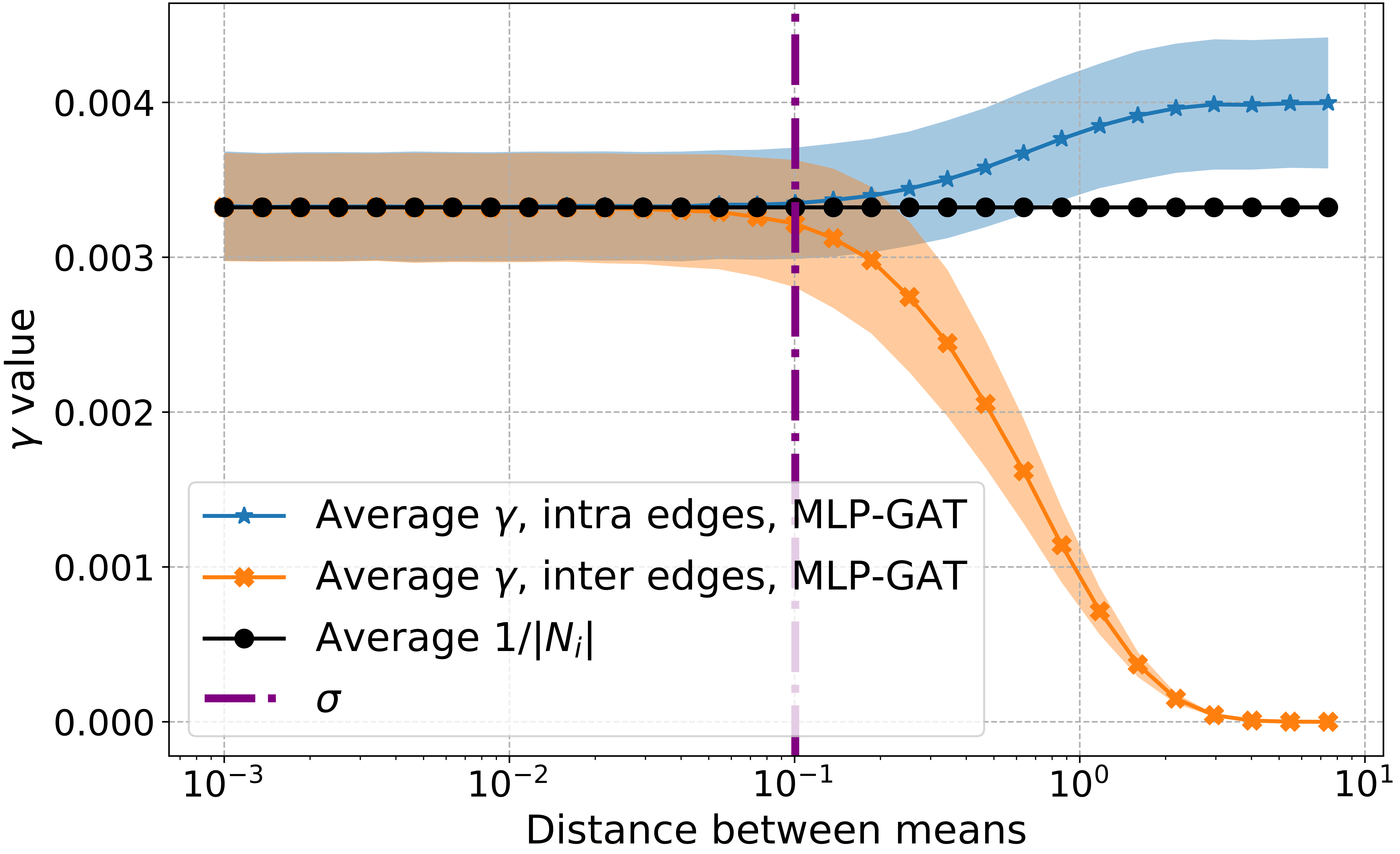
$\|\mu\|$

- GAT: significant down-weight of different-class edges

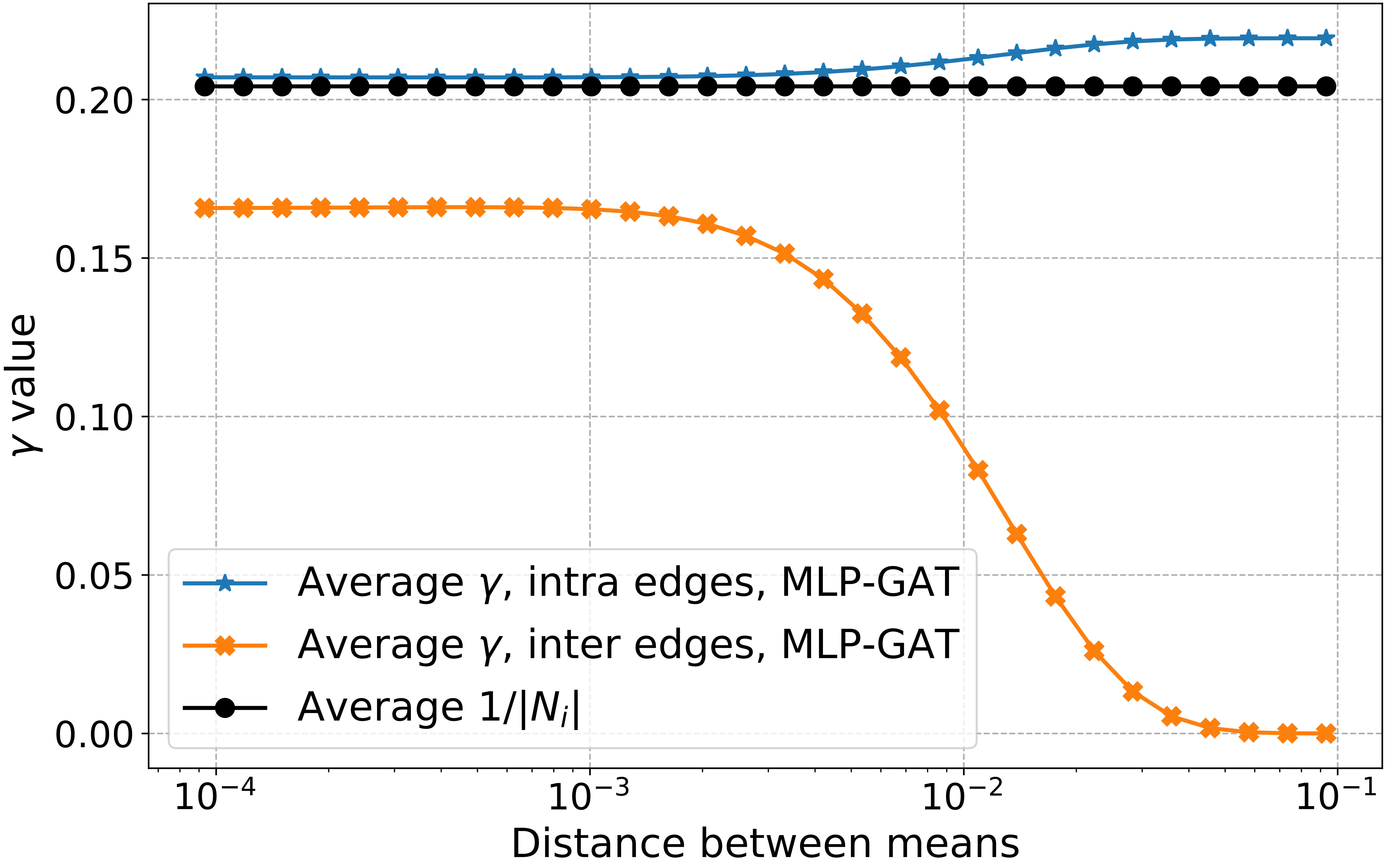




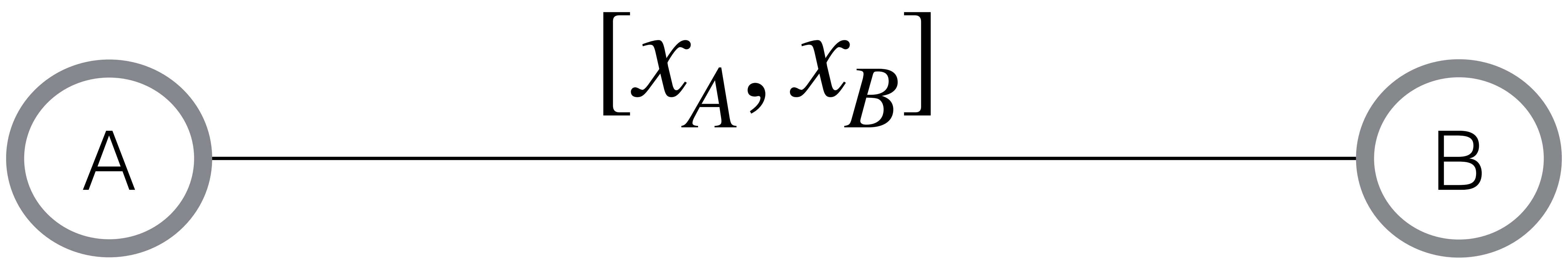
Empirical results (synthetic, fixed p and q)



