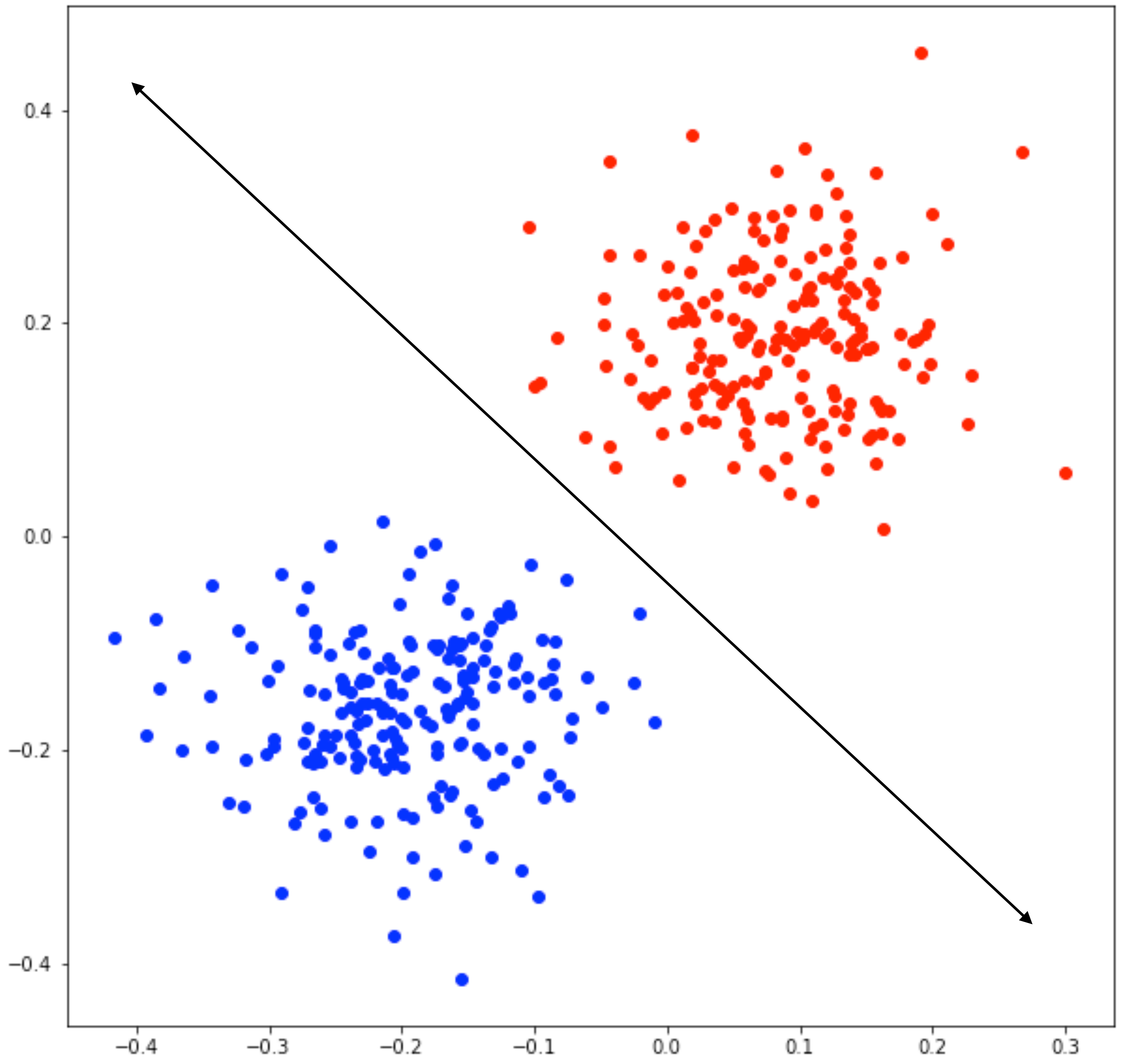
Graph Convolution for Semi-Supervised Classification: Improved Linear Separability and Out-of-Distribution Generalization

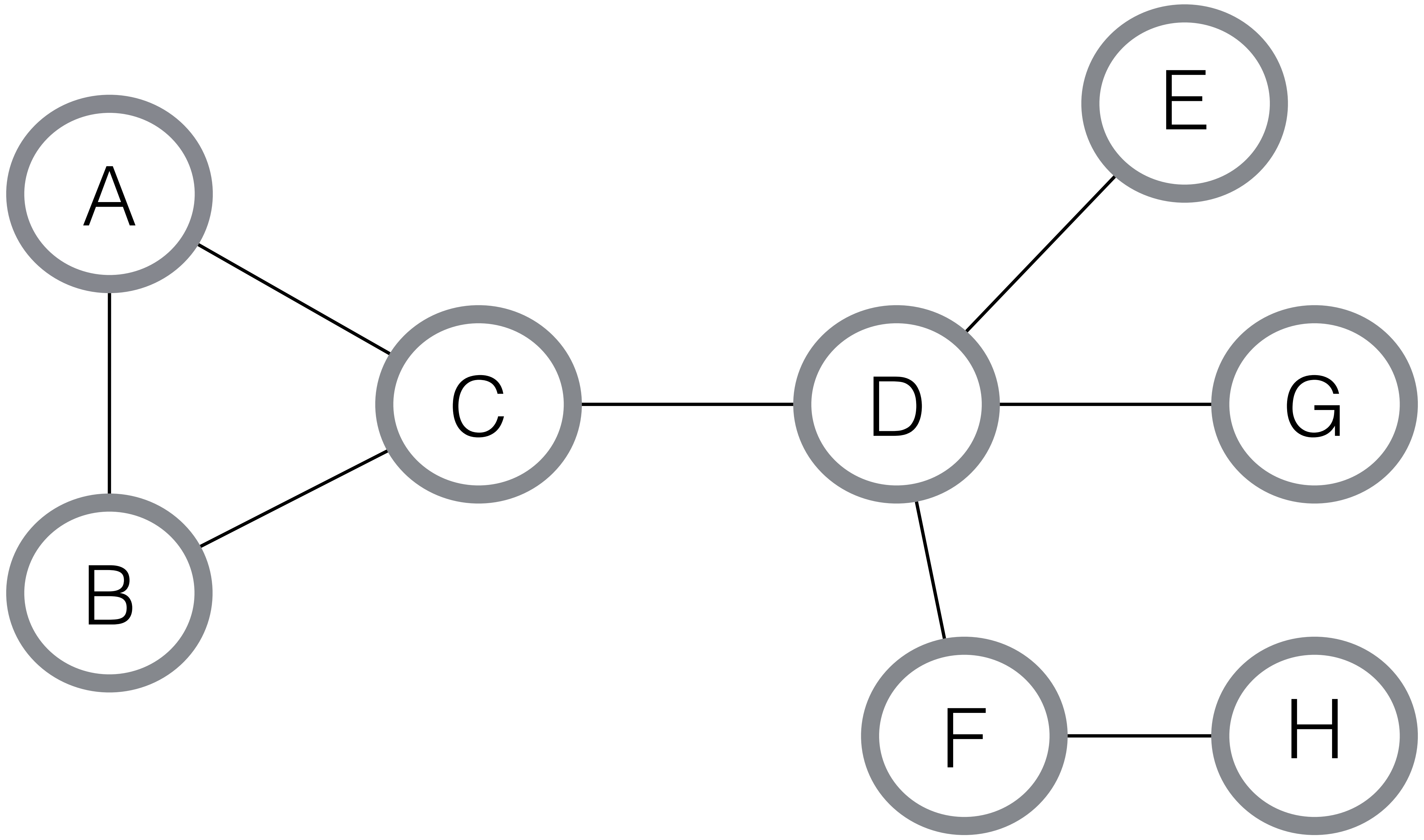
Aseem Baranwal, Kimon Fountoulakis, Aukosh Jagannath

