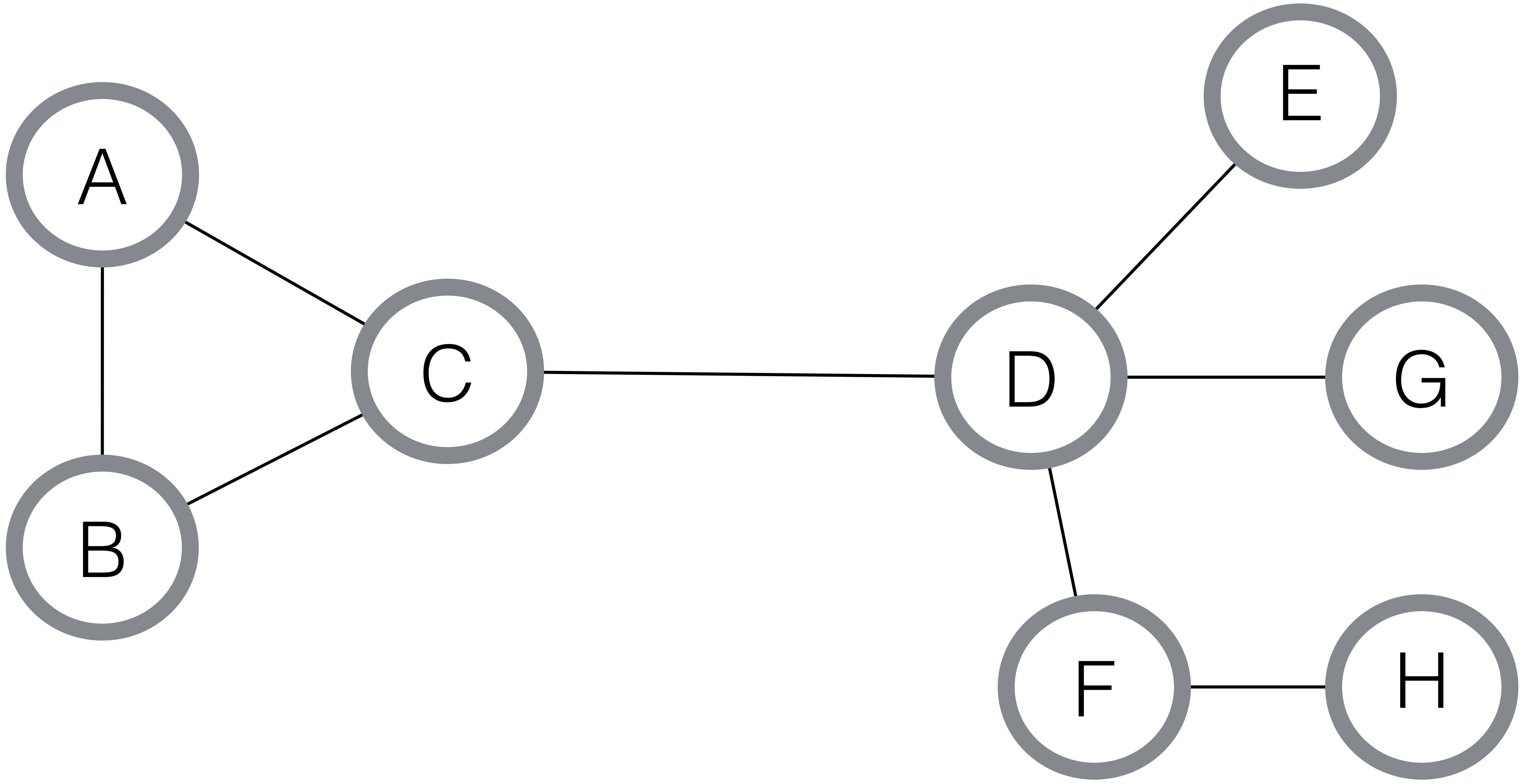
Lecture 4: Graph Attention Retrospective

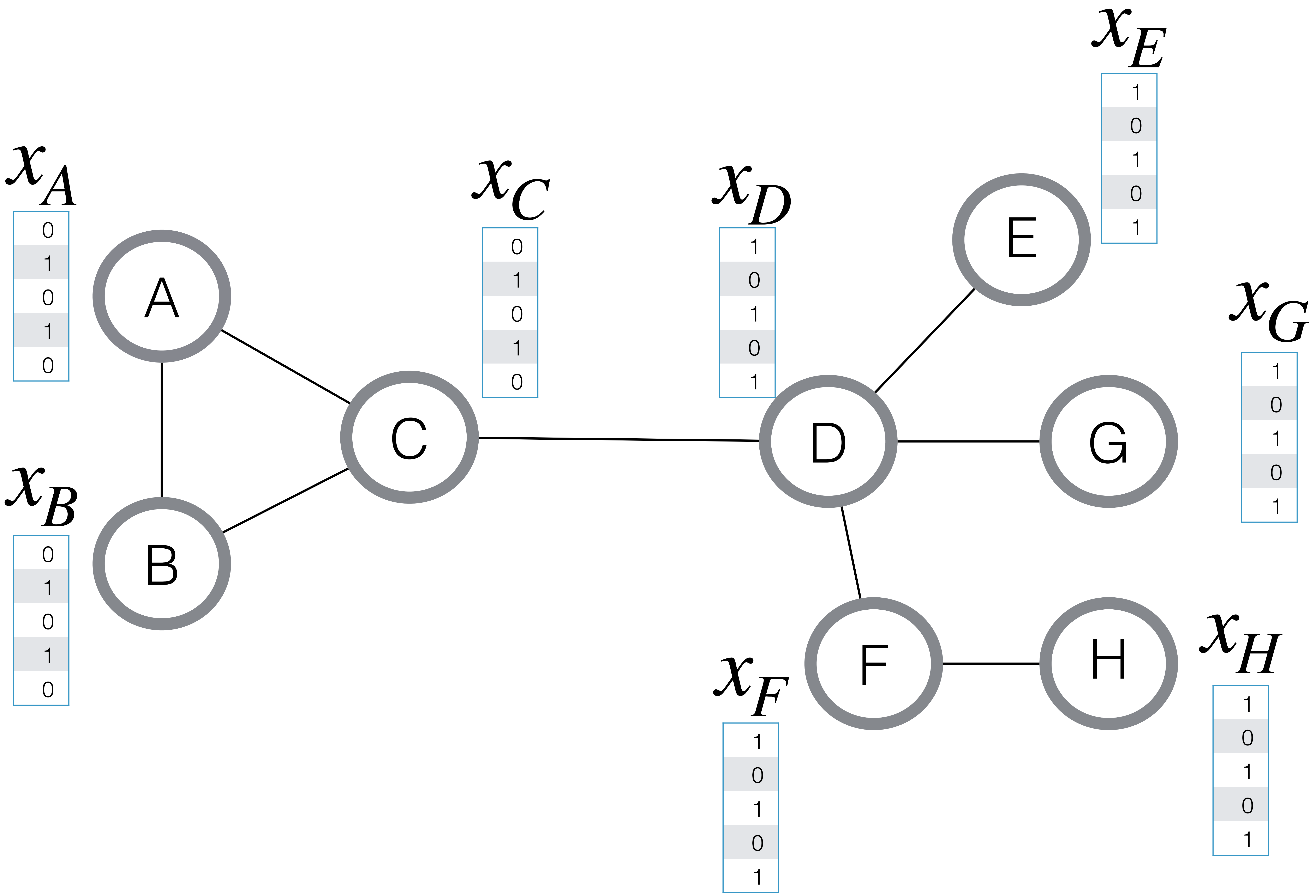
Kimon Fountoulakis

UNIVERSITY OF WATERLOO | DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Graphs