Empirical results (real)



Why does graph attention fail to discriminate?

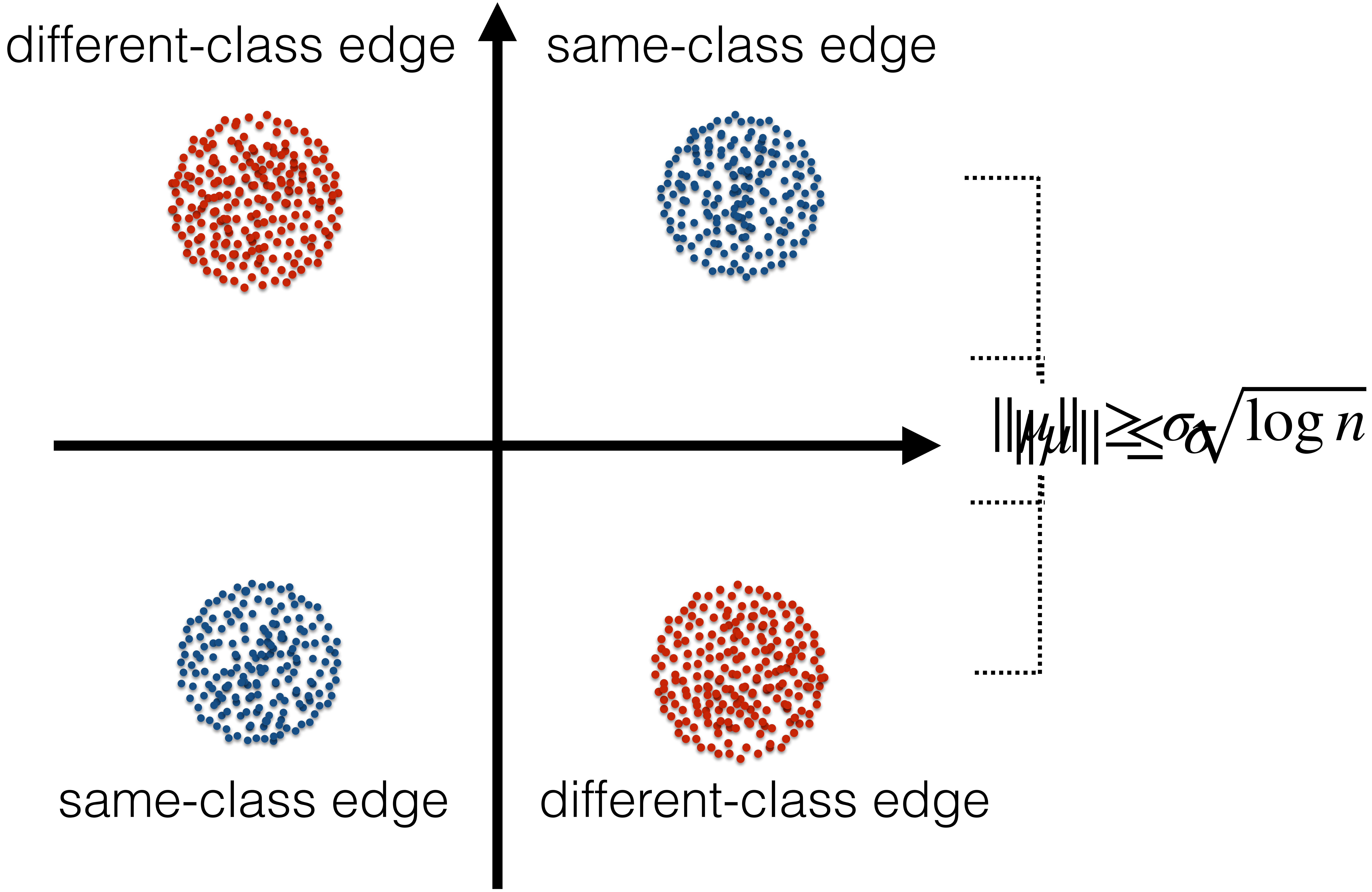


Why does graph attention fail to discriminate?

$$\gamma_{AB} = \psi\left(MLP\left([x_A, x_B]\right)\right)$$

ψ is a soft-max function

Why does graph attention fail to discriminate?



Conclusion

For our synthetic data model

- Attention is able to discriminate.
- Unfortunately, only when the graph is not needed to perfectly classify the nodes.
- This happens because current attention mechanisms rely only on utilizing the input data, which become very “noisy” faster than we start seeing any benefits from convolution.

For real data

- We demonstrate very similar observations on real data too.

Details

Assumptions

- Intra-class edge probability $p = \Omega\left(\frac{\log^2 n}{n}\right)$
- Inter-class edge probability $q = \Omega\left(\frac{\log^2 n}{n}\right)$
- $p \geq q$
- Thus, the expected number of neighbours is $\Omega(\log^2 n)$

The GAT convolution

Convolution

$$x_i' = \sum_{j \in [n]} A_{ij} \gamma_{ij} W x_j$$

Attention

$$\gamma_{ij} = \frac{\exp\left(\Psi(x_i, x_j)\right)}{\sum_{\ell \in N_i} \exp\left(\Psi(x_i, x_\ell)\right)}$$