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

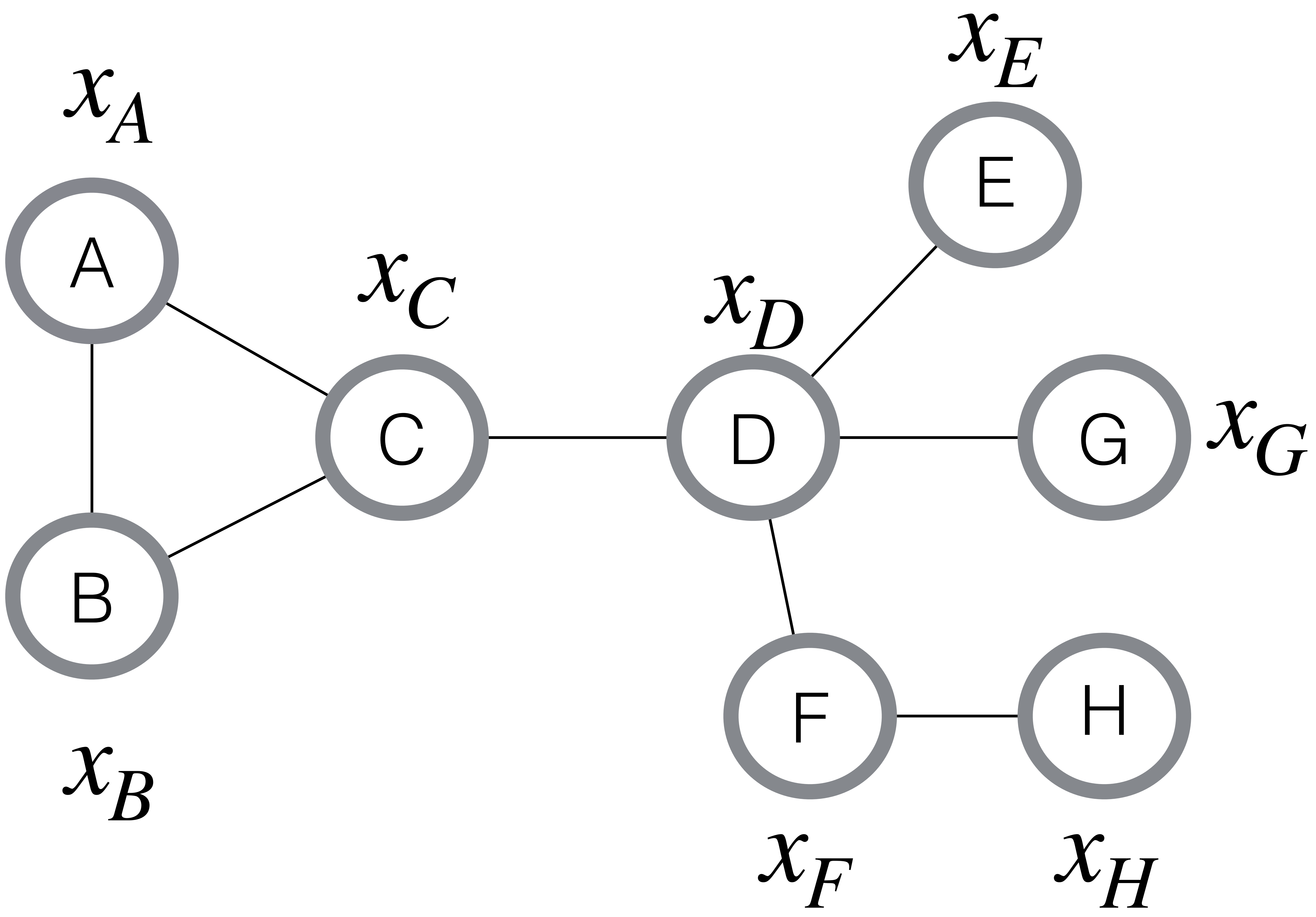
Linear classification



Spatial graph convolution of the data



Spatial graph convolution of the data