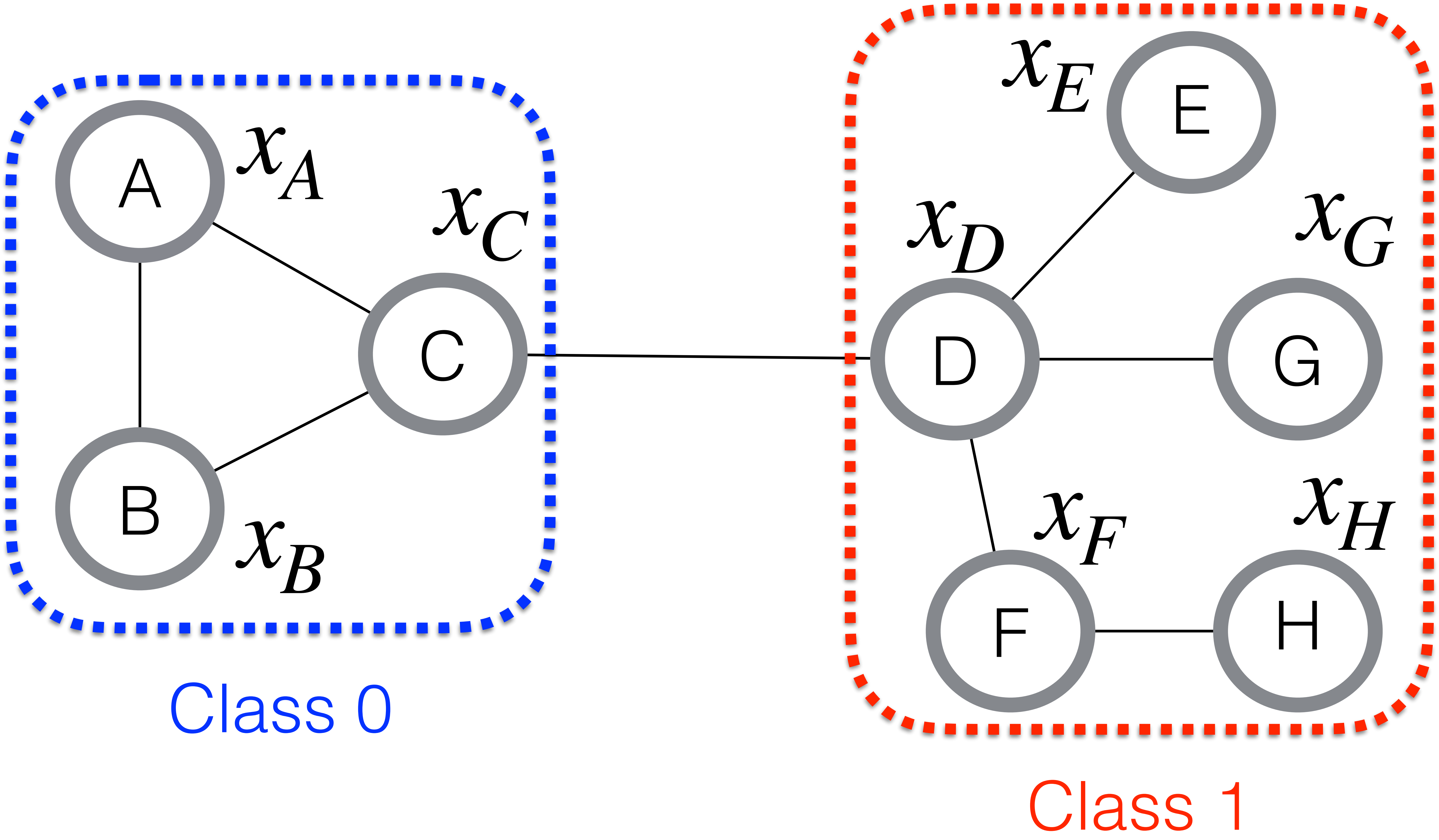


Graphs + features



- x_i is the feature vector for node i

Node classification



- x_i is the feature vector for node i

Node classification

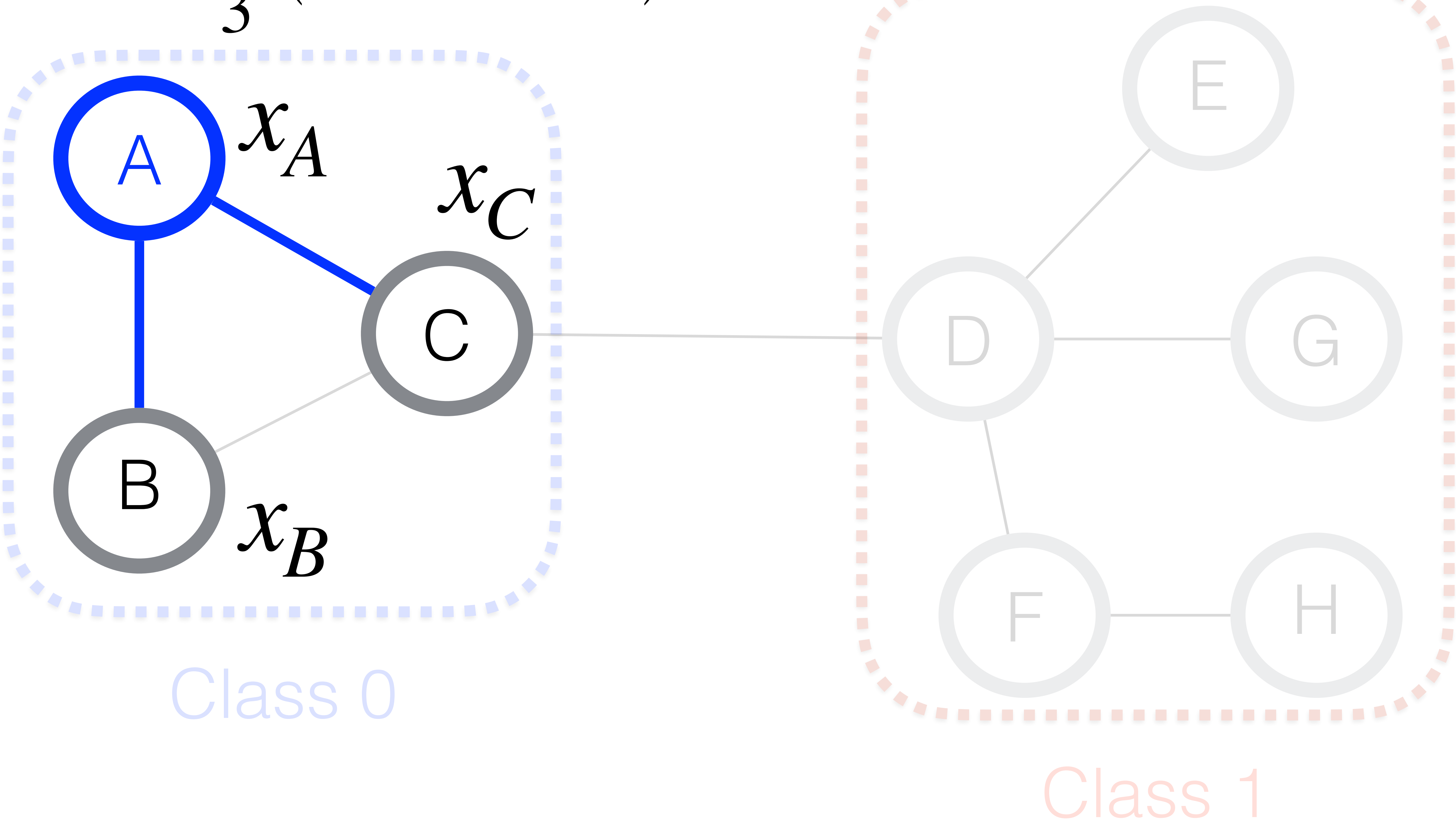
- Classification thresholds for perfect node classification (this work)
- Almost perfect or partial classification are not studied but are certainly good future directions.

Terminology

- intra-edge: an edge where its nodes are in the same class
- inter-edge: an edge where its nodes are in different class

Vanilla 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

Vanilla Graph Convolution Network (GCN)

$$X'_i := \frac{1}{D_{ii}} \sum_{i=1}^{n} A_{ij} X_j$$

Convolved data for node i

Degree of node i

Adjacency matrix

Data For node j

- A component of A is equal to 1 if two nodes are connected with an edge
- D is a diagonal matrix where each component shows the number of neighbors of a node

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

Vanilla Graph Convolution Network (GCN)

$$X' := D^{-1}AX$$

Convolved data　　Degree matrix　　Adjacency matrix

- A component of A is equal to 1 if two nodes are connected with an edge
- D is a diagonal matrix where each component shows the number of neighbors of a node

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

Vanilla Graph Convolution Network (GCN)

$$X'W := D^{-1}AXW$$

-Learning matrix W. It's value are decided by minimizing a loss function.

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

Vanilla Graph Convolution Network (GCN)

$$\sigma(X'W) := \sigma(D^{-1}AXW)$$

-Activation function σ. Examples include $\sigma(y) := \max(y,0)$ or $\sigma(y) := \text{sigmoid}(y) = 1/(1+e^{-y})$ which squeezes values in [0,1].

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

Vanilla Graph Convolution Network (GCN)

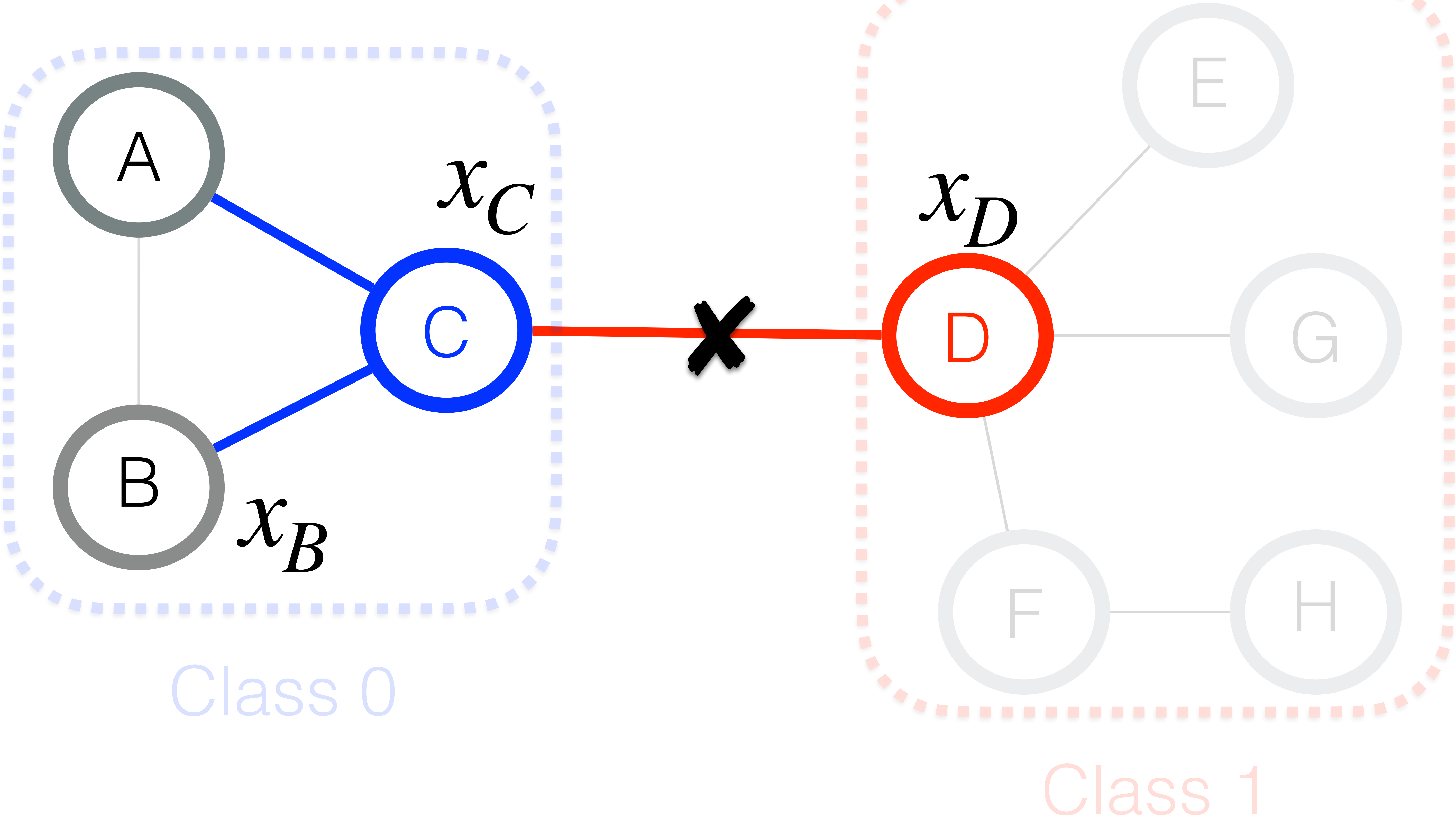
Example: 3-layer GCN

$$X' := \underbrace{\sigma_3(D^{-1}A\,\underbrace{\sigma_2(D^{-1}A\,\underbrace{\sigma_1(D^{-1}AXW_1)}_{\text{layer 1}}\,W_2)}_{\text{layer 2}}\,W_3)}_{\text{layer 3}}$$

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

Vanilla Graph Convolution Network (GCN)

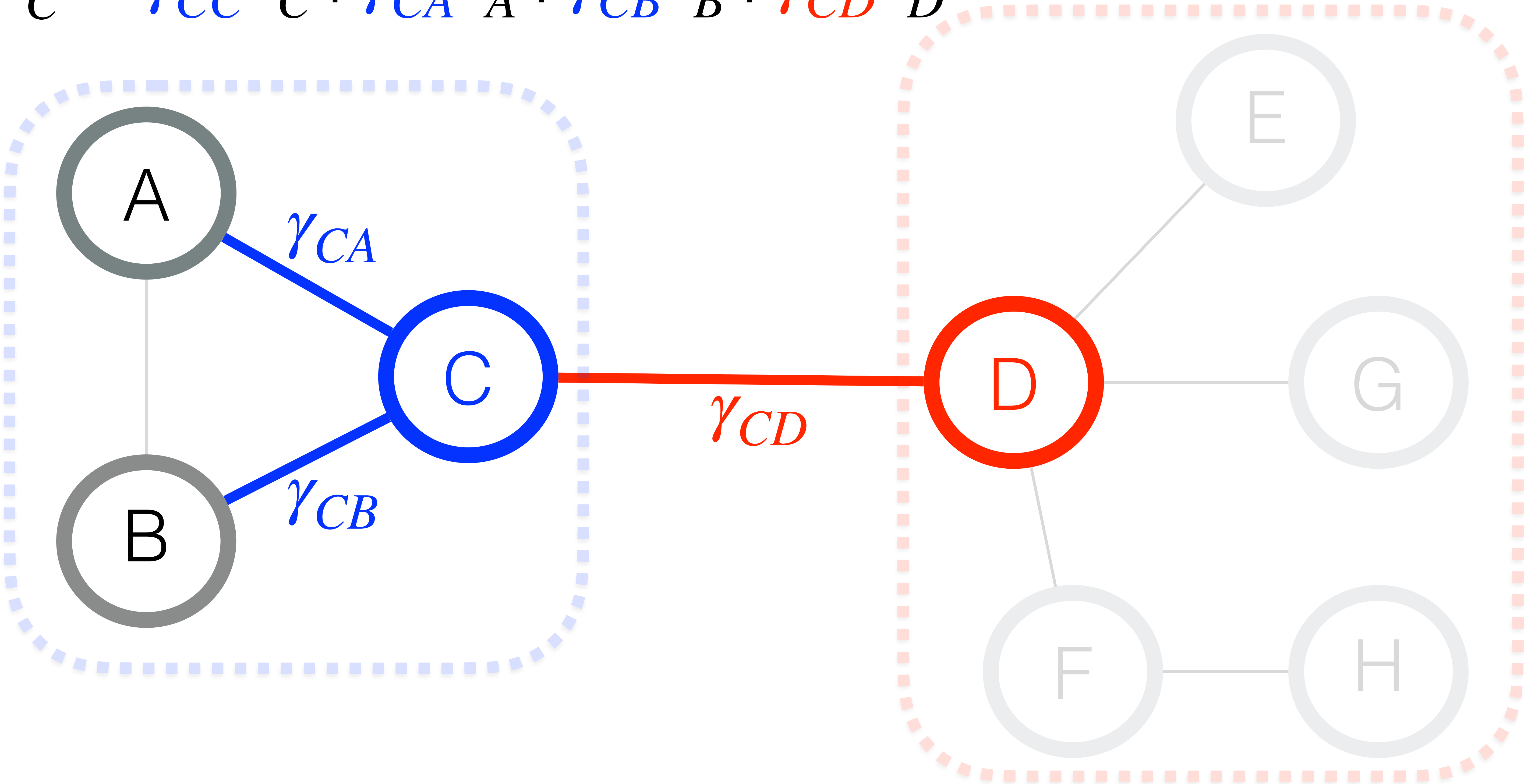
$$x'_C = \frac{1}{3}\left(x_C + x_A + x_B + \cancel{x_D}\right)$$



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

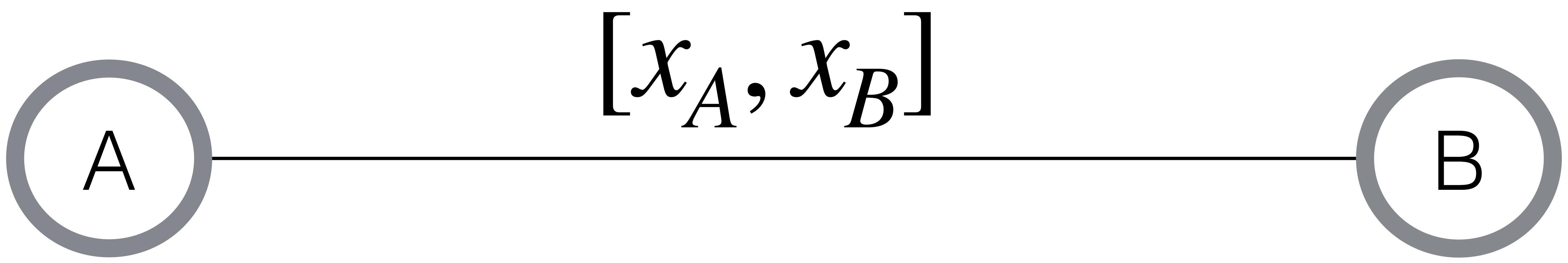
Vanilla Graph Attention Network (GAT)