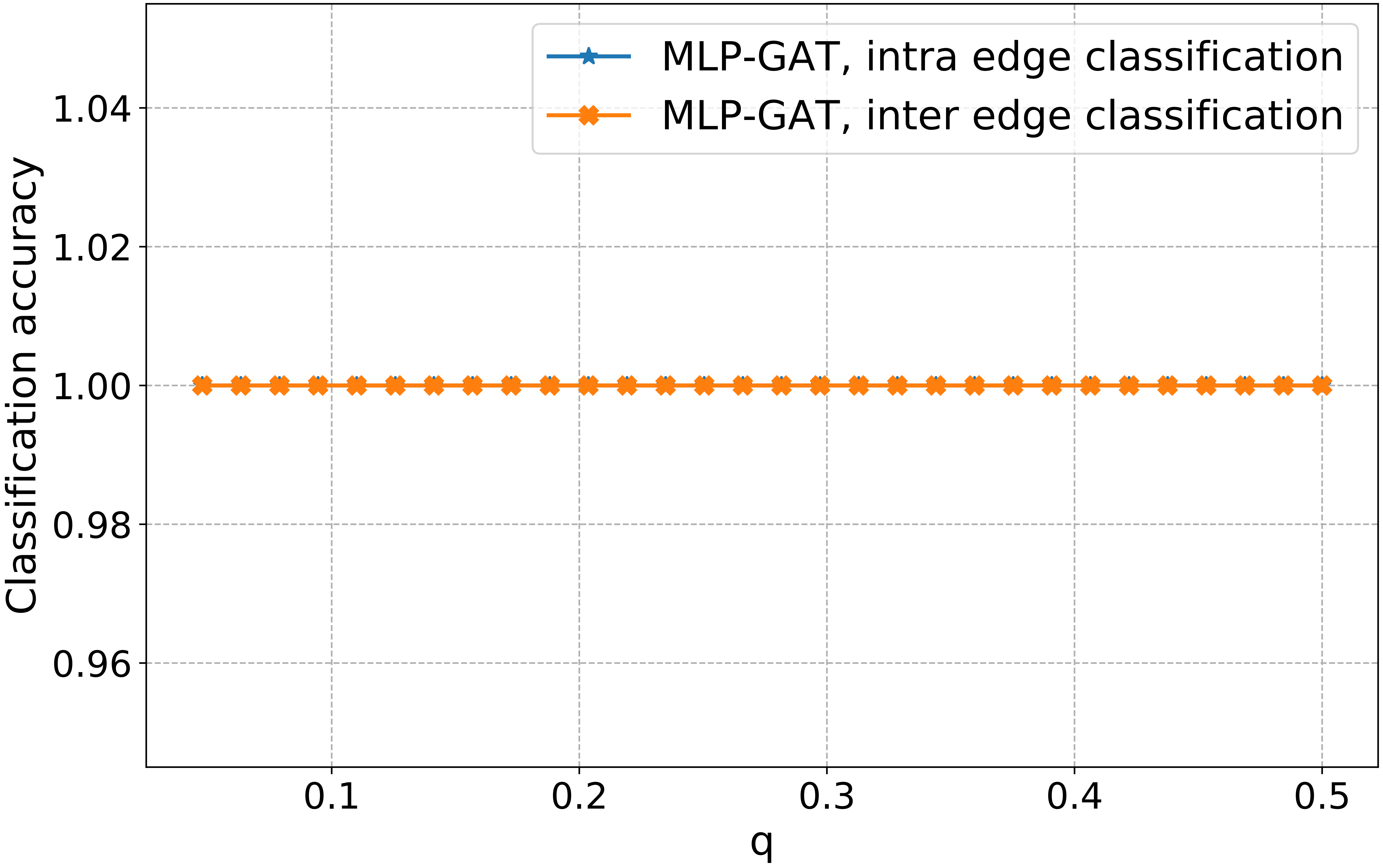
$$\Psi = \alpha\left(Wx_i, Wx_j\right)$$

where α can be an MLP

Result 1: Classification of edges, easy regime

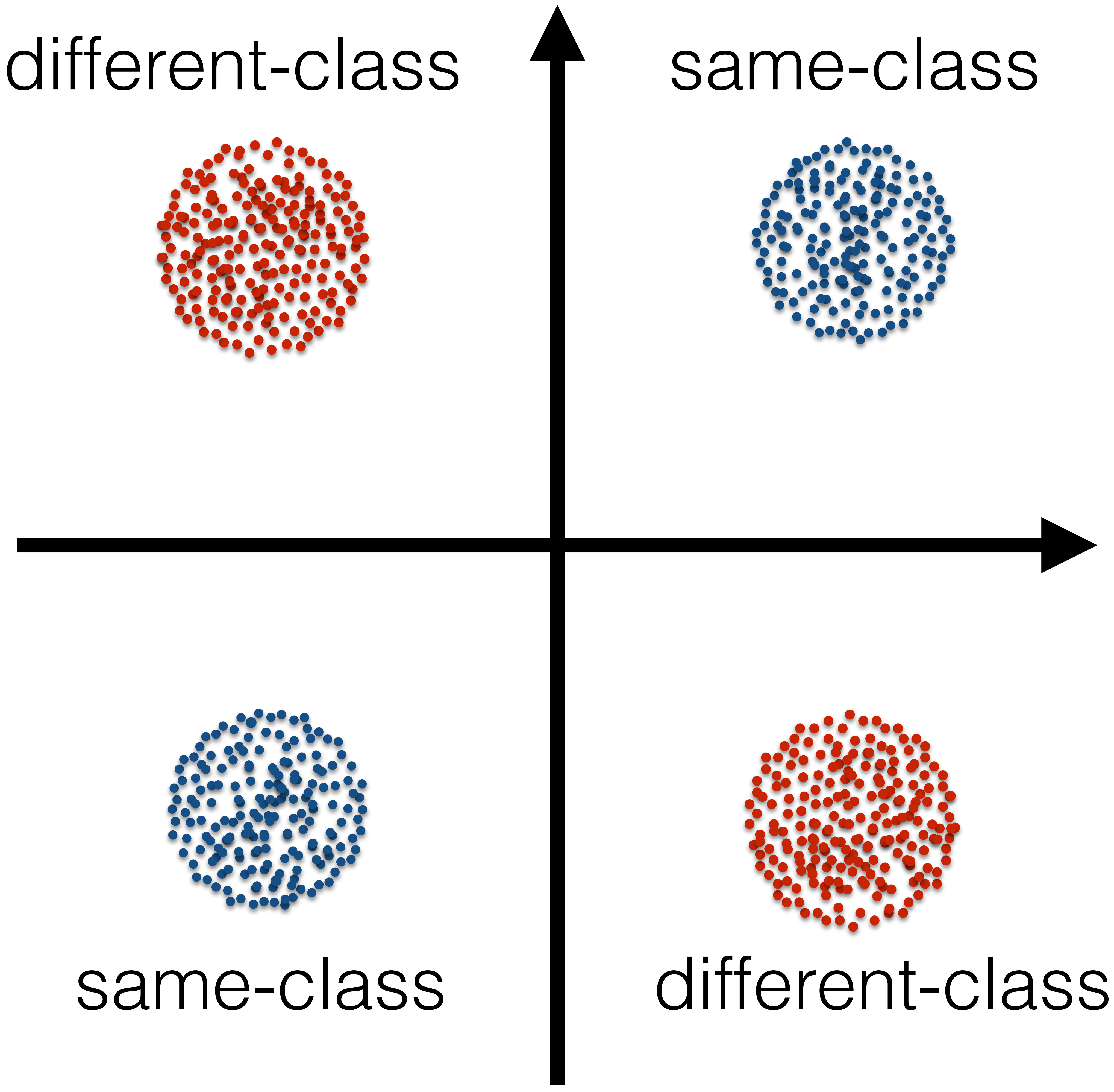
Theorem 1. *Suppose that* $\|\boldsymbol{\mu}\|_2 = \omega(\sigma\sqrt{\log n})$*. Then, there exists a choice of attention architecture* Ψ *such that with probability at least* $1 - o_n(1)$ *over the data* $(\mathbf{X}, \mathbf{A}) \sim \textsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ *it holds that* Ψ *separates intra-edges from inter-edges.*