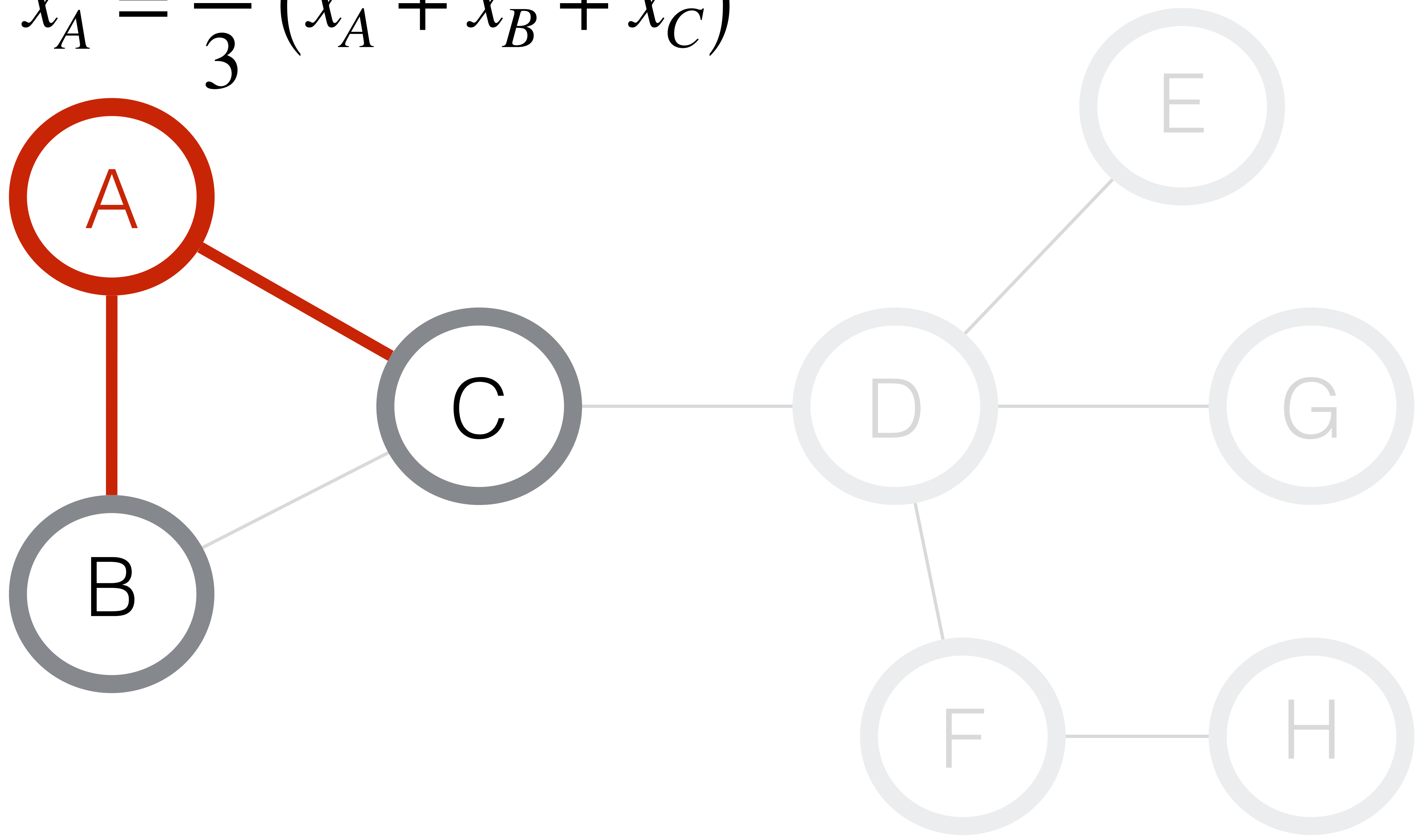


- x_i is the feature vector for node i

Spatial graph convolution of the data

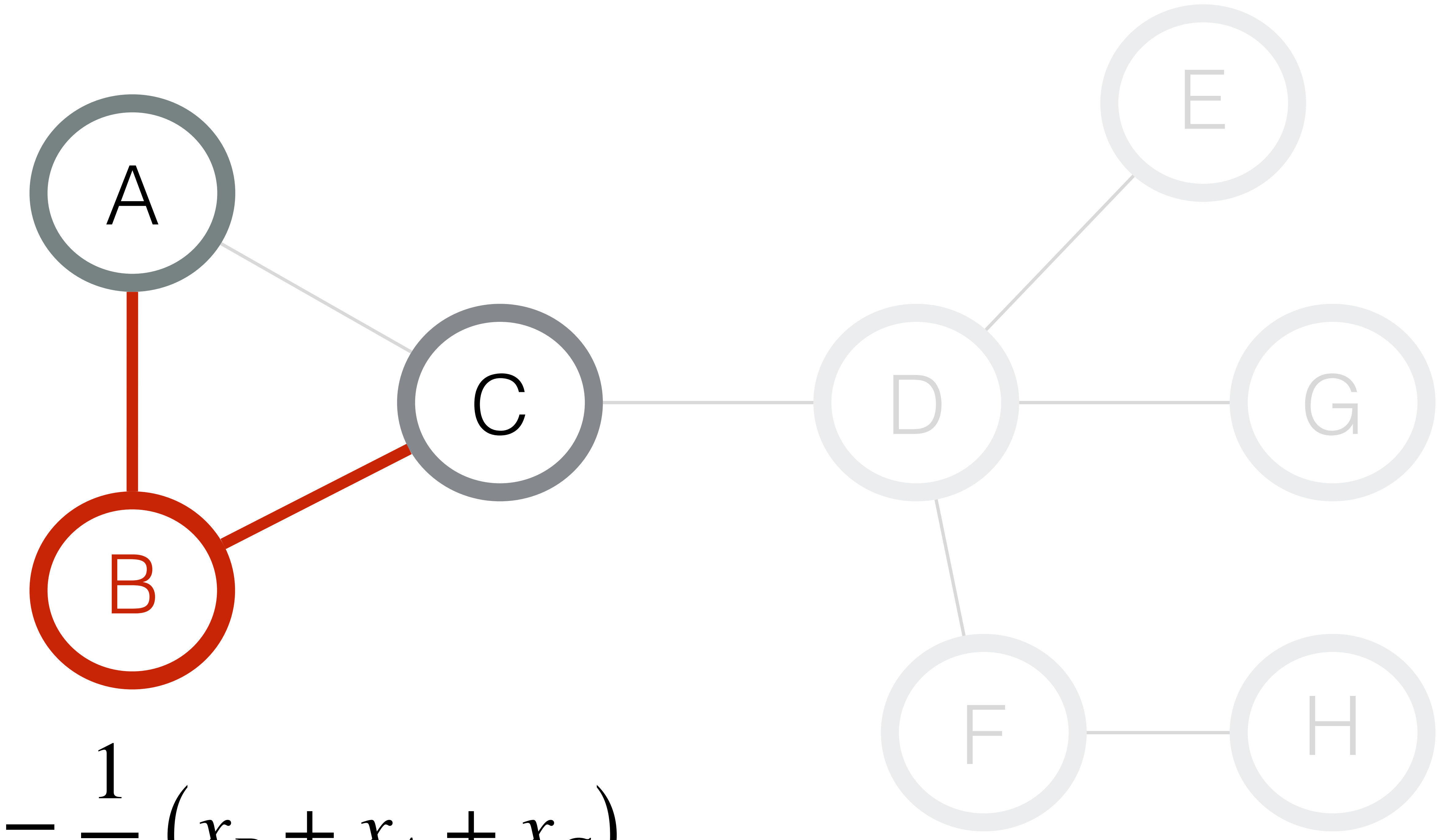
$$x'_A = \frac{1}{3}\left(x_A + x_B + x_C\right)$$