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



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

Vanilla Attention Mechanism



Vanilla Attention Mechanism

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

ψ is a soft-max function

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

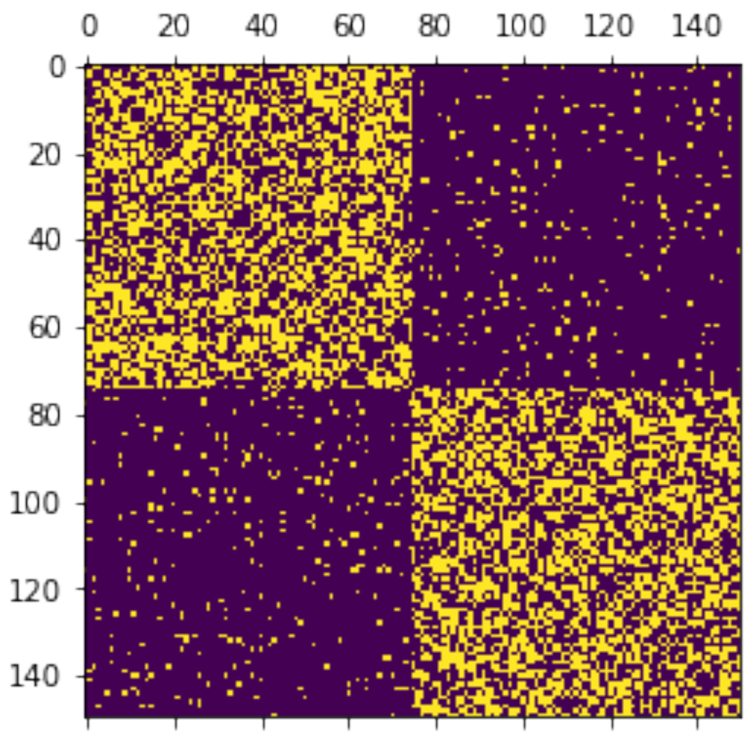
We ask:
How successfully can graph attention distinguish intra- from inter-edges?

Data model: contextual stochastic block model

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

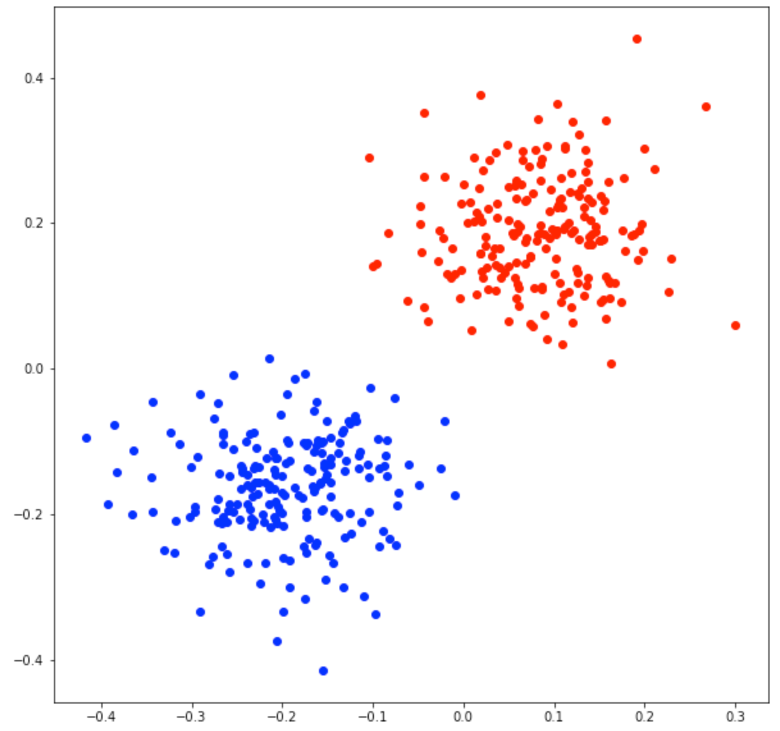
$$A \sim SBM(p, q)$$

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

$A =$ 

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

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



Results (informal)

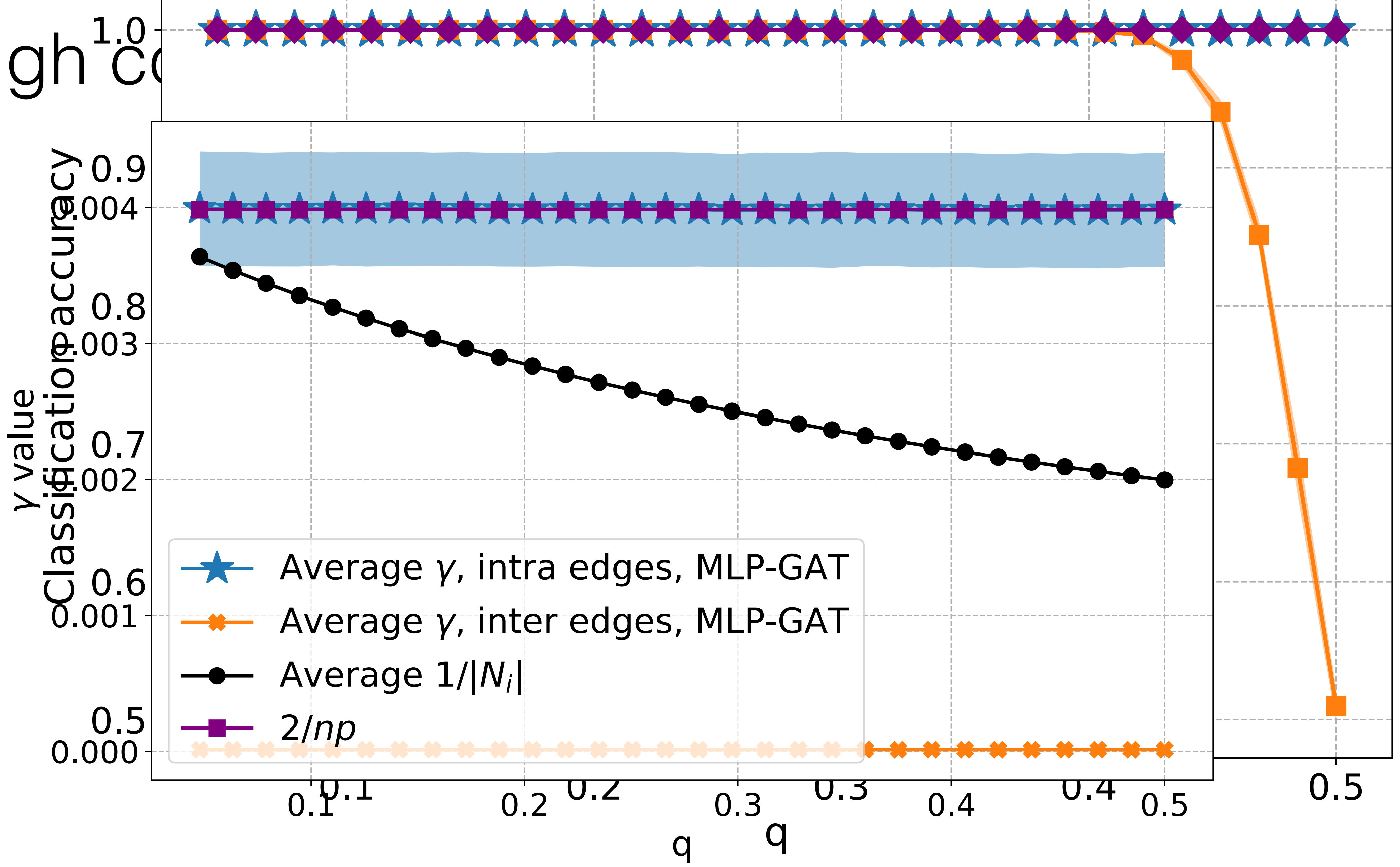
Hard regime $\|\mu\| \le K\sigma$

- 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- or intra-edge is not down-weighted

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

- MLP (no graph) achieves perfect classification
- GAT: distinguishes intra- from inter-edges with high c

Distance between means $\|\mu\|$



Results (informal)

Hard regime

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

Easy regime

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

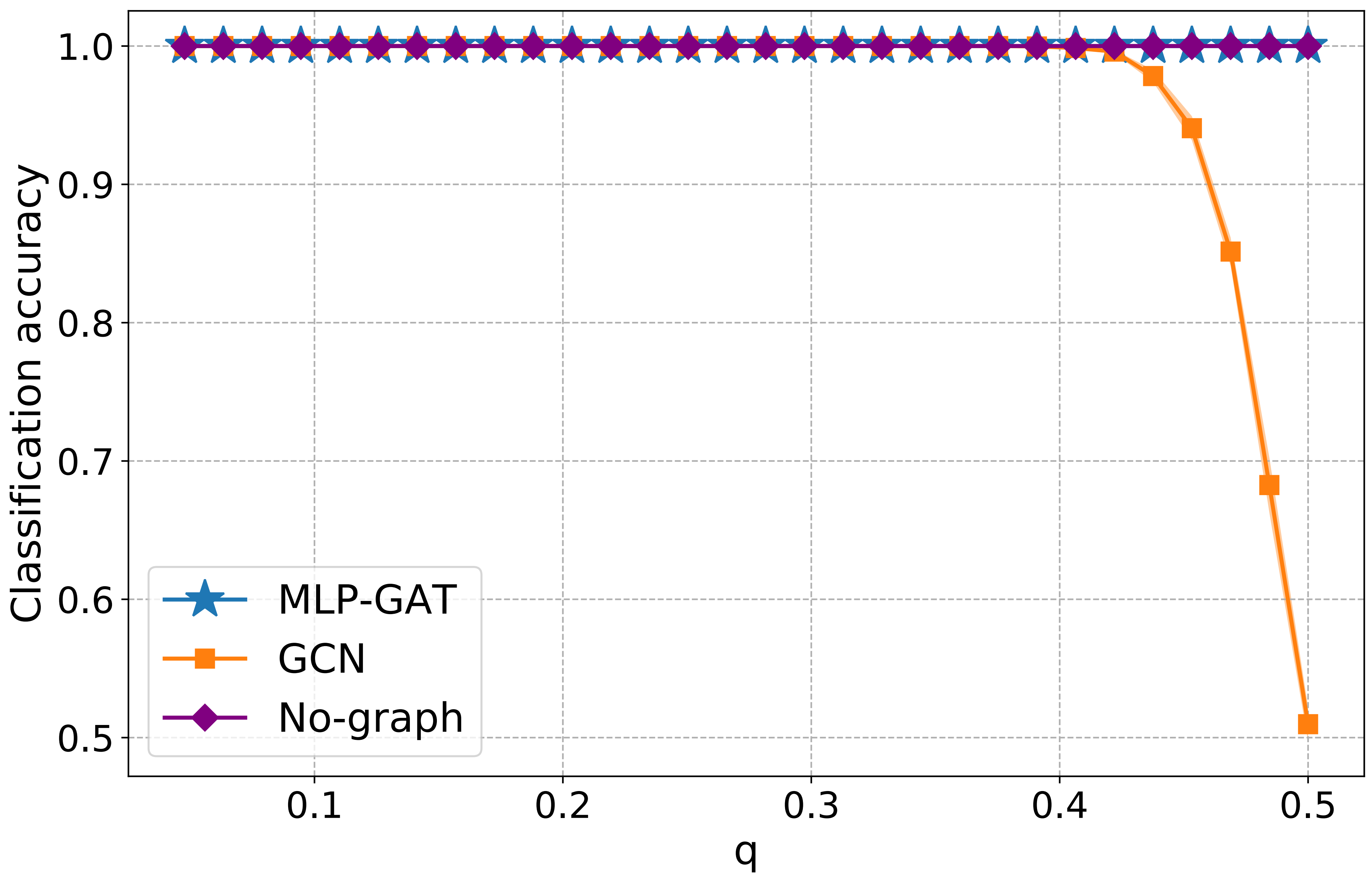
K const.

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: *perfect node* classification is possible, but it depends on p, q
- Conjecture: dependence on p, q is similar to GCN. Graph attention isn't better than GCN (more on this in subsequent slides).



Results (informal)

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

K const.