Result 1: Classification of edges, easy regime



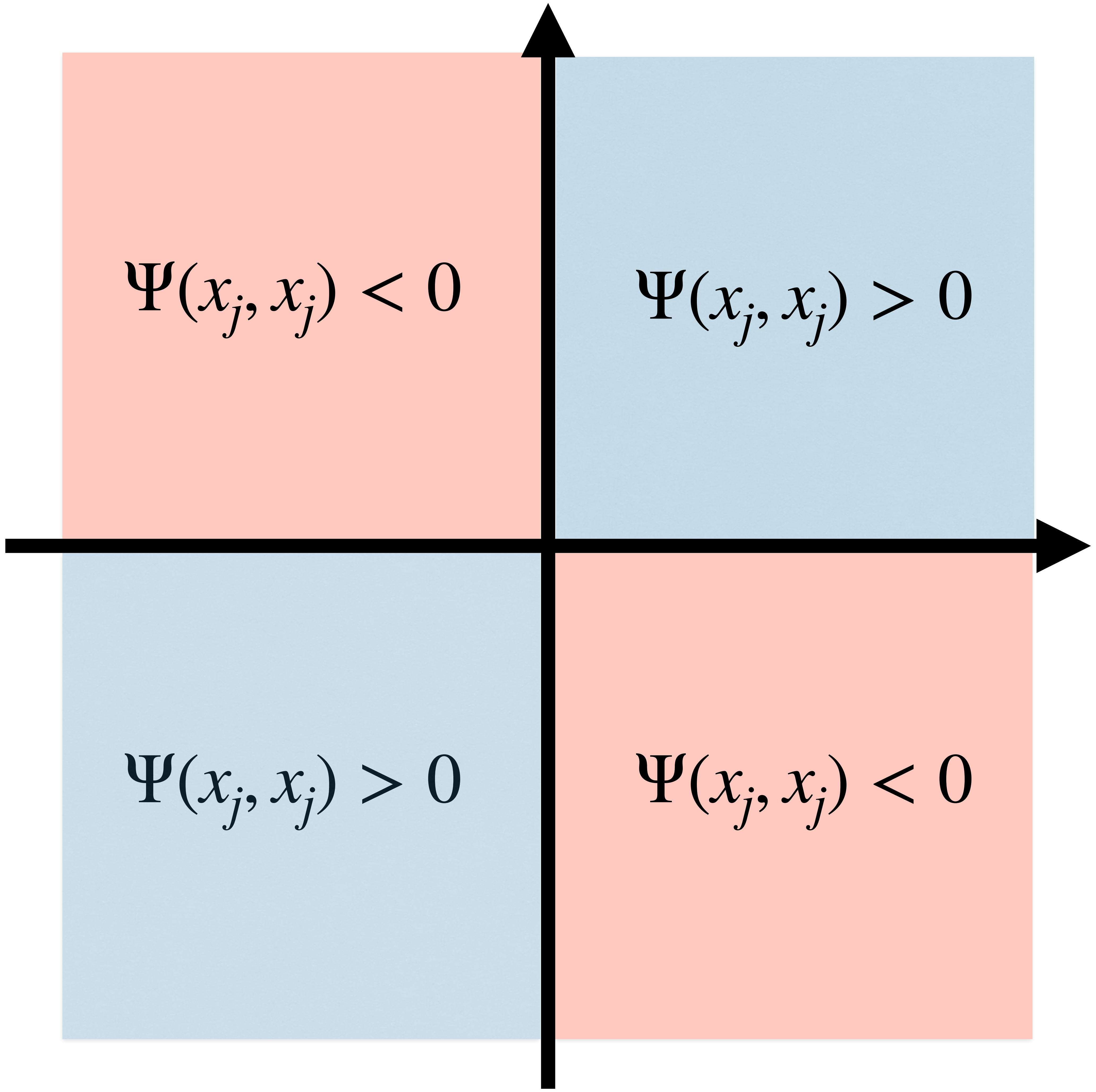
Proof sketch

- Our goal is to find an attention architecture Ψ that classifies the XOR problem

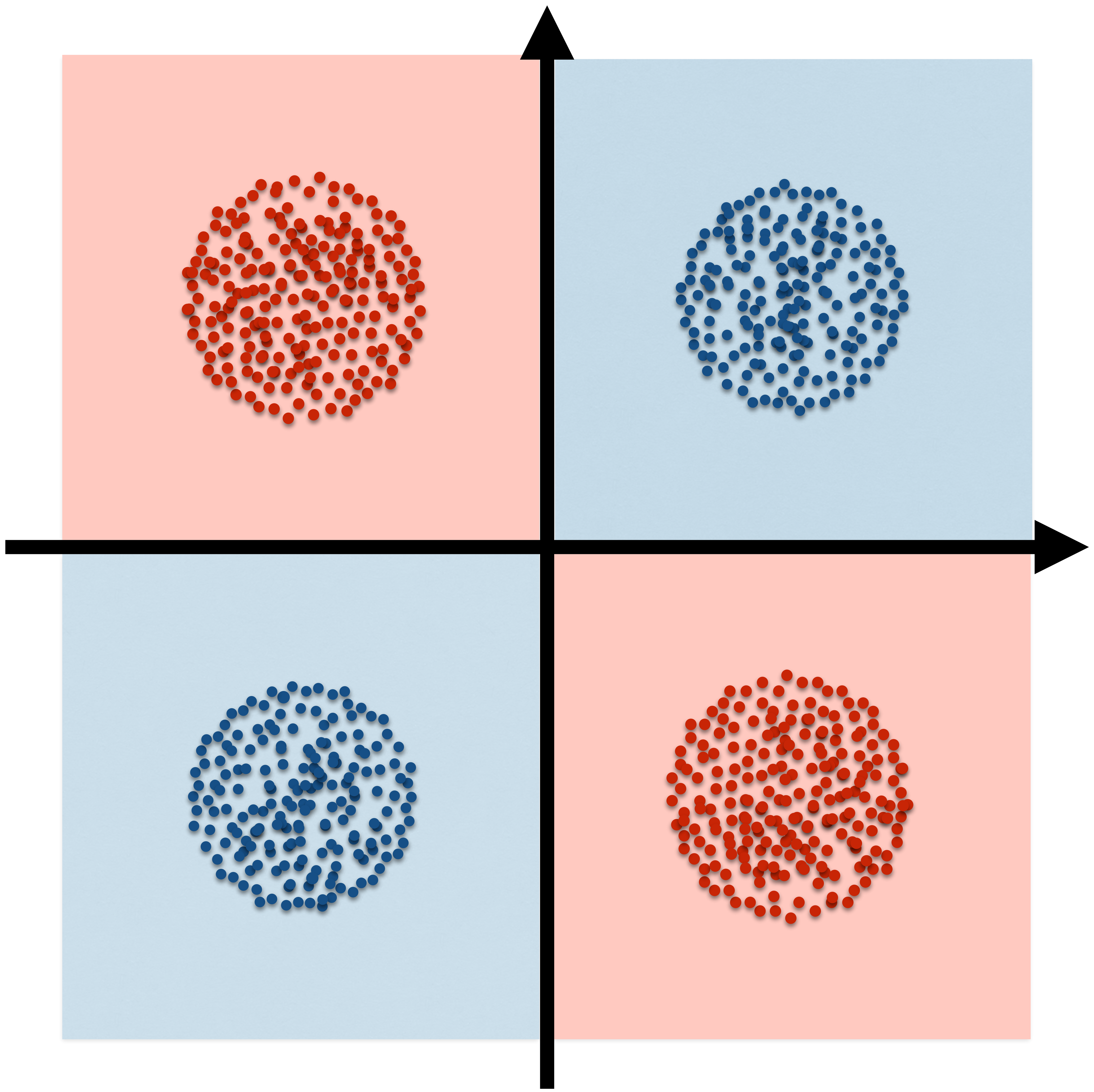


Proof sketch

- Goal: construct a Ψ with the following classification regions



Proof sketch



Proof sketch

- Construct Ψ that measures correlation with the means of the XOR problem.

$$\Psi(x_i, x_j) = r \cdot \text{LeakyReLU}\left(S \cdot \begin{bmatrix} w^T x_i \\ w^T x_j \end{bmatrix} \right)$$