Node A gets messages
from nodes B and C



Spatial graph convolution of the data

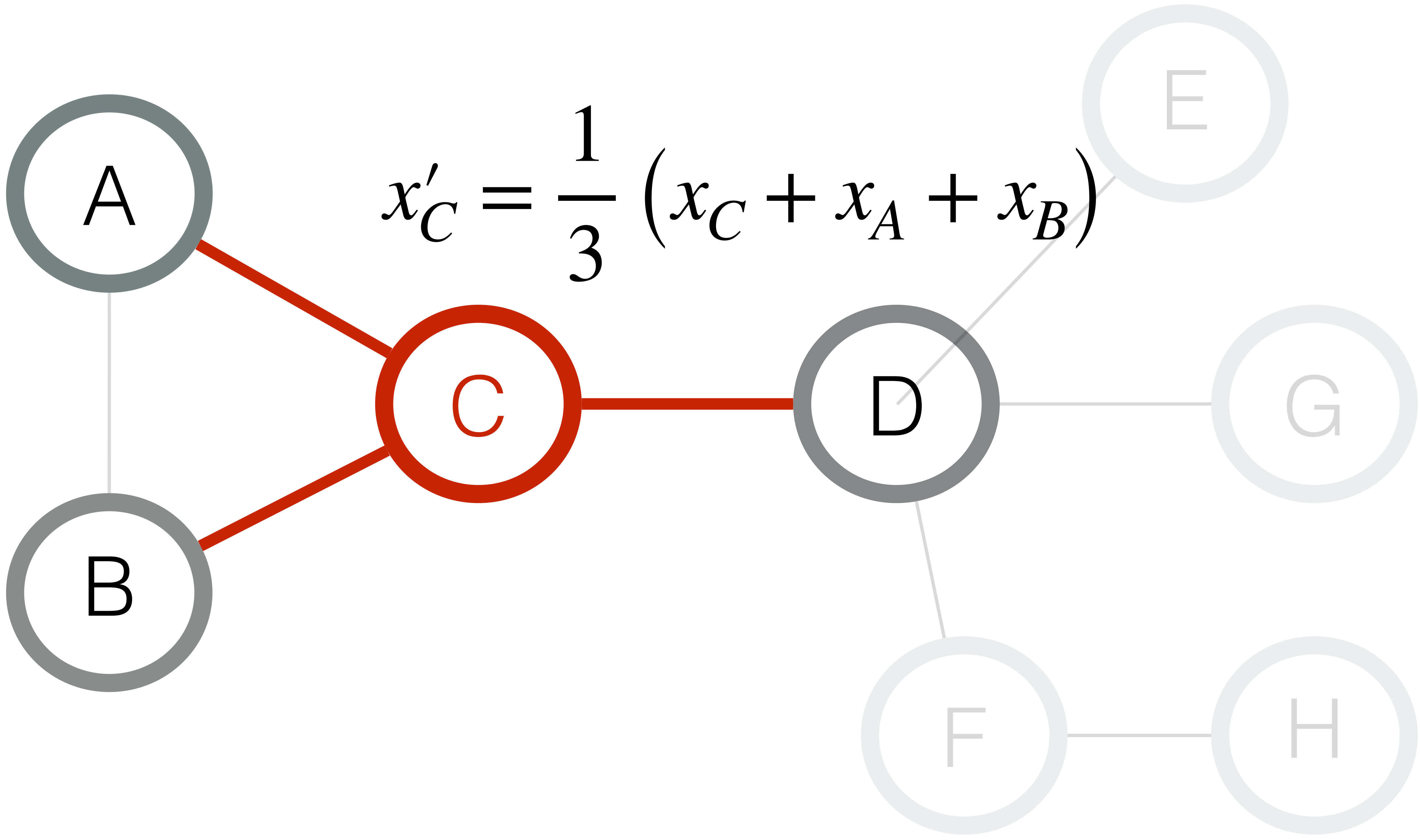
Node B gets messages
from nodes A and C



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

Spatial graph convolution of the data

Node B gets messages
from nodes A, B and C



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

Spatial graph convolution of the data

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

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