- MLP: constant fraction of misclassified nodes

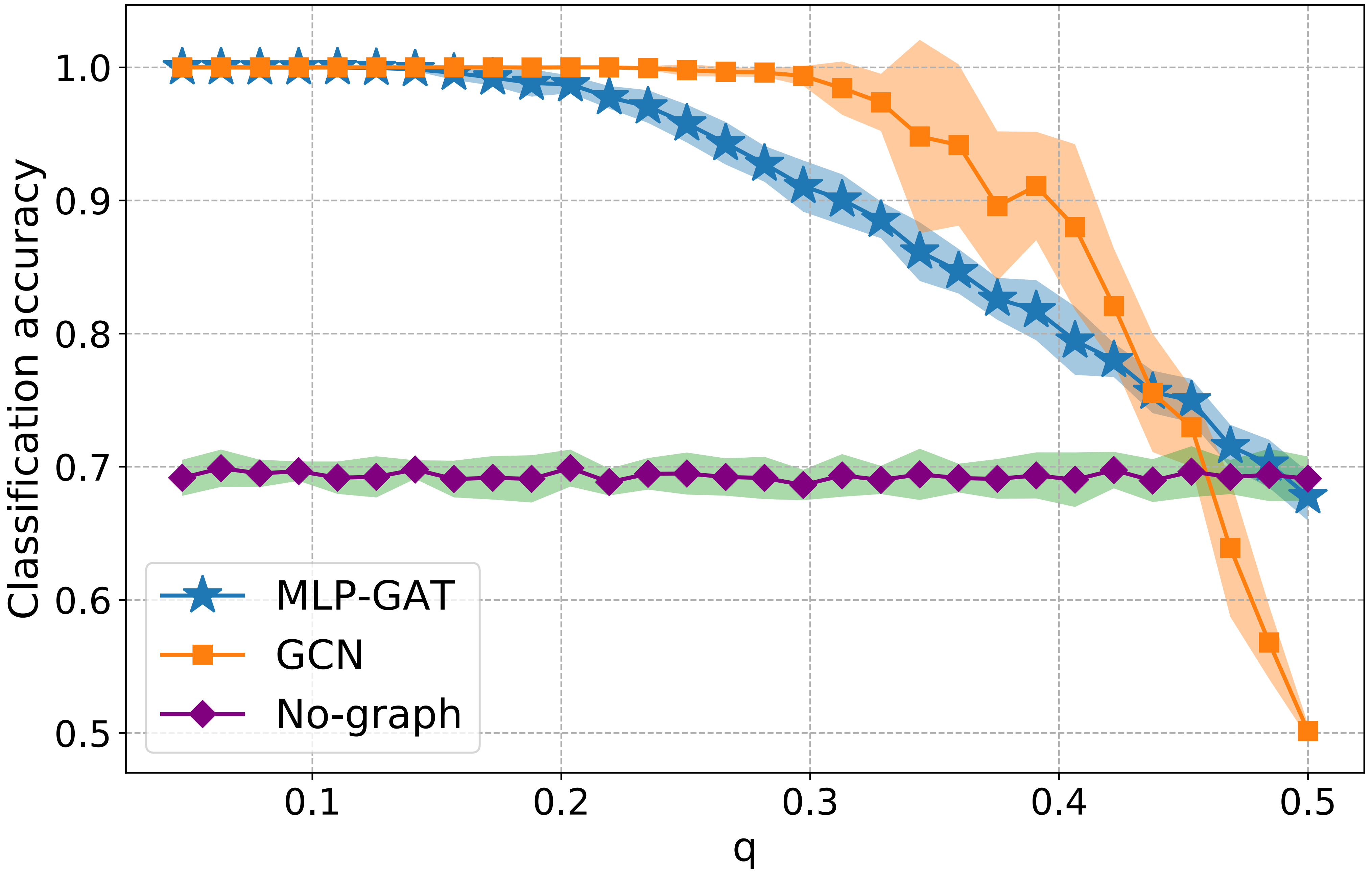
K non const.

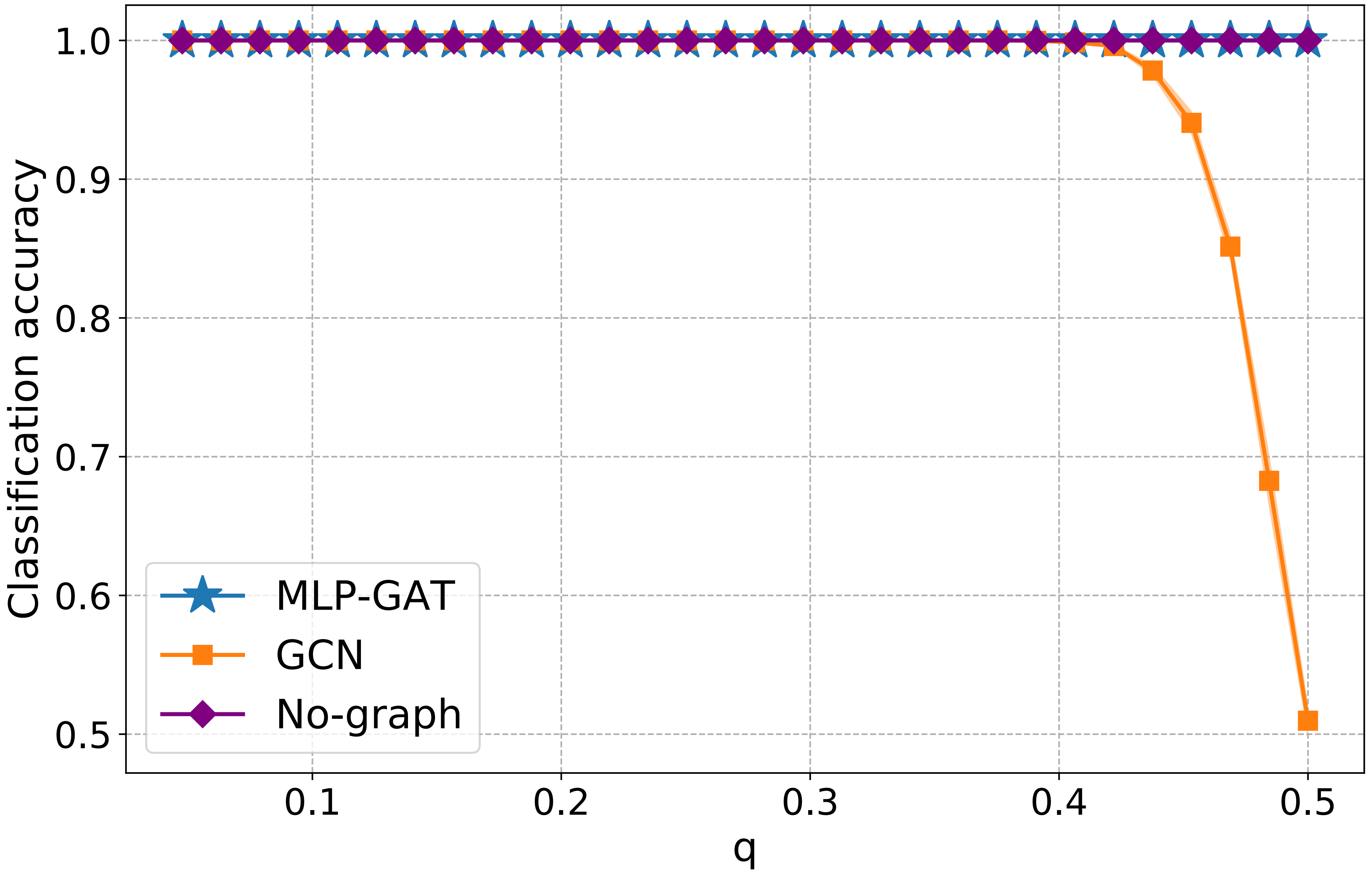
- MLP: at least one misclassified node

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

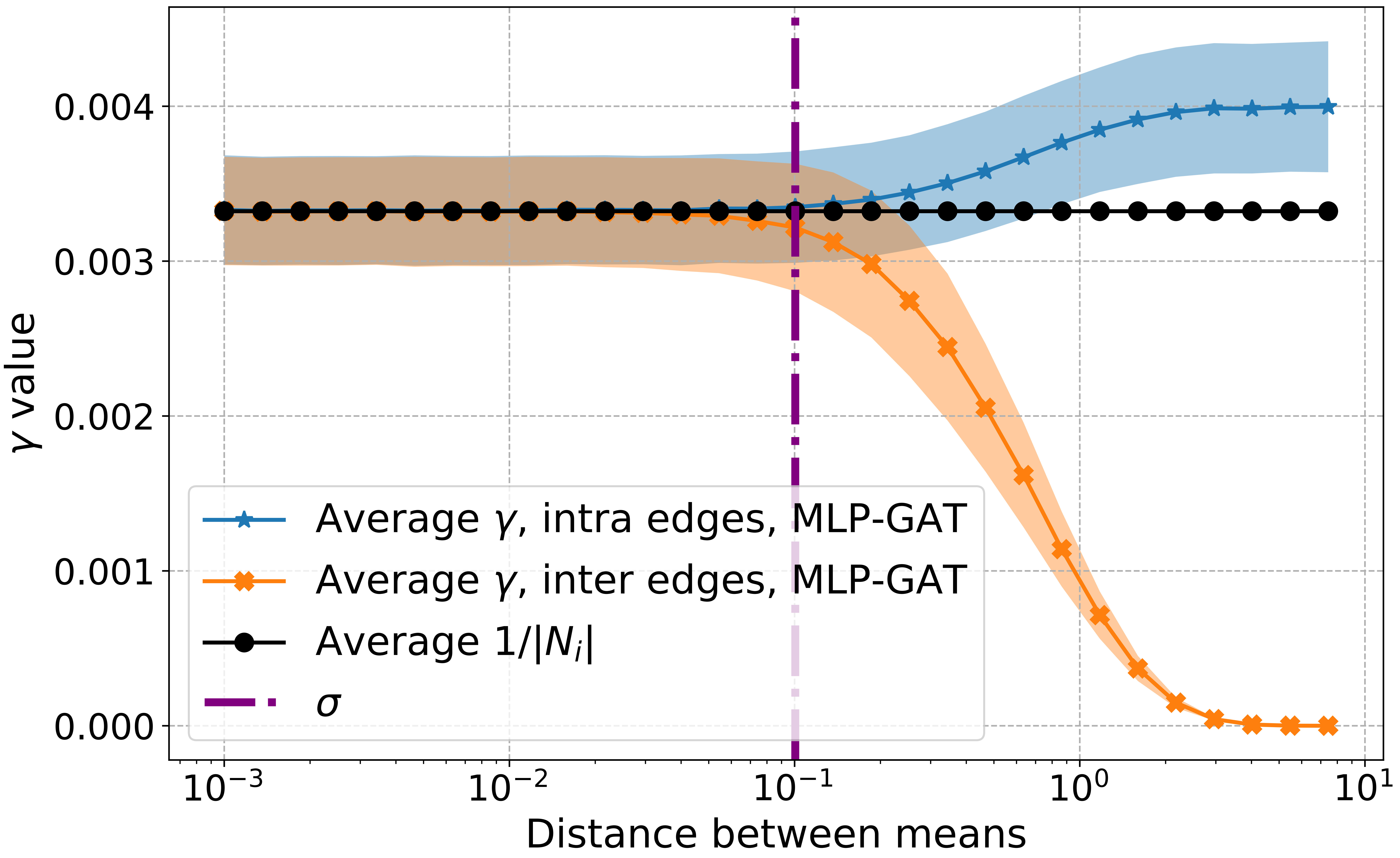
- MLP (no graph) achieves perfect classification