$$S = \begin{bmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}$$

$$r = R \cdot [1 \quad 1 \quad -1 \quad -1]$$

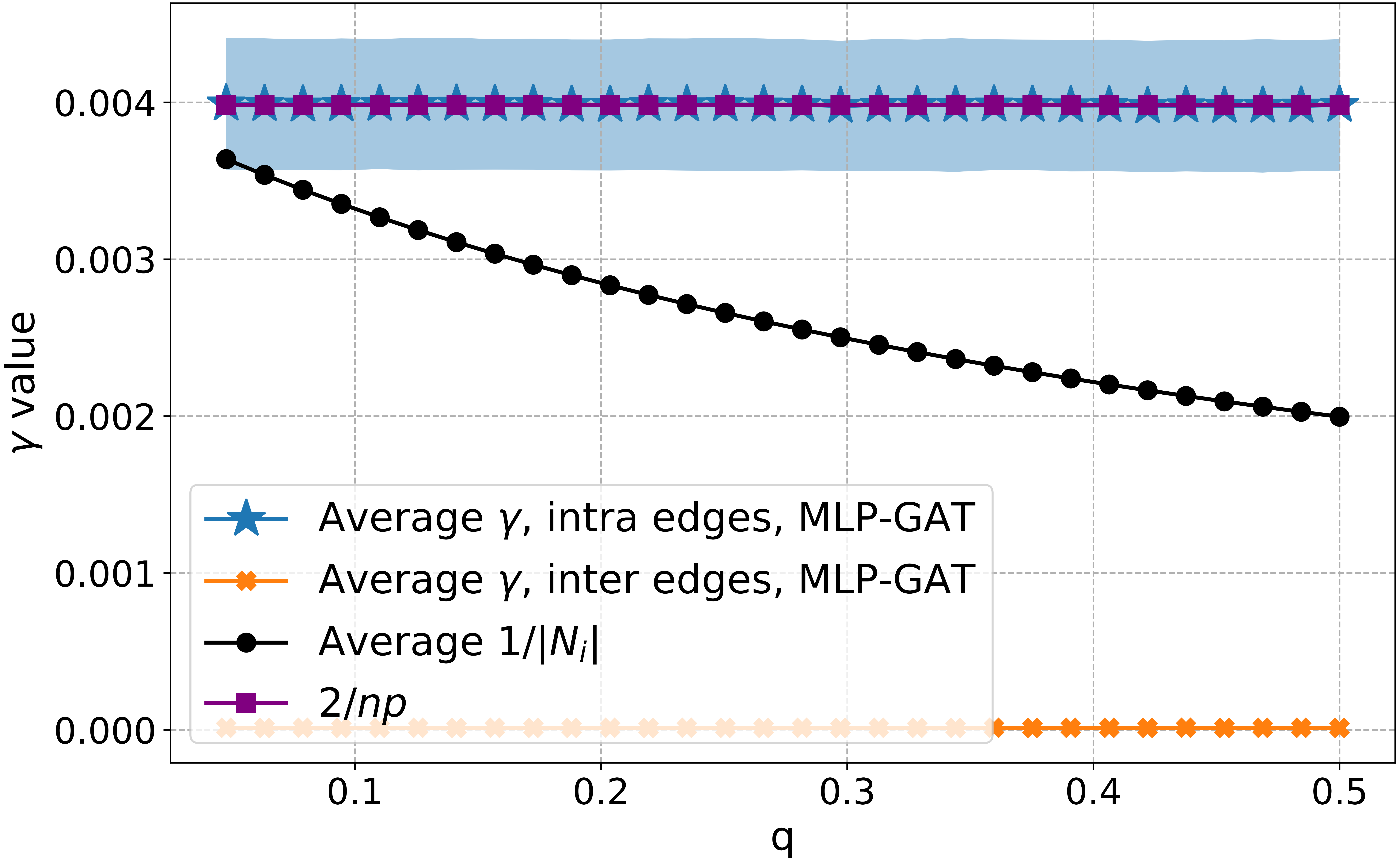
R controls the margin of classification

$$w = \mu / \|\mu\|_2$$

Result 2: Gammas, easy regime

Corollary 2. *Suppose that* $\|\boldsymbol{\mu}\|_2 = \omega(\sigma\sqrt{\log n})$. *Then there exists a choice of attention architecture* Ψ *such that with probability at least* $1 - o_n(1)$ *over the data* $(\mathbf{X}, \mathbf{A}) \sim \textsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$ *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.*

Result 2: Gammas, easy regime



Proof sketch

- From the edge classification result we have that

$$\Psi(x_i, x_j) \overset{whp}{=} \begin{cases} 2R\|\mu\|_2(1-\beta)(1 \pm o(1)) & \text{if } i, j \in C_1 \\ 2R\|\mu\|_2(1-\beta)(1 \pm o(1)) & \text{if } i, j \in C_0 \\ -2R\|\mu\|_2(1-\beta)(1 \pm o(1)) & \text{if } i \in C_1, j \in C_0 \\ -2R\|\mu\|_2(1-\beta)(1 \pm o(1)) & \text{if } i \in C_0, j \in C_1 \end{cases},$$

- Using the above the definition of gammas we obtain the result.

$$\gamma_{ij} = \frac{\exp\left(\Psi(x_i, x_j)\right)}{\sum_{\ell \in N_i} \exp\left(\Psi(x_i, x_\ell)\right)}$$

Proof sketch

- Example of an intra-class edge

$$\gamma_{ij} \overset{whp}{=} \frac{\exp\left(2R\|\mu\|_2\right)}{\sum_{intra\ (i,j)} \exp\left(2R\|\mu\|_2\right) + \cancel{\sum_{inter\ (i,j)} \exp\left(-2R\|\mu\|_2\right)}_{\approx 0}} \overset{whp}{=} \frac{2}{np}$$

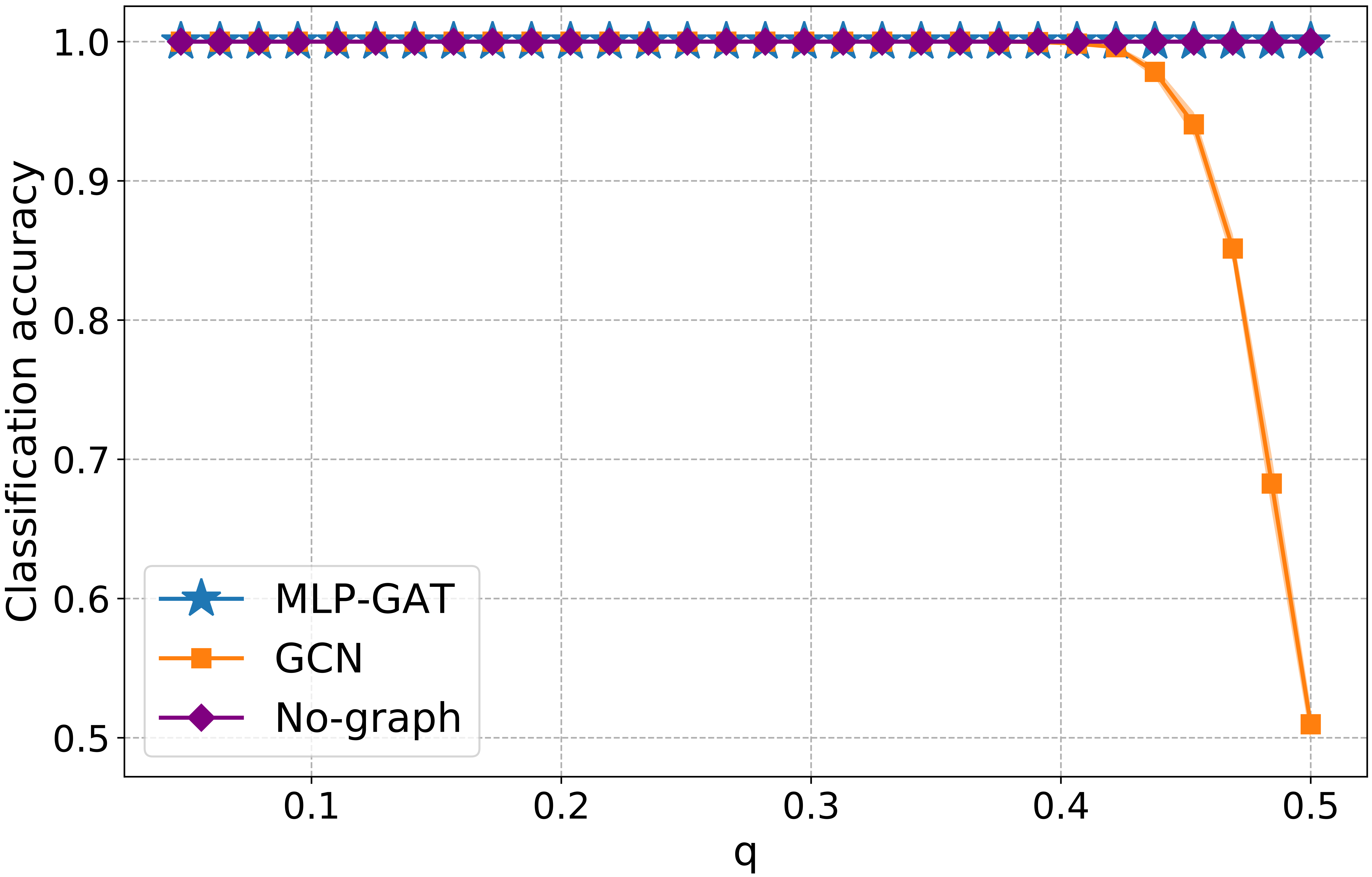
- Example of an inter-class edge

$$\gamma_{ij} \overset{whp}{=} \frac{\exp\left(-2R\|\mu\|_2\right)}{\sum_{intra\ (i,j)} \exp\left(2R\|\mu\|_2\right) + \sum_{inter\ (i,j)} \exp\left(-2R\|\mu\|_2\right)} = o\left(\frac{1}{N_i}\right) \overset{whp}{=} o\left(\frac{1}{n(p+q)}\right)$$