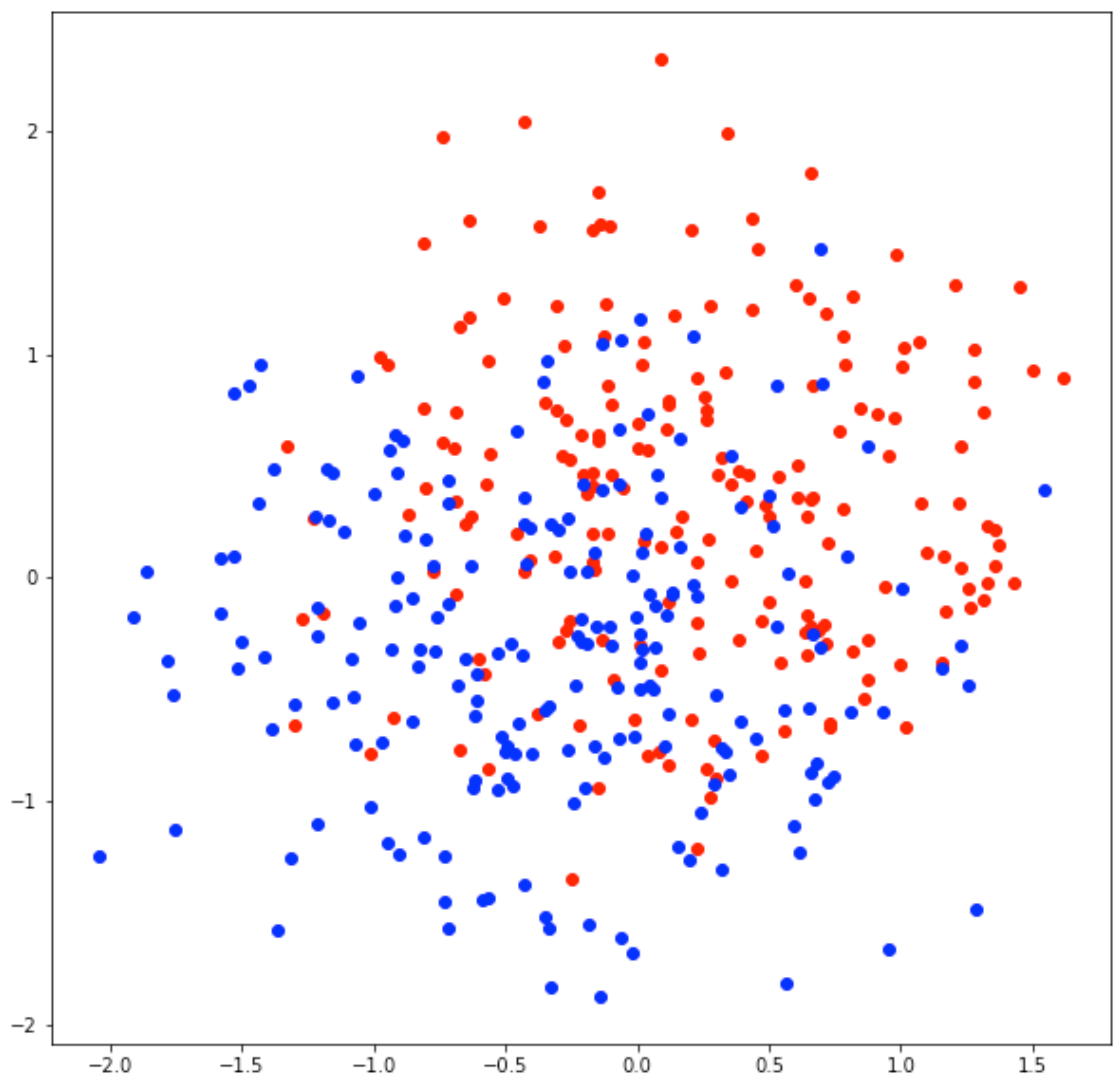
Spatial graph convolution of the data

$$X' := D^{-\frac{1}{2}} A D^{-\frac{1}{2}} X$$

We want to answer the following for linear classifiers

- Does graph convolution of the data help generalization?
- Which graphs are good graphs and which are not?
- What if we test on a graph that comes from a distribution with different parameters?

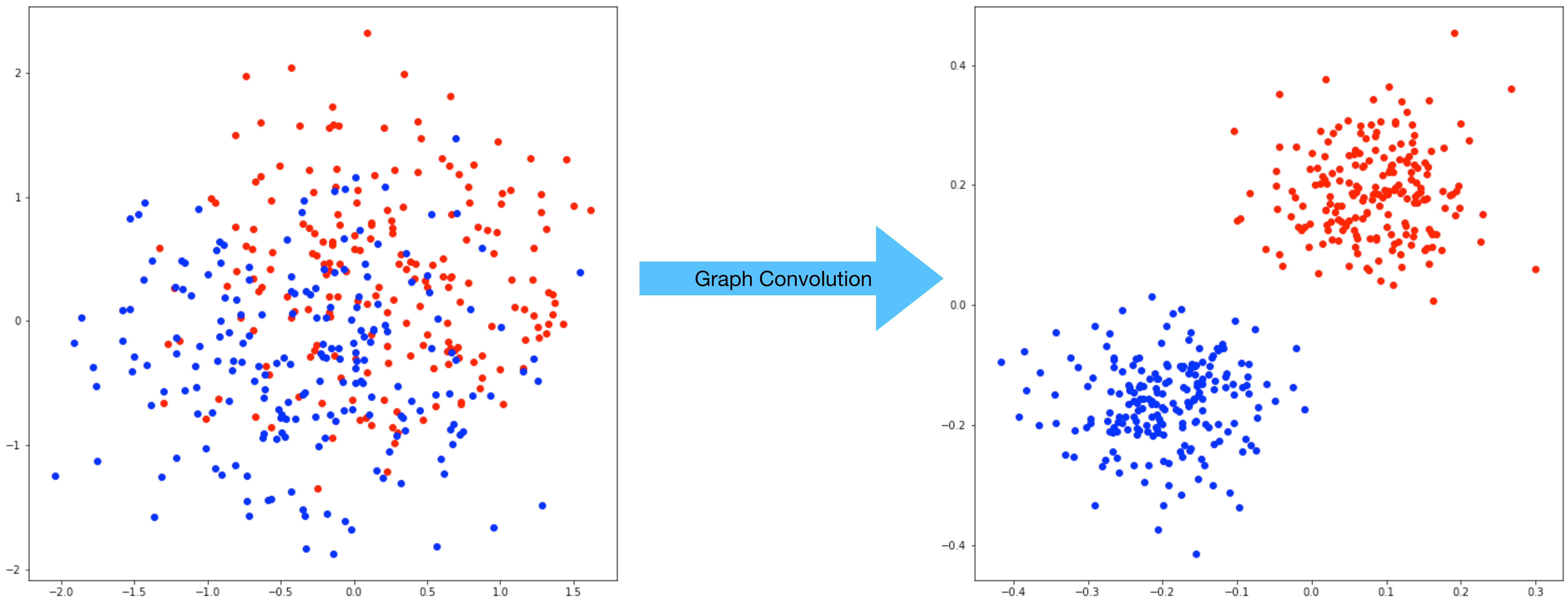
What can graph convolution do?