Distance between means
$\|\mu\|$

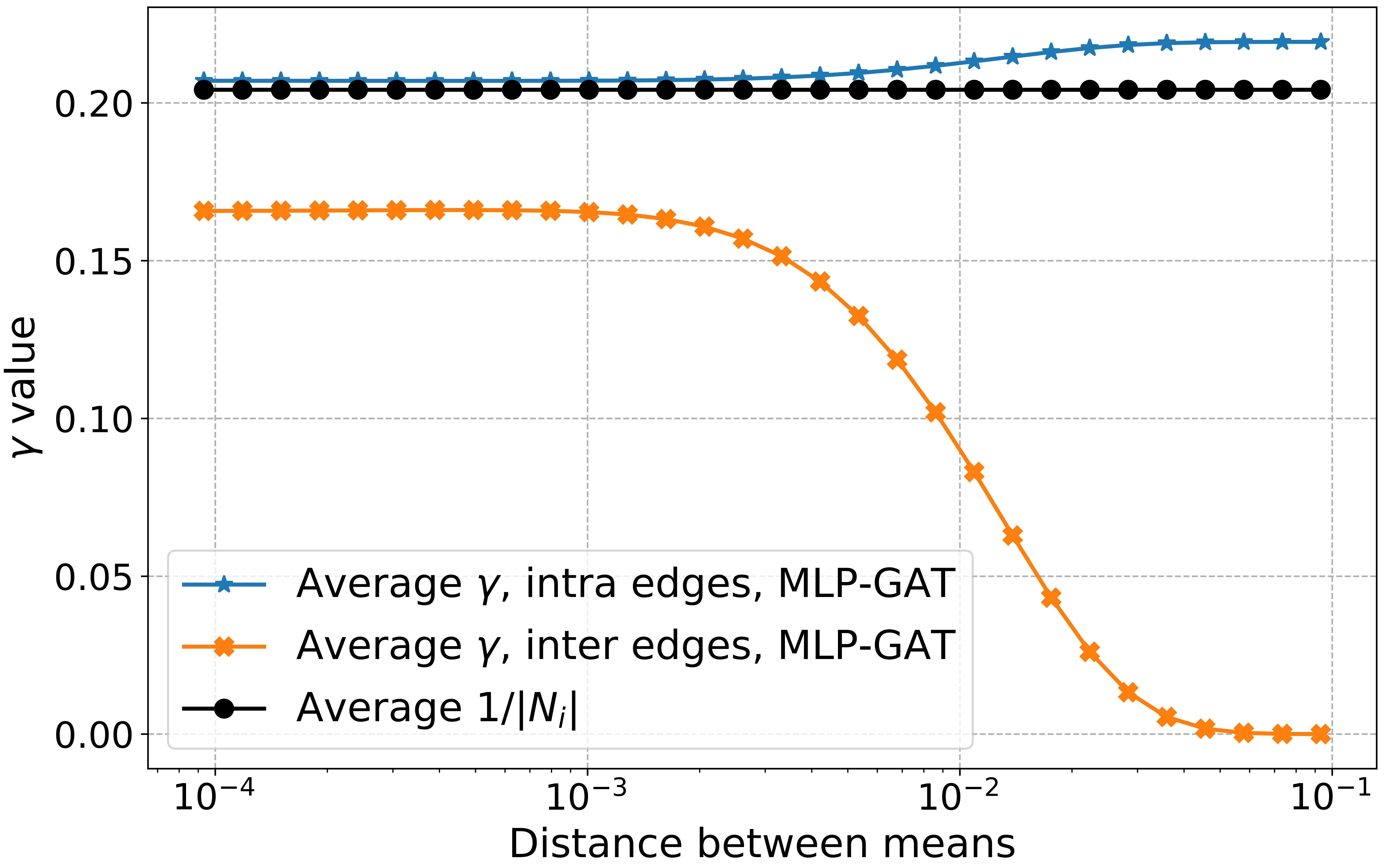




Empirical results (synthetic, fixed p and q, and $p \geq q$)



Empirical results (real)

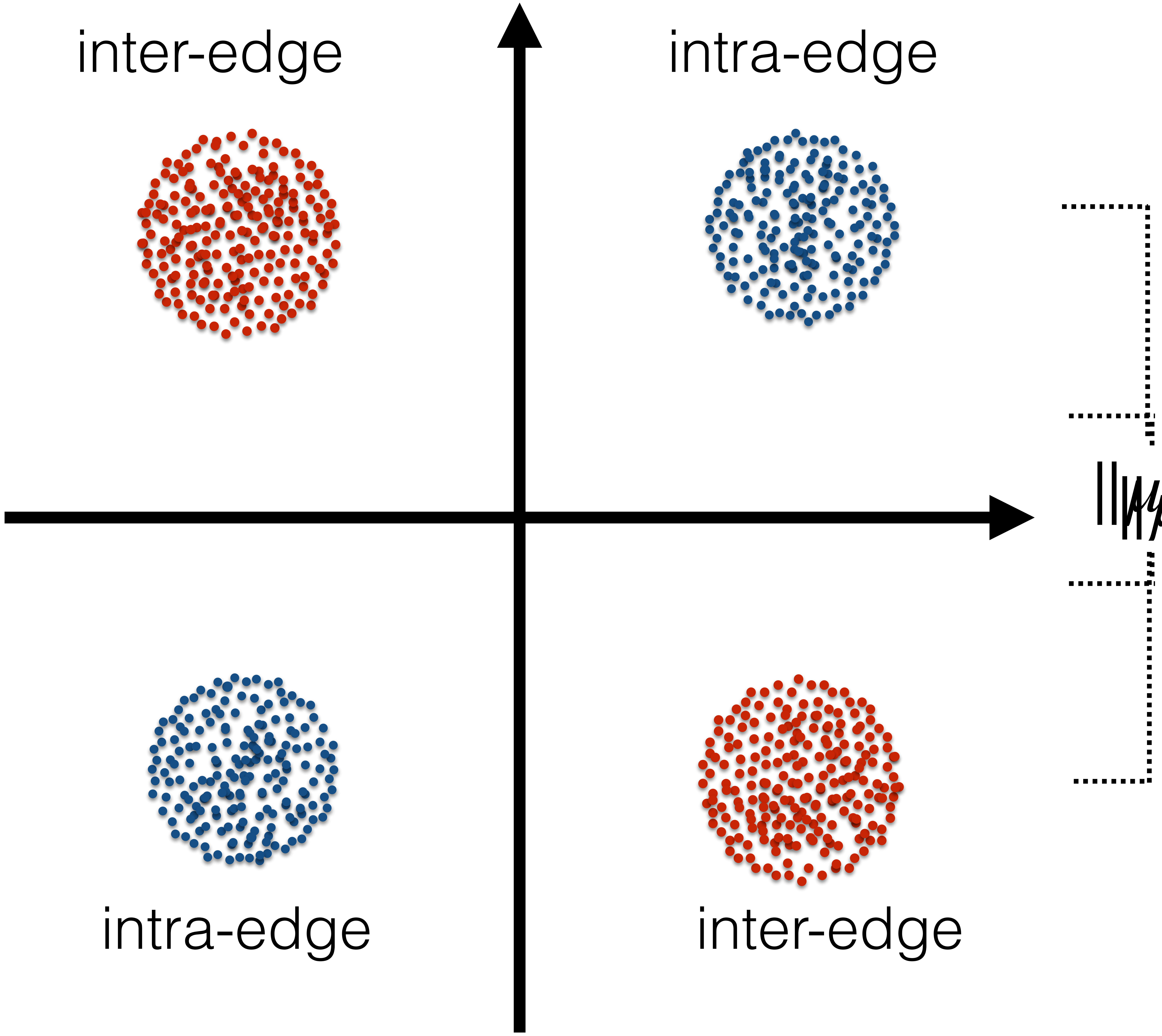


Why does graph attention fail to discriminate?

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

ψ is a soft-max function

Why does graph attention fail to discriminate?



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

$\|\mu\| \lesssim \sigma$

Conclusion

For our synthetic data model

- Attention is able to distinguish intra- from inter-edges. This results in perfect classification.
- Unfortunately, only when the graph is not needed to perfectly classify the nodes.
- This happens because the attention mechanism relies only on utilizing node features in attention.

For real data

- We demonstrate very similar observations on real data too.

Part 2: Details

Assumptions

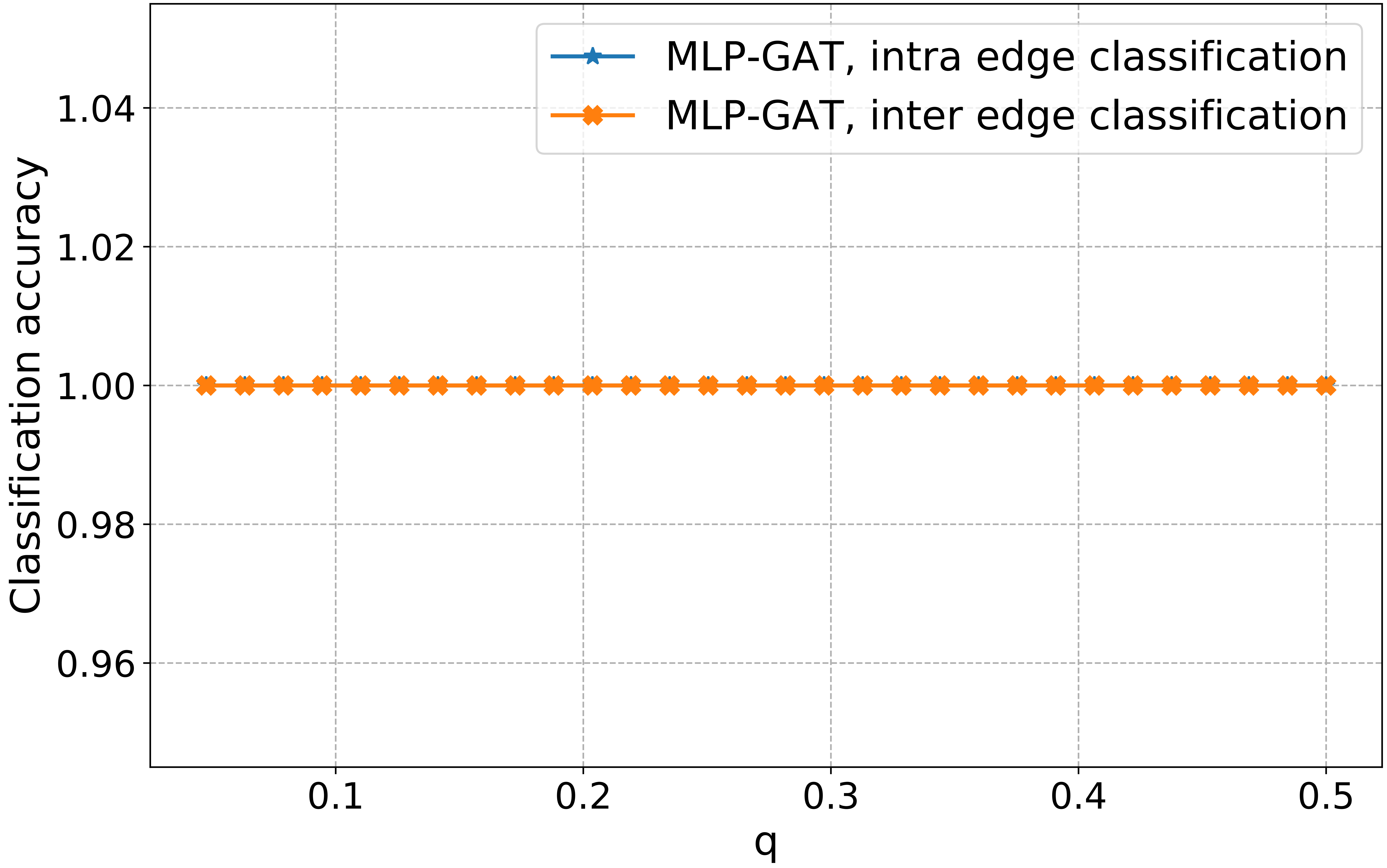
- Intra-edge probability $p = \Omega\left(\dfrac{\log^2 n}{n}\right)$
- Inter-edge probability $q = \Omega\left(\dfrac{\log^2 n}{n}\right)$
- Thus, the expected number of neighbours is $\Omega(\log^2 n)$ and we have degree concentration.

Super sparse cases where $p, q = a/n, b/n$, where a, b are constants aren't studied in this work.
We work on this direction currently.