Result 3: node classification, easy regime

Corollary 3. *Suppose that* $\|\boldsymbol{\mu}\|_2 = \omega(\sigma\sqrt{\log n})$. *Then, there exists a choice of attention architecture* Ψ *such that with probability at least* $1 - o_n(1)$ *over the data* $(\mathbf{X}, \mathbf{A}) \sim \textsf{CSBM}(n, p, q, \boldsymbol{\mu}, \sigma^2)$*, the model separates the nodes for any* p, q *satisfying Assumption 1.*

Result 3: node classification, easy regime



Proof sketch

- From the previous result we have that

intra-class

$$\gamma_{ij} = \frac{2}{np}(1 \pm o_n(1))$$

inter-class

$$\gamma_{ij} = o\left(\frac{2}{n(p+q)}\right)$$

- Convolution reduces to

$$x_i' = \sum_{intra\ (i,j)} \frac{2}{np}(1 \pm o_n(1))w^T x_j + \cancel{\sum_{inter\ (i,j)} o\left(\frac{2}{n(p+q)}\right)}_{\approx 0} w^T x_j$$

Proof sketch

- The simplification of convolution implies that the new standard deviation is

$$\frac{\sigma}{\sqrt{np}}$$

- While the distance between the means is much larger

$$\|\mu\|_2 = \omega(\sigma\sqrt{\log n})$$

- And this implies perfect node classification with high probability

Result 4: classification of edges, hard regime

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

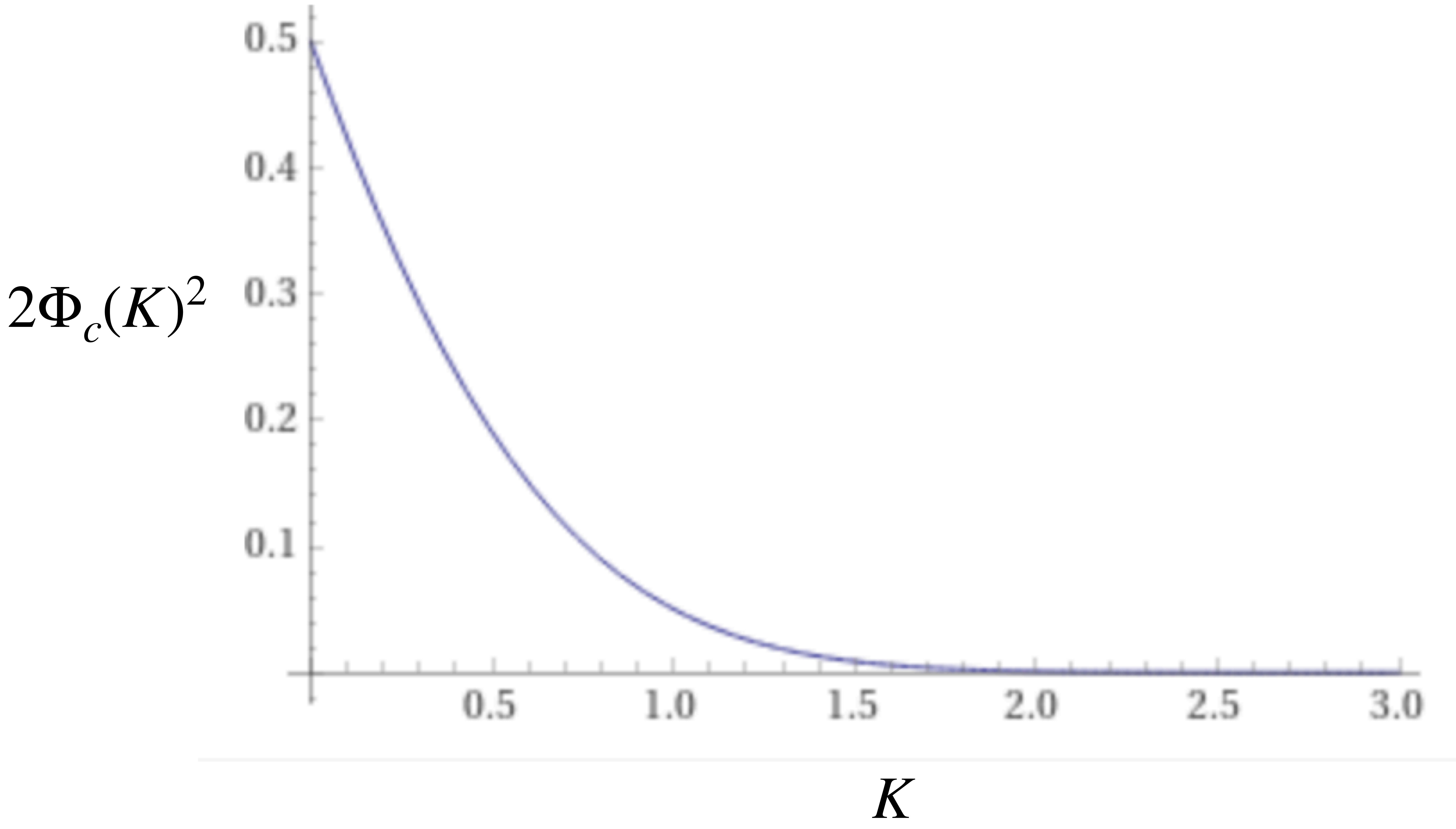
1. *For any* $c' > 0$*, with probability at least* $1 - O(n^{-c'})$*,* Ψ *fails to correctly classify at least a* $2 \cdot \Phi_{\mathrm{c}}(K)^2$ *fraction of the inter-edges.*
2. *For any* $\kappa > 1$ *if* $\boxed{q > \frac{\kappa \log^2 n}{n\Phi_{\mathrm{c}}(K)^2},}$ *then with probability at least* $1 - O\left(\frac{1}{n^{\frac{\kappa}{4}\Phi_{\mathrm{c}}(K)^2 \log n}}\right)$*,* Ψ *misclassify at least one inter-edge.*