Original Data

- Consider distributions with 2D features
- We cannot separate this data linearly
- Can graph convolution help?

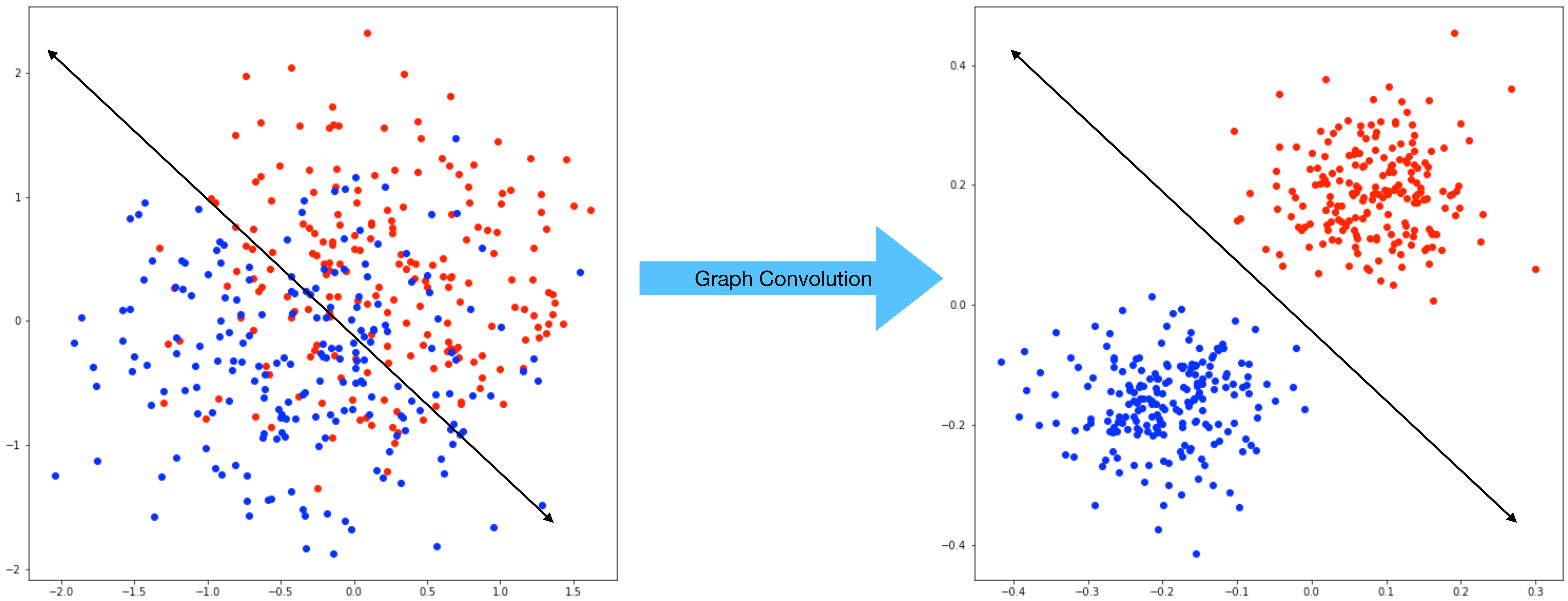
What can graph convolution do?



Original Data

After graph convolution

What can graph convolution do?



Original Data

After graph convolution

Graph convolution makes the data linearly separable

Model

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

- Two classes C_0, C_1
  n data points with features $(X_i)_{i=1}^n \in \mathbb{R}^d$

- $X_i \sim \mathcal{N}(\mu, \sigma^2 I)$ if $i \in C_0$
  $X_i \sim \mathcal{N}(\nu, \sigma^2 I)$ if $i \in C_1$