Result 1: Classification of edges, easy regime

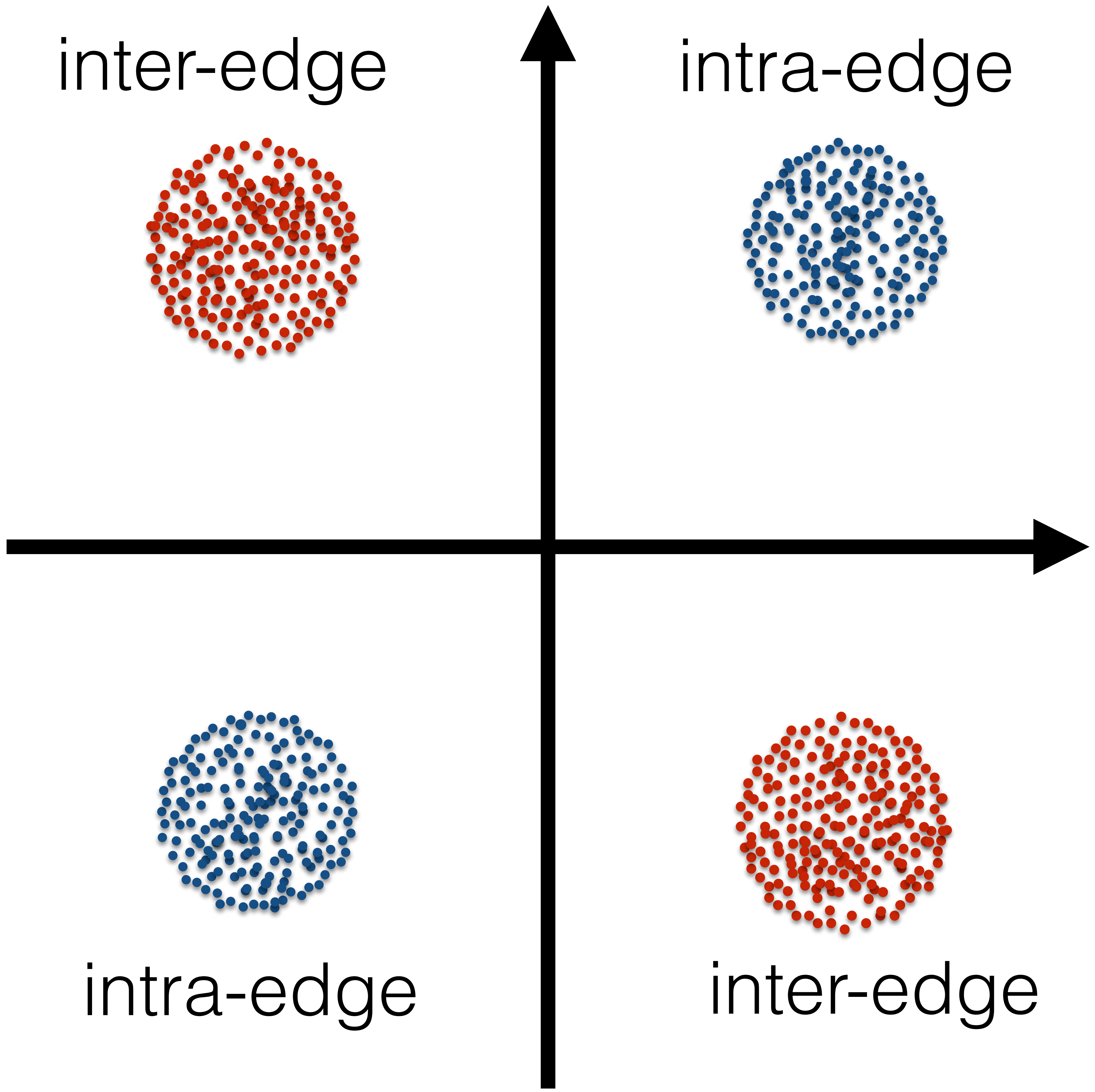
Theorem. Suppose that $\|\mu\|_2 = \omega(\sigma\sqrt{\log n})$. Then, there exists a choice of attention architecture Ψ such that with probability at least $1 - o_n(1)$ over the data $(X, A) \sim CSBM(n, p, q, \mu, \sigma^2)$ it holds that Ψ separates intra- from inter-edges.

Result 1: Classification of edges, easy regime ($p \geq q, p = 0.5$)



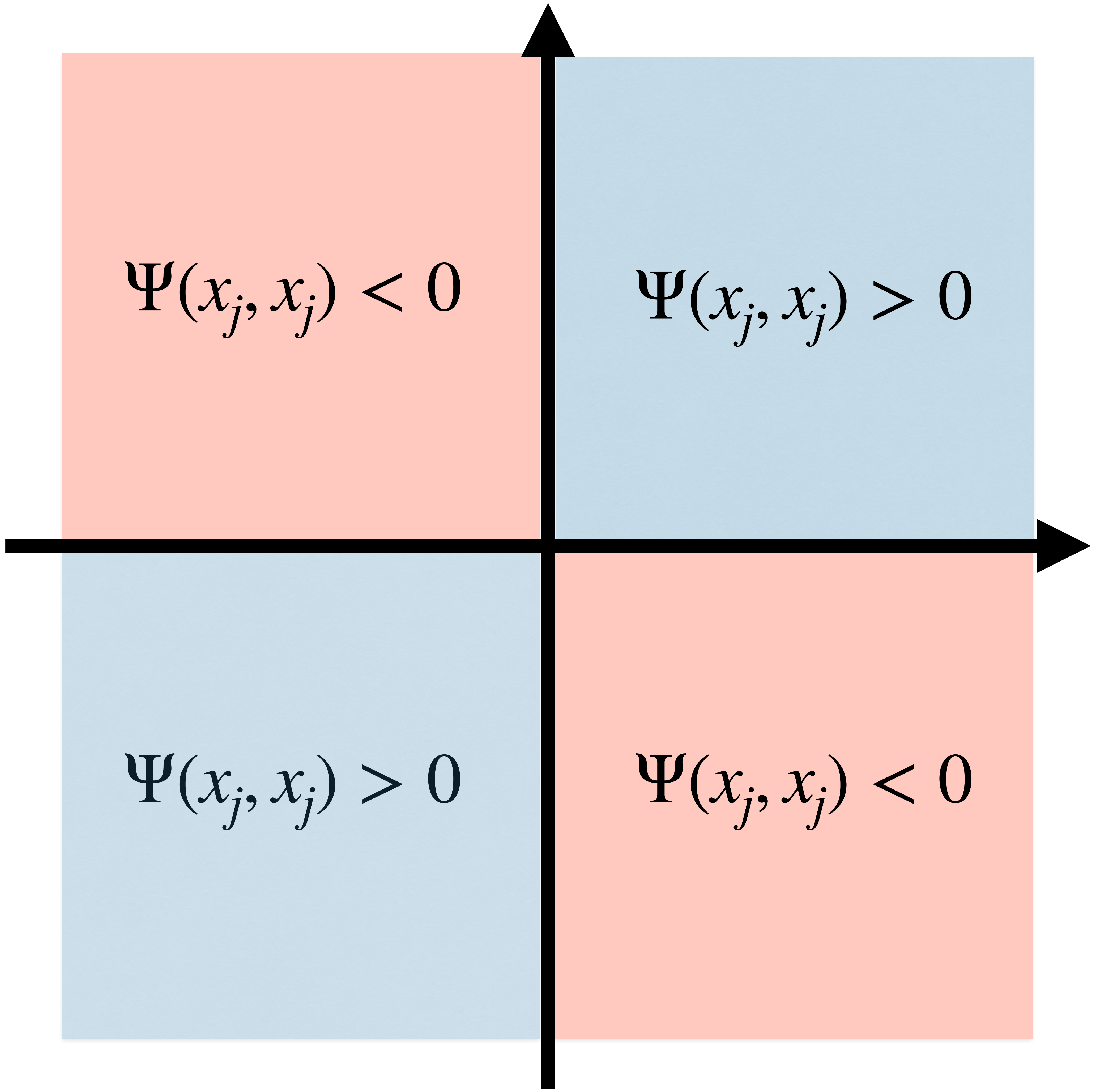
Proof sketch ($p \geq q$)

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

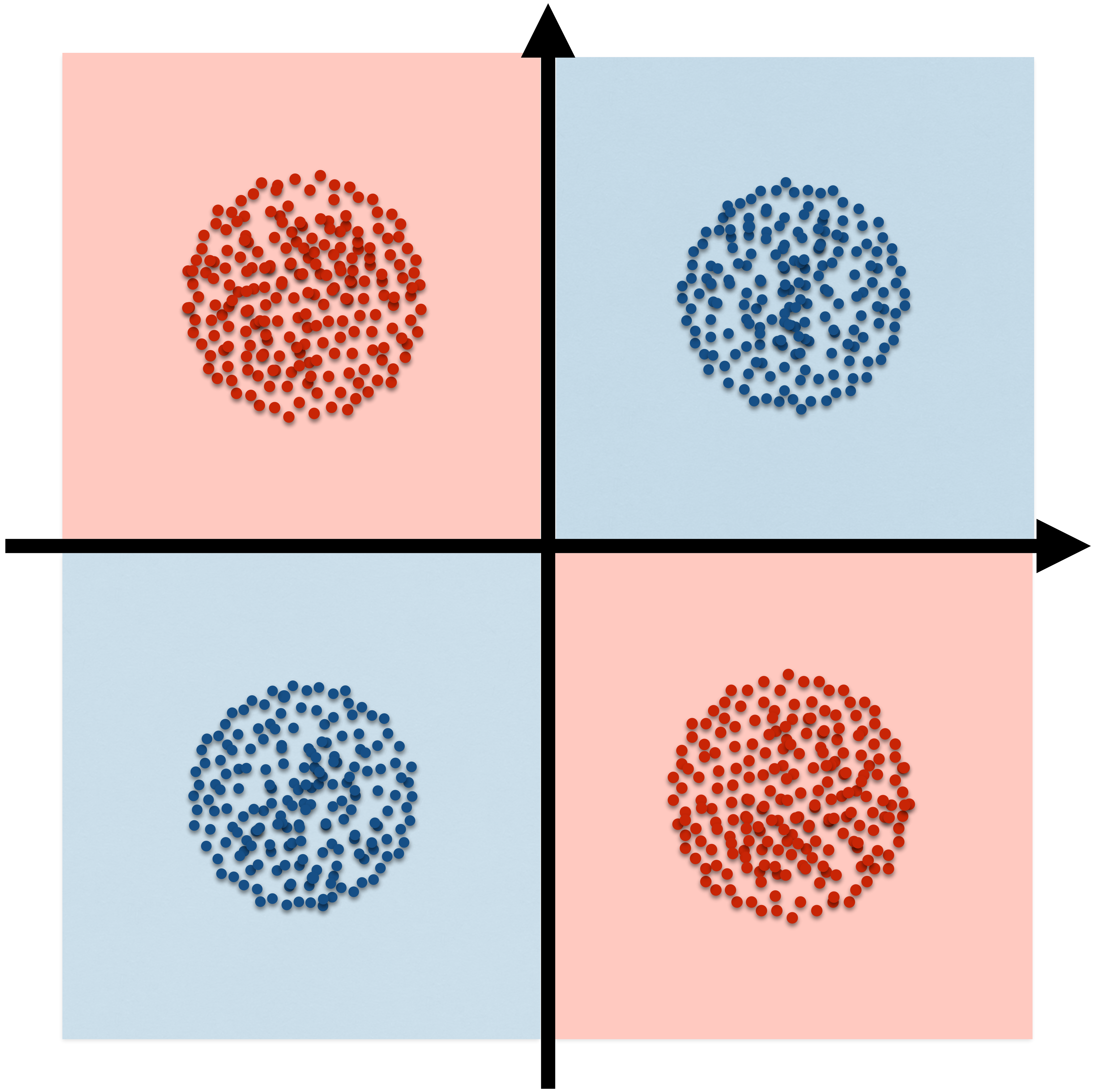


Proof sketch ($p \geq q$)

- Goal: construct a Ψ with the following classification regions



Proof sketch ($p \geq q$)



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

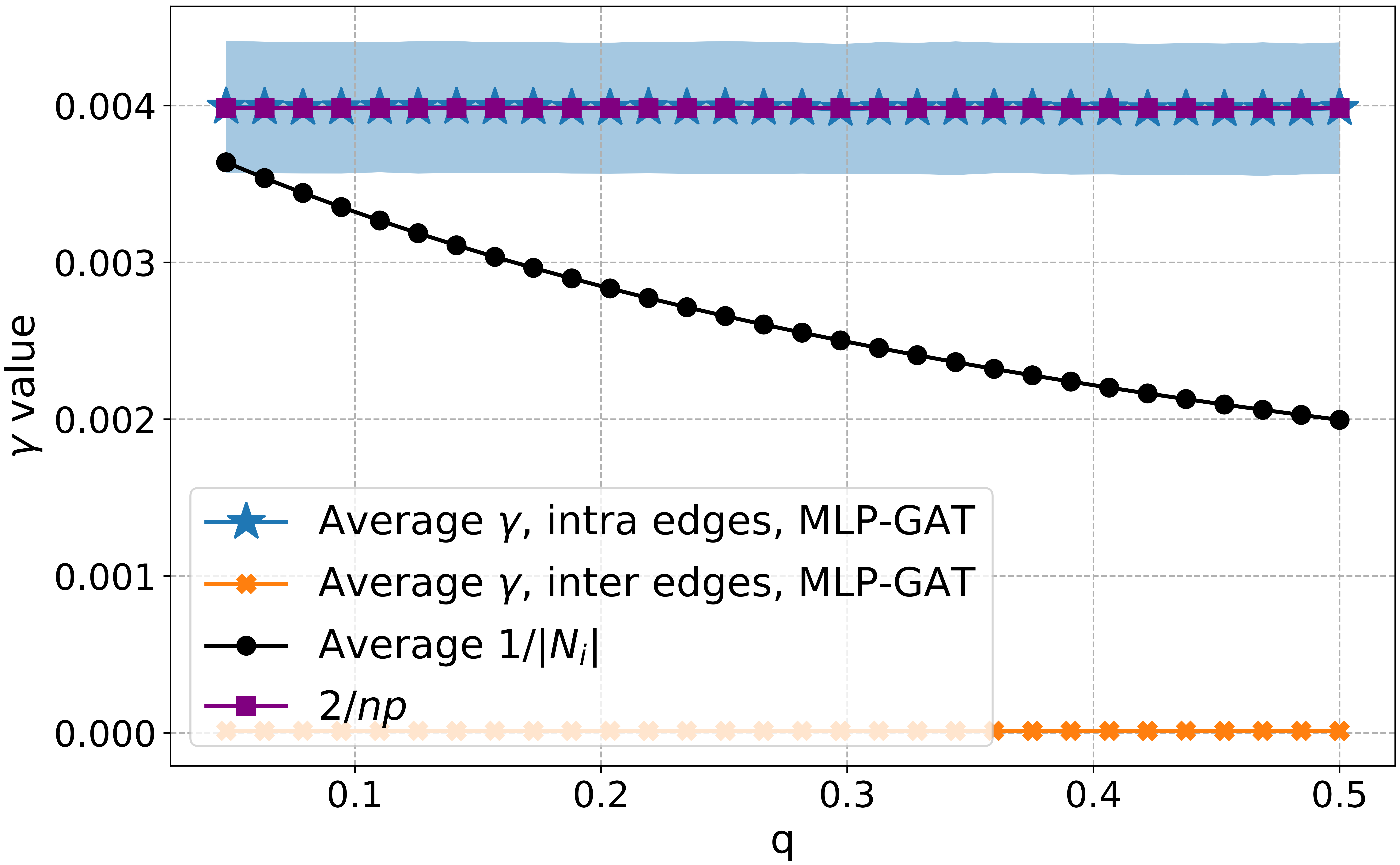
$$w = \frac{\mu}{\|\mu\|_2}$$

Result 2: Attention coefficients, easy regime

Corollary. Suppose that $\|\mu\|_2 = \omega(\sigma\sqrt{\log n})$. Then with probability at least $1 - o_n(1)$ over the data $(X, A) \sim CSBM(n, p, q, \mu, \sigma^2)$, a two-layer MLP attention architecture Ψ gives attention coefficients such that:

1. If $p \geq q$, then $\gamma_{ij} = \dfrac{2}{np}(1 \pm o_n(1))$ if (i, j) is an intra-edge and $\gamma_{ij} = o\left(\dfrac{1}{n(p+q)}\right)$ otherwise
2. If $p < q$, then $\gamma_{ij} = \dfrac{2}{np}(1 \pm o_n(1))$ if (i, j) is an inter-edge and $\gamma_{ij} = o\left(\dfrac{1}{n(p+q)}\right)$ otherwise

Result 2: Attention coefficients, easy regime ($p \geq q, p = 0.5$)



Proof sketch ($p \geq q$)

- 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 ($p \geq q$)

- Example of an intra-class edge

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

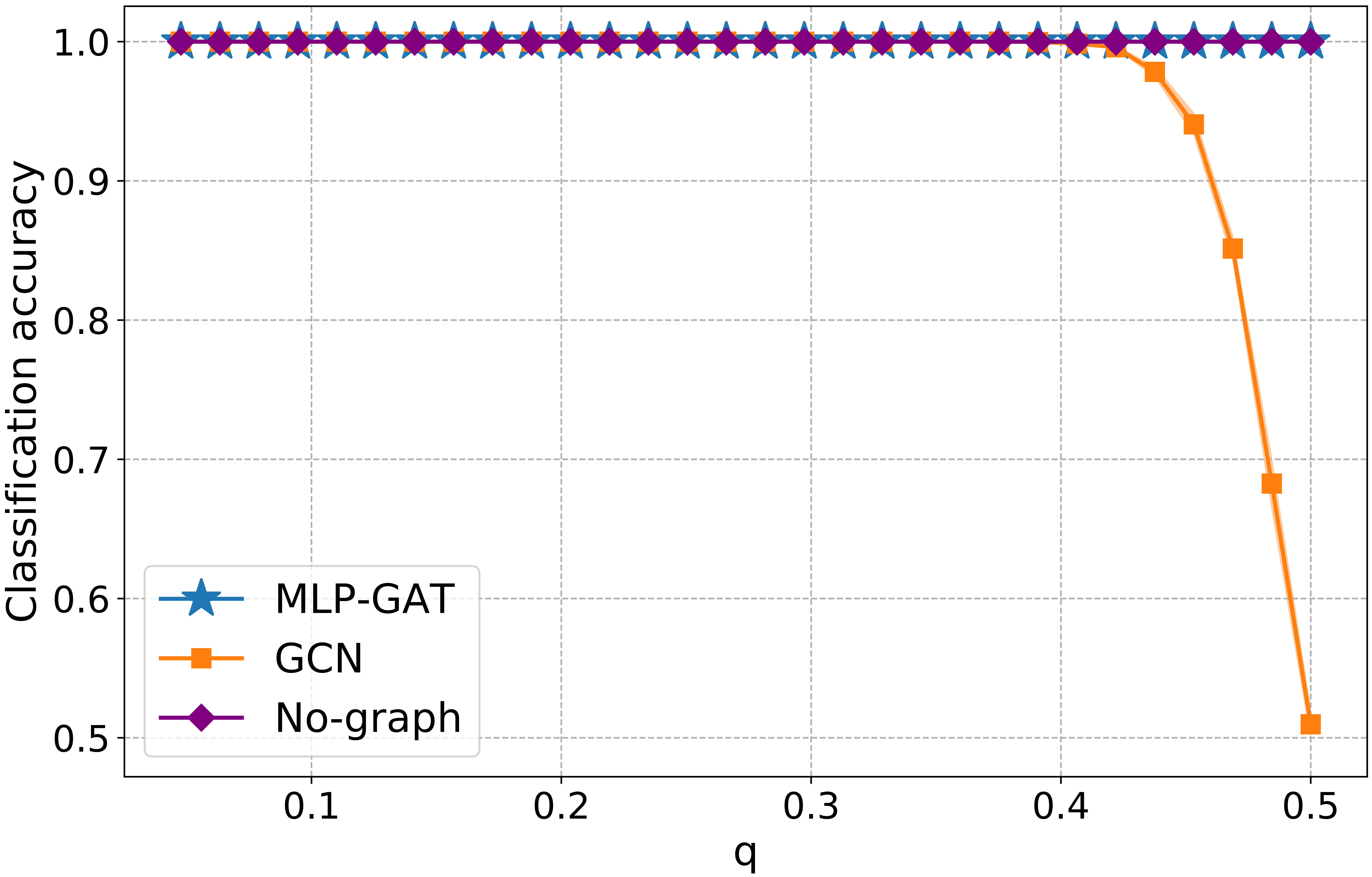
- Example of an inter-class edge

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

Result 3: node classification, easy regime

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

Result 3: node classification, easy regime ($p \geq q, p = 0.5$)



Proof sketch ($p \geq q$)

- 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

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

≈ 0

Proof sketch ($p \geq q$)

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

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

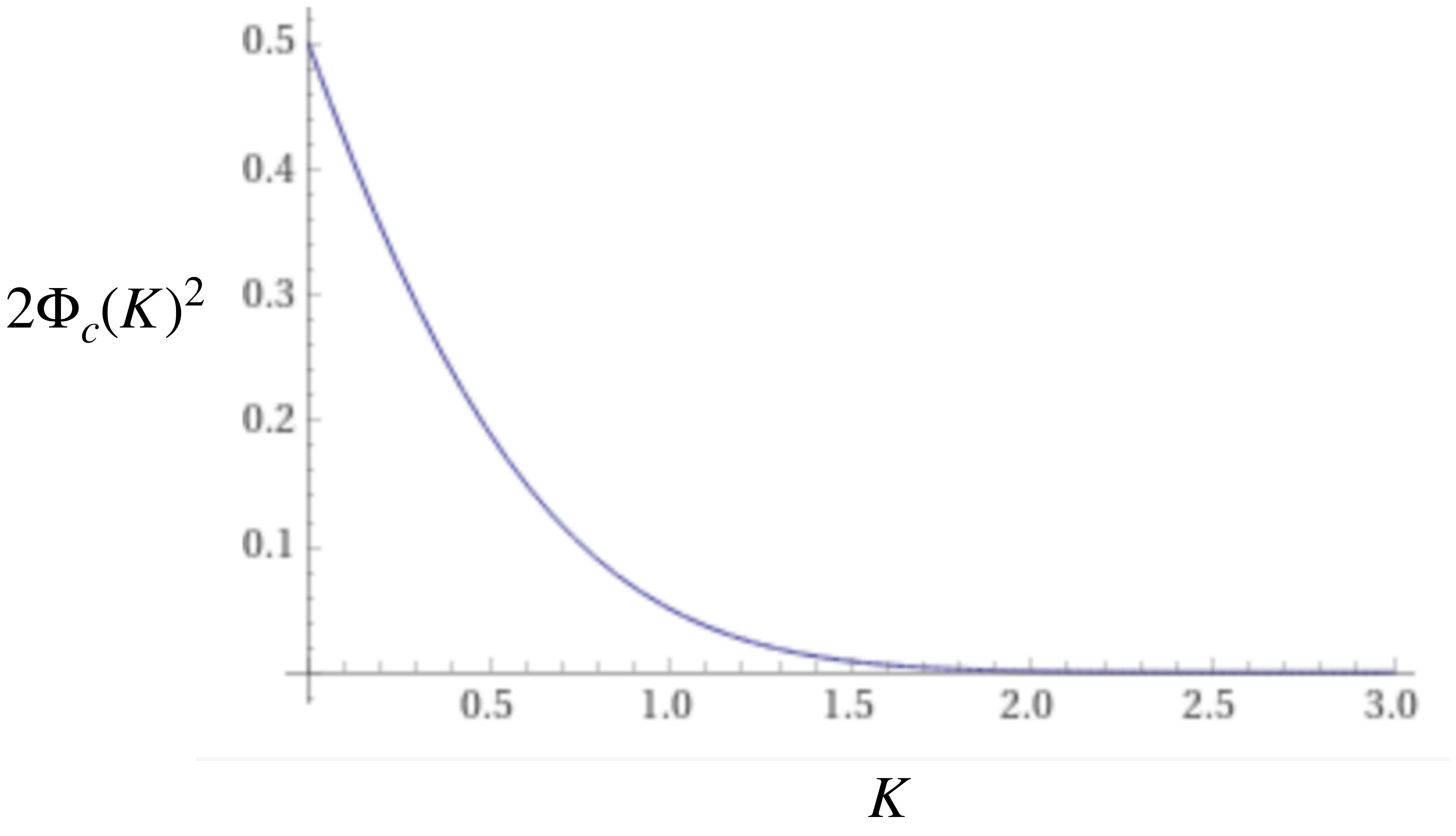
- While the distance between the means is much larger

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

- And this implies perfect node classification with high probability

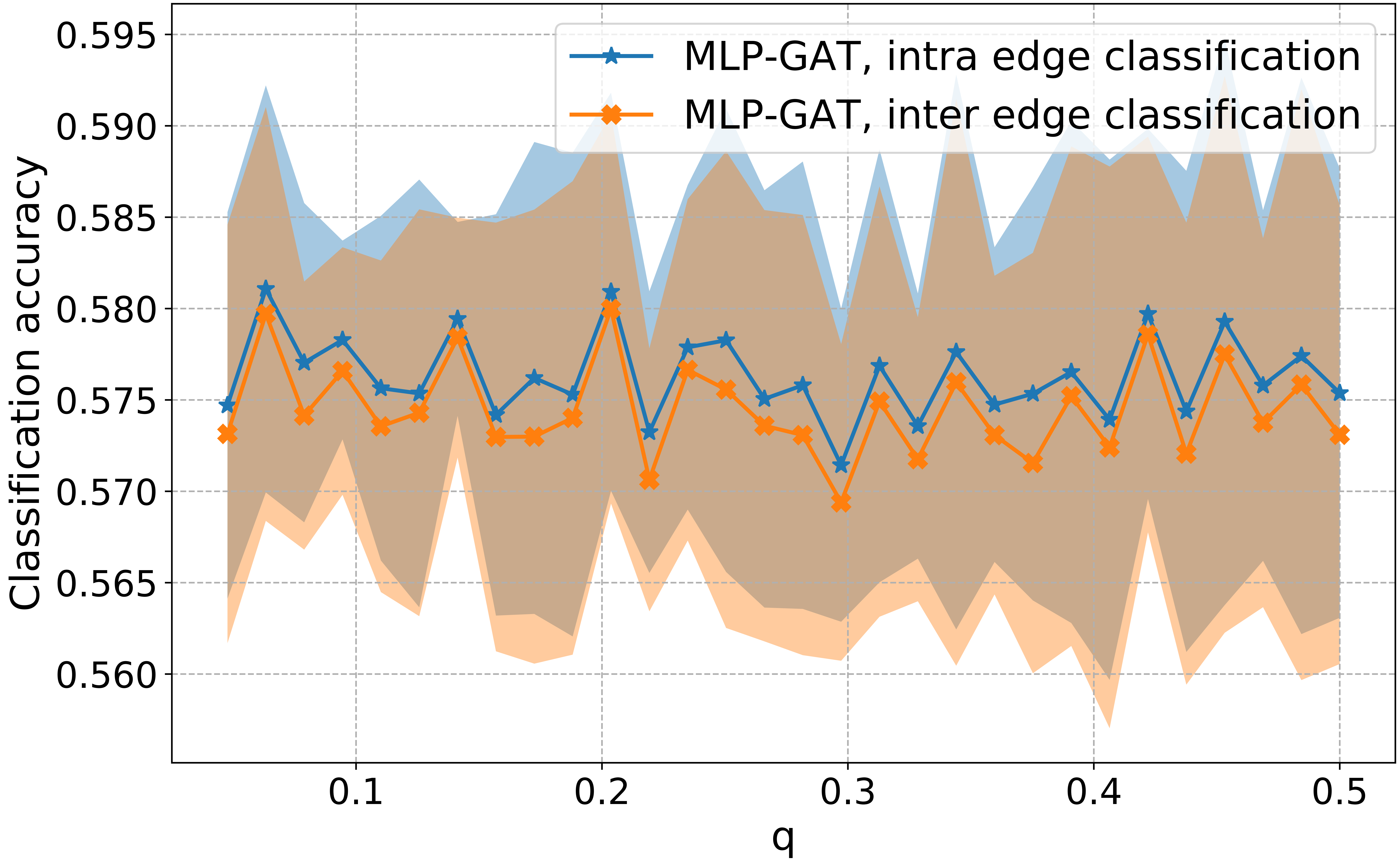
Result 4: classification of edges, hard regime