Slightly denser than our initial assumption

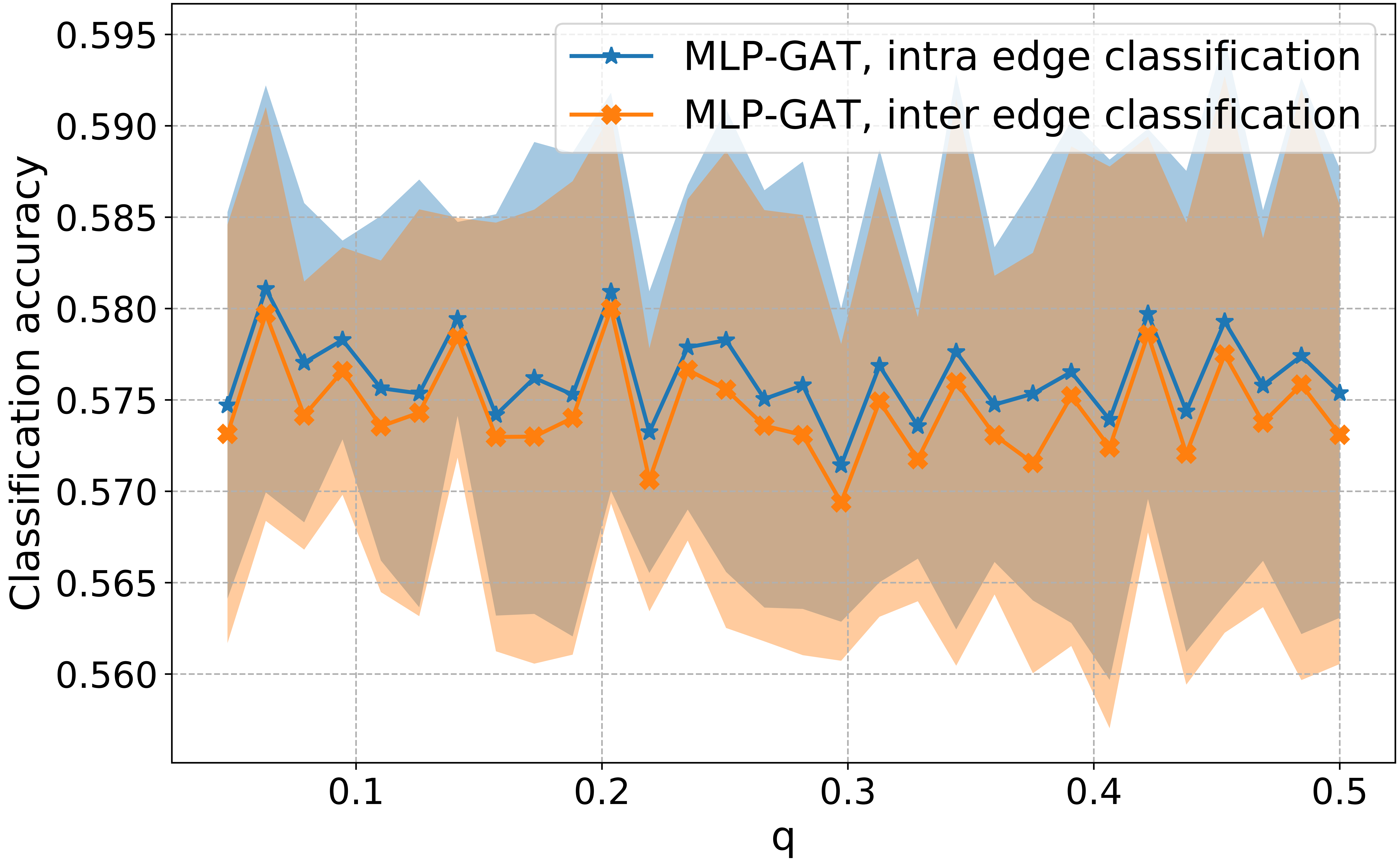
$$q = \Omega\left(\frac{\log^2 n}{n}\right)$$

$\Phi_c(K) = 1 - \Phi(K)$, where Φ is the cumulative density of standard normal

Result 4: classification of edges, hard regime



Result 4: classification of edges, hard regime

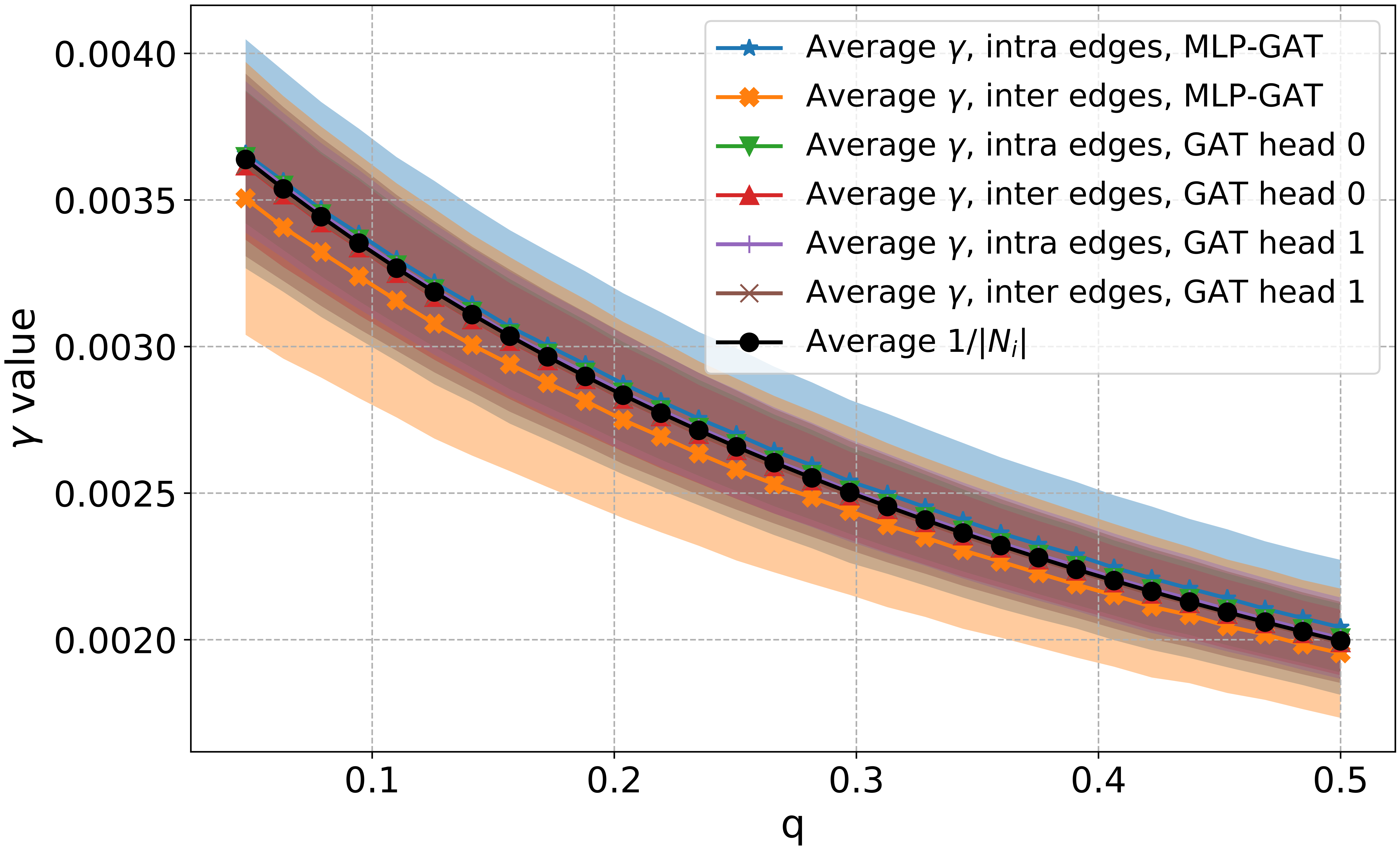


Result 5: gammas for a popular GAT model, hard regime

P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y. Bengio. Graph Attention Networks, ICLR 2018

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* $(\boldsymbol{w}, \boldsymbol{a}, b)$ *are bounded by a constant. Then, with probability at least* $1 - o_n(1)$ *over the data* $(\mathbf{X}, \mathbf{A}) \sim$ *CSBM*$(n, p, q, \boldsymbol{\mu}, \sigma^2)$, *at least* 90% *of* γ_{ij} *are* $\Theta(1/|N_i|)$.