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

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

Graph convolution improves linear separability

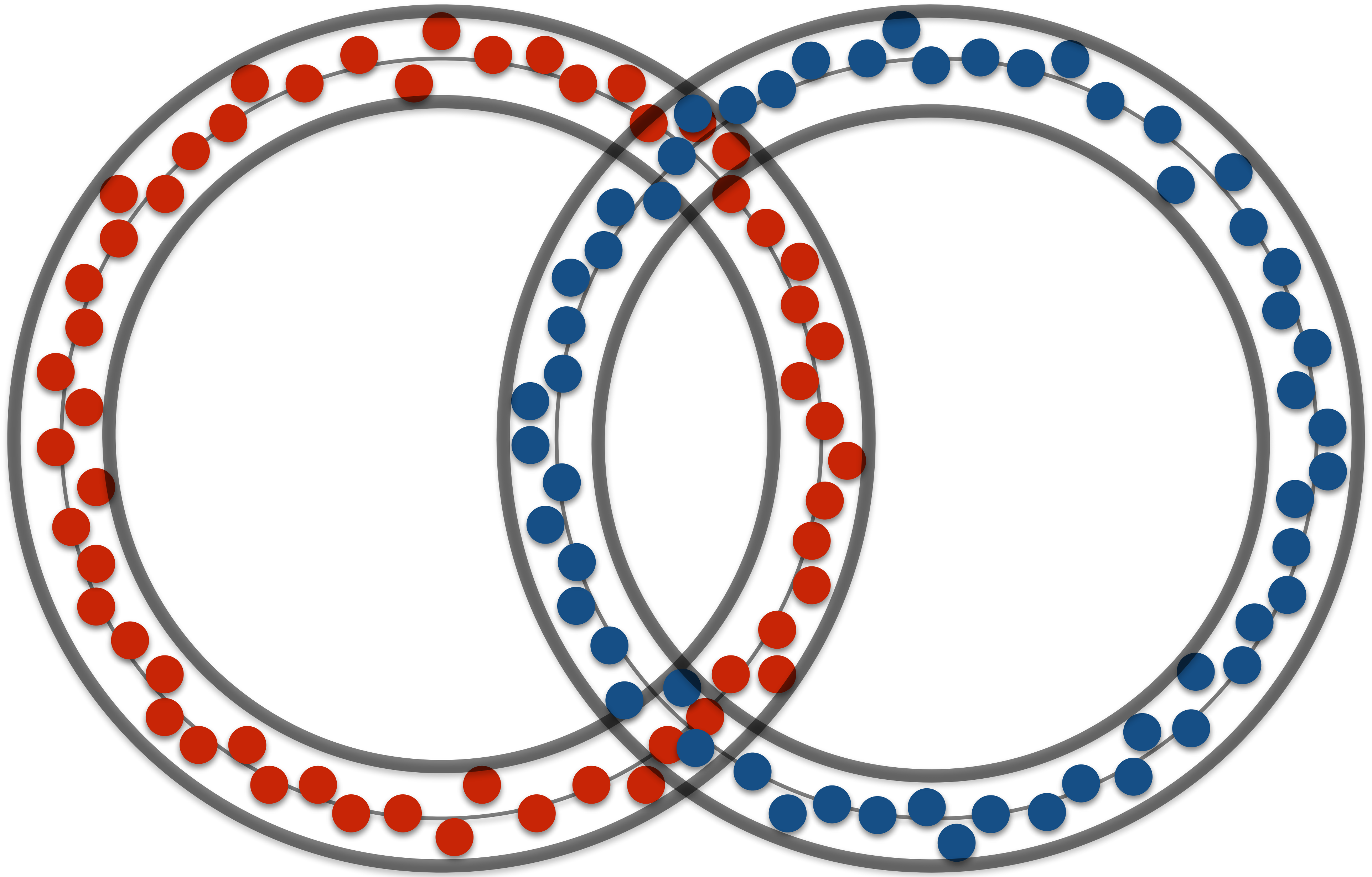
- Without the graph, no hyperplane can separate a binary GMM if means are $\mathcal{O}(\sigma)$ apart, i.e., $\|\mu - \nu\|_2 = \mathcal{O}(\sigma)$
- With graph convolution, this threshold changes to

$$\|\mu - \nu\| = \mathcal{O}\left(\frac{\sigma}{\sqrt{\mathbb{E}[D]}}\right)$$

Expected degree of a node

Proof sketch for no-convolution

Intuitive representation of Gaussian data in high dimensions



$\|\mu - \nu\|_2 = \mathcal{O}(\sigma)$

- We can actually show that there will be a constant fraction of misclassified data

Proof sketch for graph convolution

- After graph convolution the means move closer by a factor $\Gamma(p, q) = \dfrac{p - q}{p + q}$
- But the variance is reduced by $\mathbb{E}[D] = \mathcal{O}(n(p + q))$
- Thus the separability threshold changes from $\|\mu - \nu\|_2 = \mathcal{O}(\sigma)$ to

$$\|\mu - \nu\| = \mathcal{O}\left(\frac{\sigma}{\sqrt{\mathbb{E}[D]}}\right)$$

- Then we can show that the hyperplane that passes through the mid-point of the two means separates the data with high probability.

Assumptions

- *Assumption 1*: the graph is not too sparse $p = \omega(\log^2(n)/n)$
- *Assumption 2*: there exists notable difference between the amount of edges within a class, as opposed to between classes $\Gamma(p,q) = \frac{p-q}{p+q} = \Omega(1)$
- *Assumption 3*: Let d be the number of features. If $d \to \infty$ and $n \to \infty$, then $\omega(d \log d) \leq n \leq \mathcal{O}(poly(d))$. If d is fixed then n can grow with any rate.