Corollary. Suppose that $||\mu||_2 = K\sigma$ for some $K > 0$ and let Ψ any attention mechanism on concatenated pairs of node features. Then, Ψ fails to correctly classify at least a $2\Phi_c(K)^2$ fraction of intra- and inter-edges with probability $1 - O(n^{-c})$ for any $c > 0$.



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

Result 4: classification of edges, hard regime

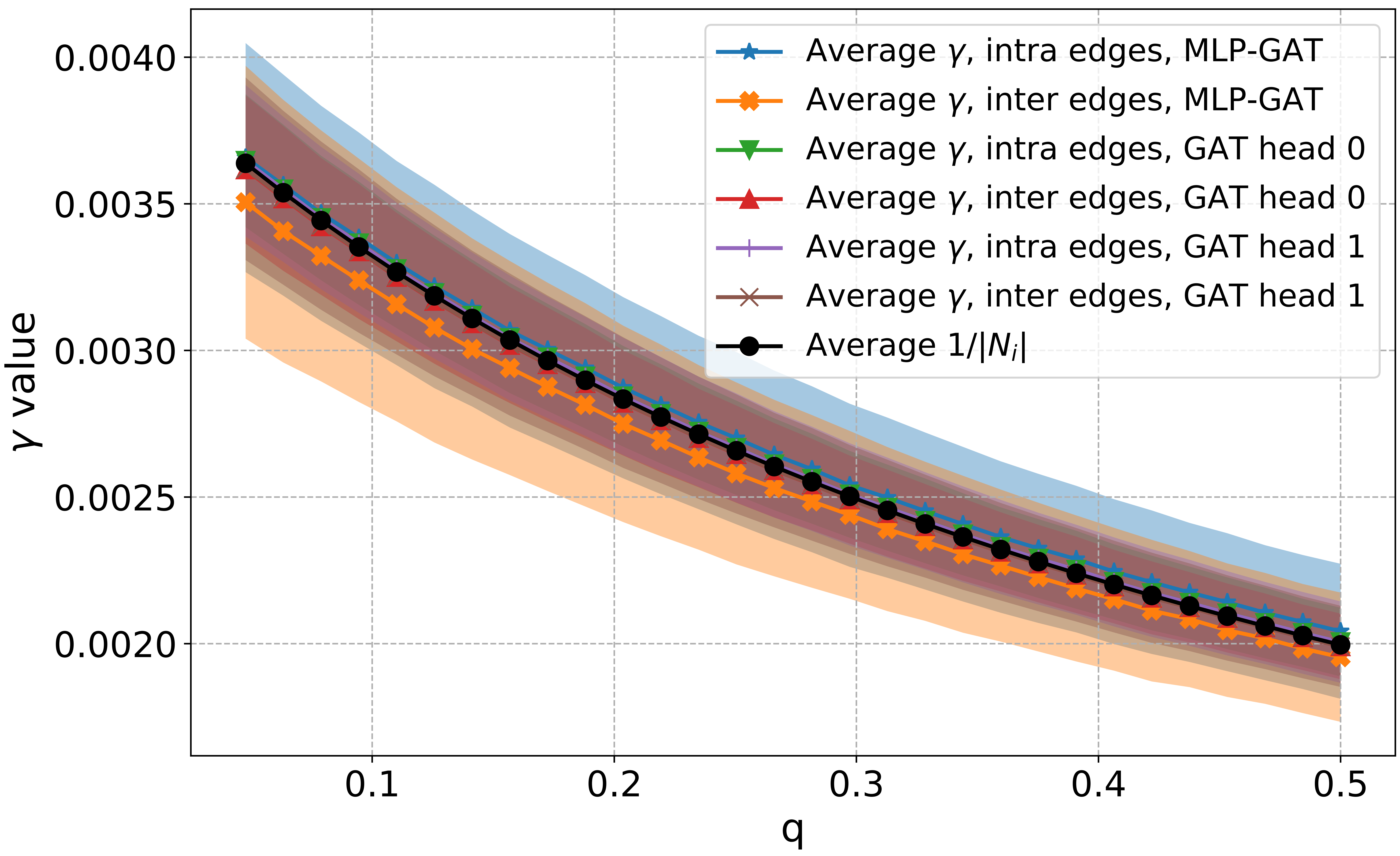


Result 4: Attention coeff. for a popular GAT model, hard regime

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

Theorem. Suppose that $\|\mu\|_2 \leq K\sigma$ and $\sigma \leq K'$ for some constants K and K'. Moreover, assume that the learnable parameters are bounded by a constant. Then, with probability at least $1 - o_n(1)$ over the data $(X, A) \sim CSBM(n, p, q, \mu, \sigma^2)$, at least 90% of intra- and inter-edge attention coefficients are $\gamma_{ij} = \Theta(1/|N_i|)$.

Result 4: classification of edges, hard regime



Proof sketch

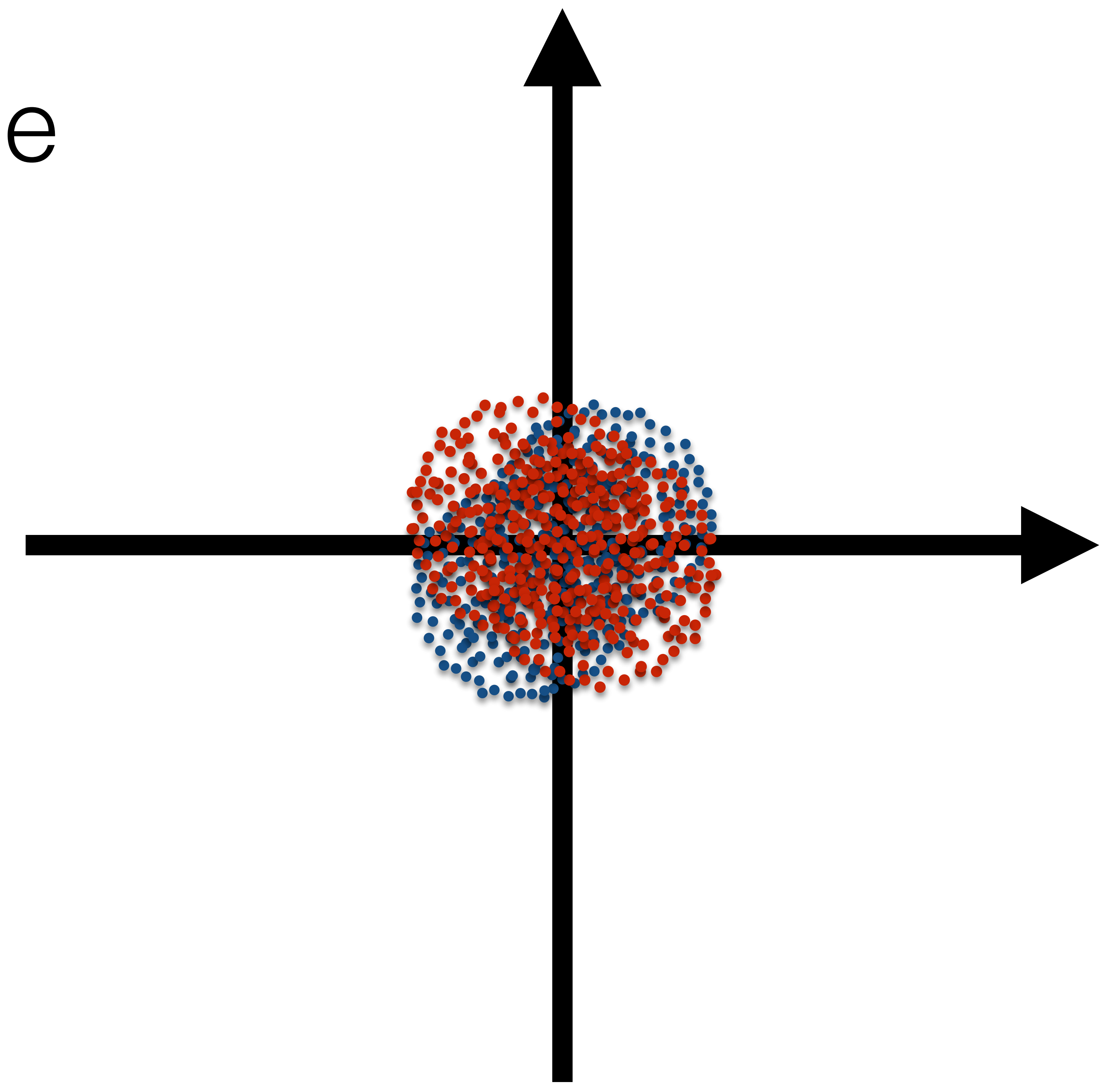
- The standard deviation is comparable to the distance between the means.

$+$

- Data act like Gaussian noise.

$=$

- Not all 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, using this in the definition of γ we get that most γ are $\Theta(1/|N_i|)$.

Conjecture

- We don't have a proof for node classification in the regime where $\|\mu\|_2 \leq K\sigma$, where $K \leq \mathcal{O}(\sqrt{logn})$.
- But we conjecture that graph attention doesn't have a better threshold than GCN in this regime.

For thresholds for GCN see:

1. A. Baranwal, K. Fountoulakis, A. Jagannath, Graph Convolution For Semi-Supervised Classification (ICML 2021)
2. K. Fountoulakis, D. He, S. Lattanzi, B. Perozzi, A. Tsitsulin, S. Yang, On Classification Thresholds for Graph Attention with Edge Features, arXiv:2210.10014

Difficulty in proving the conjecture

Graph attention convolution: $x_i' = \underbrace{\sum_{j\in[n]} A_{ij}\gamma_{ij}W\mu_j}_{convolved\ means:\ signal_i} + \sigma \underbrace{\sum_{j\in[n]} A_{ij}\gamma_{ij}Wz_j}_{conv.\ noise:\ noise_i}$

- We need to lower bound the expected maximum noise:

$$\mathbb{E}[max_{i\in C_0}\ noise_i]$$

- Seems like a classical Sudakov argument, but $noise_i$ is not Gaussian…

Beyond vanilla attention

- What if we set the attention mechanism Ψ using ground truth information?

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

- If $t = \mathcal{O}(1)$ the threshold is $\|\mu\|_2 \leq \sigma \frac{p+q}{|p-q|}\sqrt{\frac{\log n}{n(p+q)}}$ (our conjecture)

- If $t = \omega(1)$ the threshold is $\|\mu\|_2 \leq \sigma\sqrt{\frac{\log n}{n(p+q)}}$ (Better than our conjecture)

Can the “good” attention mechanism realized?

- Yes, use the eigenvectors of the adjacency matrix in attention function Ψ.
- Only works when $|\sqrt{p} - \sqrt{q}| > \sqrt{2 \log n / n}$.
- But… in this regime, one can simply achieve perfect classification using the eigenvector of the adjacency.

Additional edge features

- GAT can have better threshold than GCN
- But it requires additional clean edge features
- Which should allow us to show that GAT is better than classical methods, e.g., using eigenvectors of the adjacency/Laplacian matrices.
- See: K. Fountoulakis, D. He, S. Lattanzi, B. Perozzi, A. Tsitsulin, S. Yang, *On Classification Thresholds for Graph Attention with Edge Features*, arXiv:2210.10014 (Oct. 2022)

Thank you!