Result 4: classification of edges, hard regime



Proof sketch

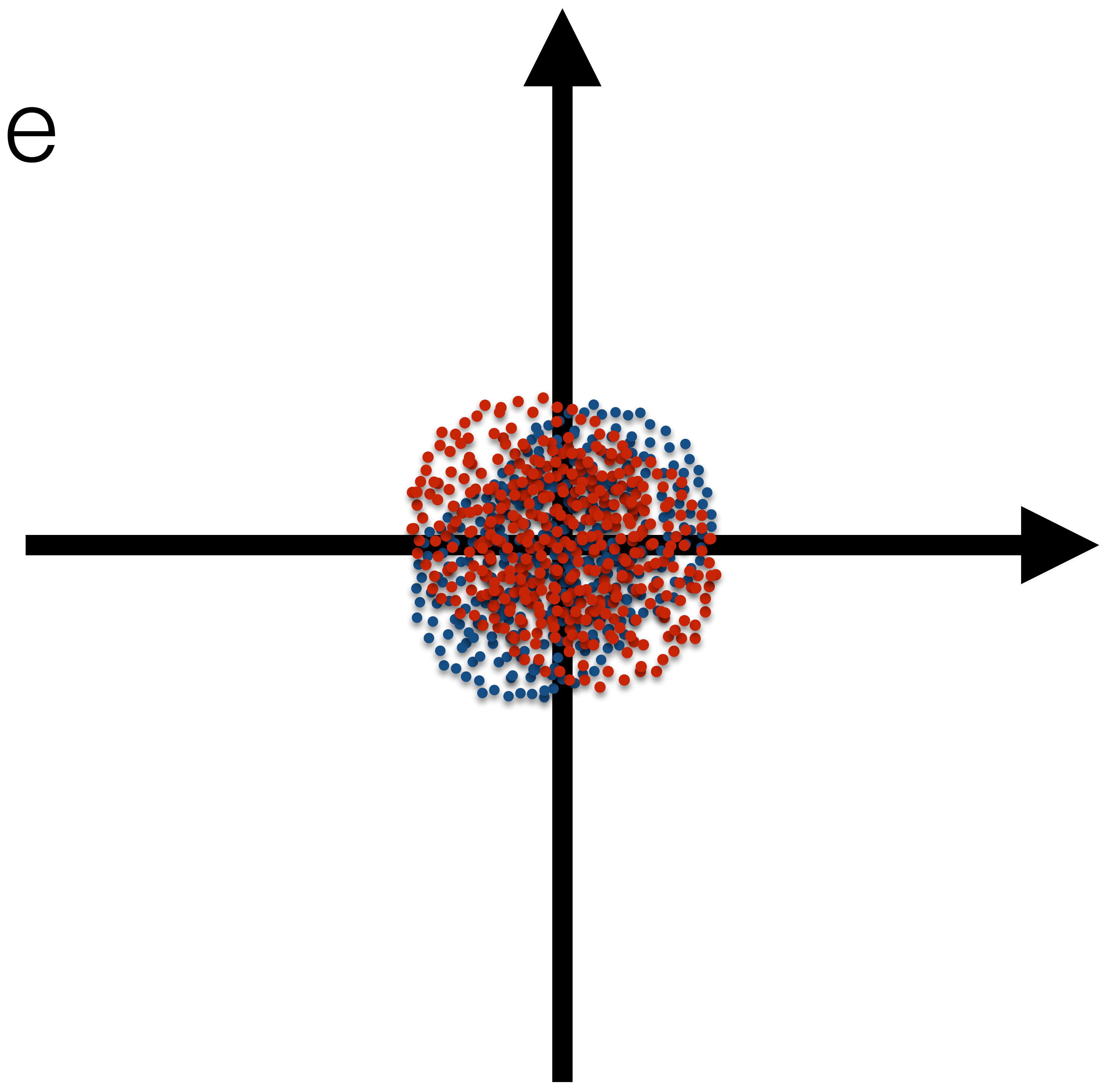
- The standard deviation is comparable to the distance between the means.

$$+$$

- Data act like Gaussian noise.

$$=$$

- The data are not indicative of class membership.

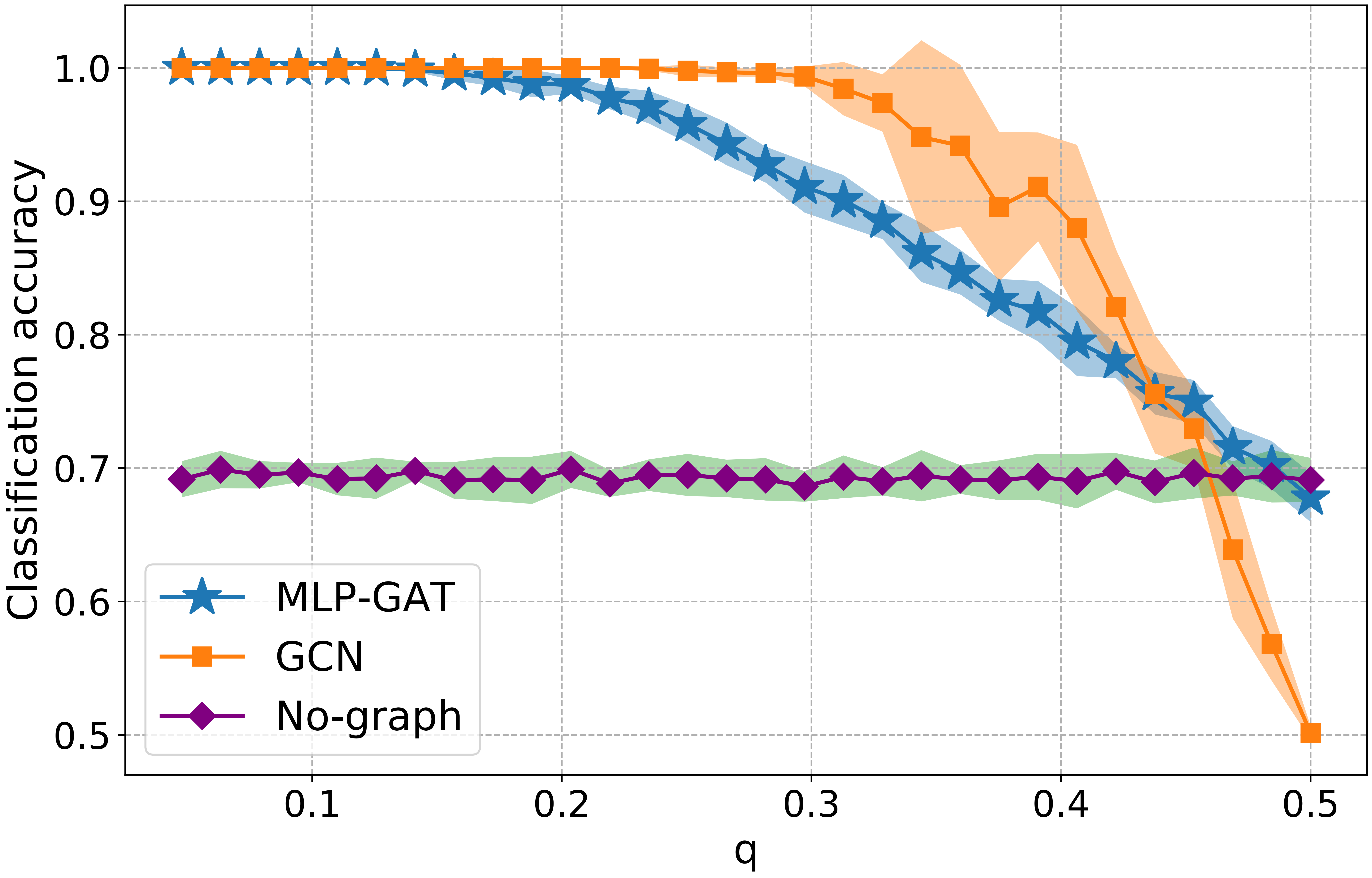


- Since the data behave like random noise, then a large fraction of $\Psi(x_i, x_j)$ are of constant magnitude, and this implies that γ is $\Theta(1/|N_i|)$.

Conjecture

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* Δ *is the expected degree.*

Conjecture



Classification accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.1
0.2
0.3
0.4
0.5
q
MLP-GAT
GCN
No-graph

Can the problem be fixed?

- To some extend, yes.

- Solution: convolve the data using GCN before applying attention.

- This will improve the threshold of edge separability to $\dfrac{\sigma\sqrt{\log n}}{\sqrt{n(p+q)}}$ from $\sigma\sqrt{\log n}$

A. Baranwal, K. Fountoulakis. A. Jagannath. Graph Convolution for Semi-Supervised Classification: Improved Linear Separability and Out-of-Distribution Generalization. ICML 2021.

Can the problem be fixed?

- But whenever we involve GCN, then all results will depend on parameter q.

- Conjecture: The improved version of GAT won't be better for node classification compared to GCN, since they will both depend on noise q in the same way.

Thank you!