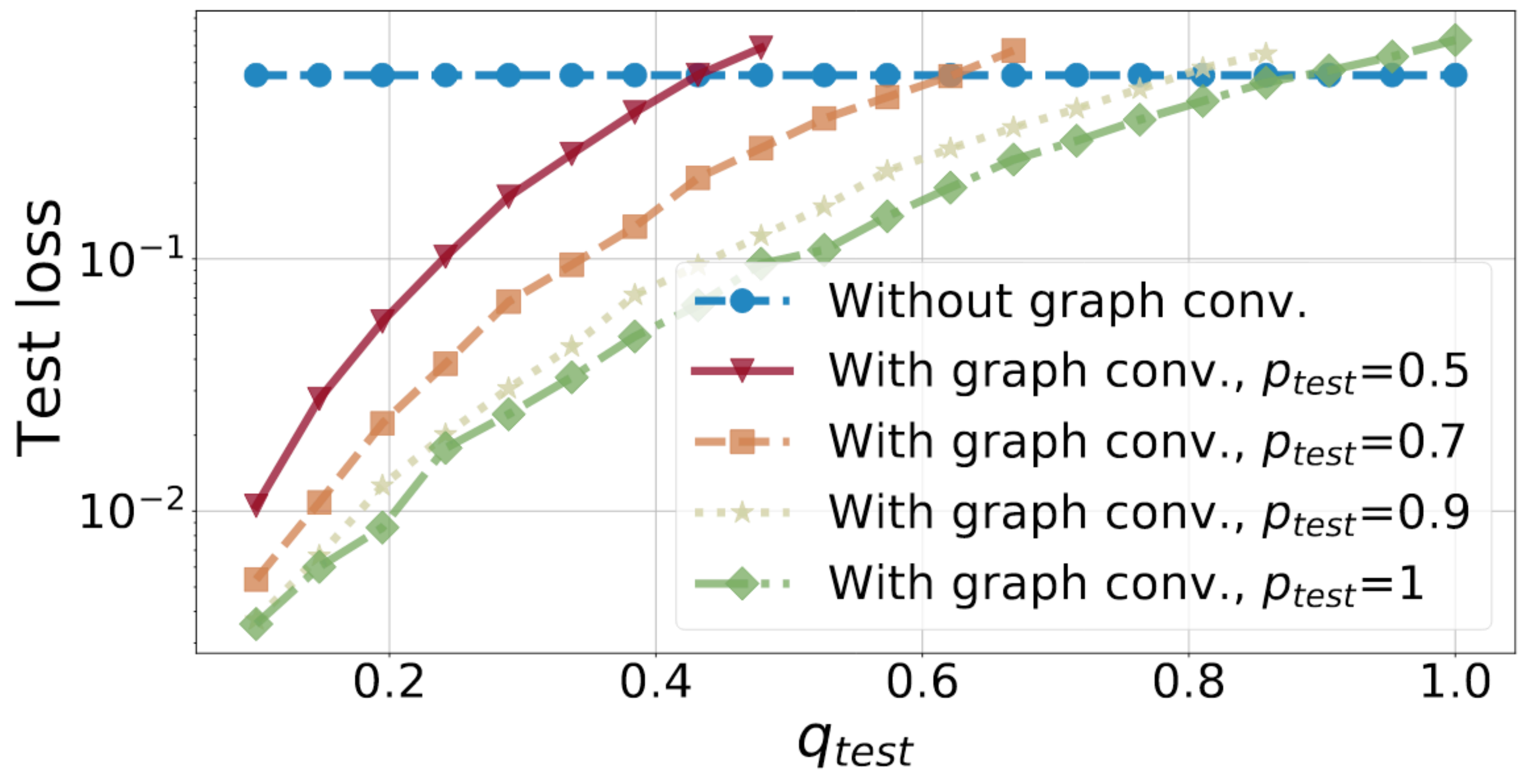
Bounds on training loss

- We use binary cross entropy loss to learn the classifier
- Without graph convolution, if $||\mu - \nu||_2 = K\sigma$,
  then the loss is lower bounded by $(2\log 2)\Phi(-K/2)$
- In the regime where the convolved data is separable,
  the loss decays exponentially
  $Loss(A, X) \leq C\exp\left(-d||\mu - \nu||\Gamma(p, q)\right)$,
  where $\Gamma(p, q) = \dfrac{p - q}{p + q}$

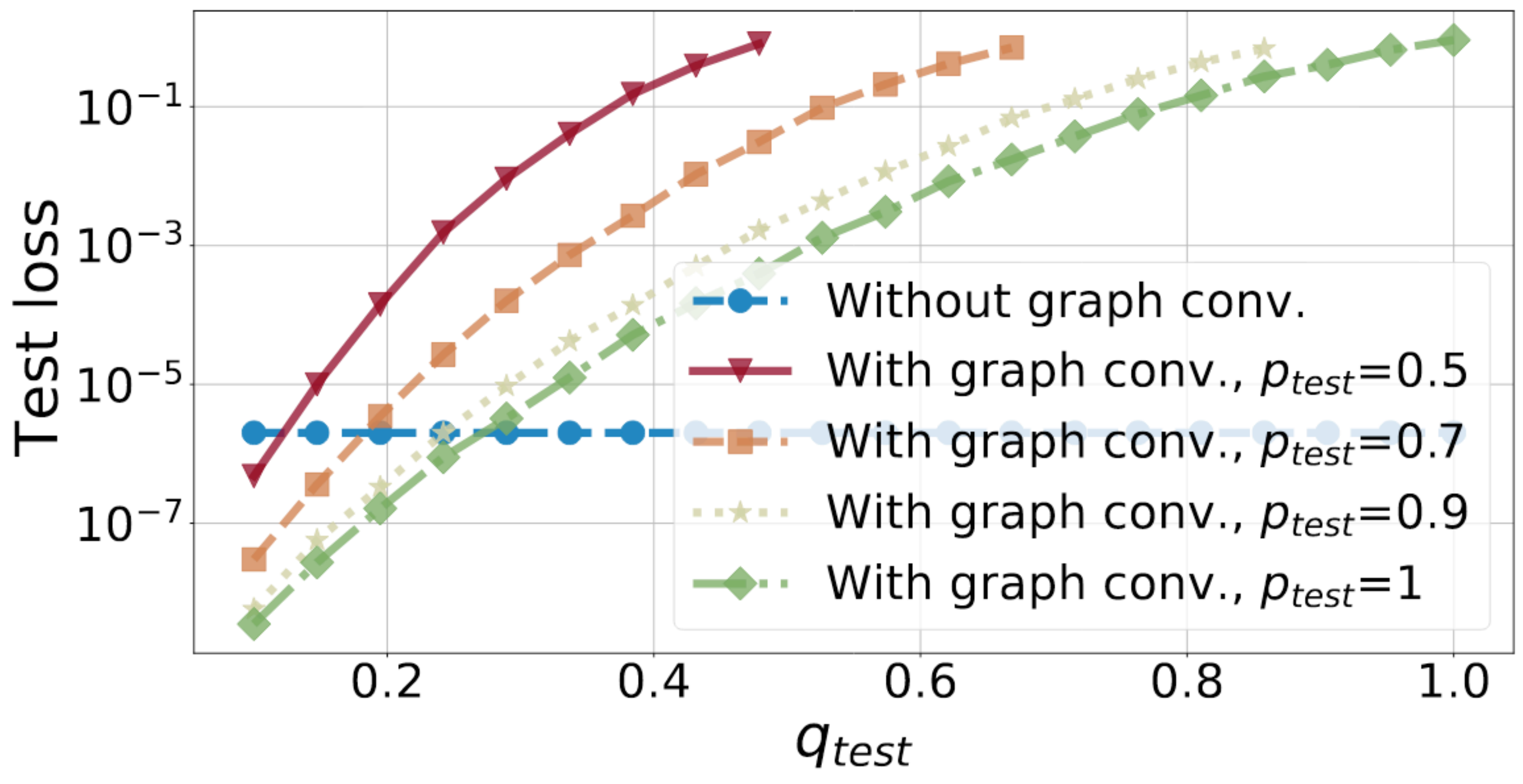
Generalization

- If the graph is not sparse, then for any new dataset A, X with different n, p, q, the test loss is bounded above
$$Loss(A, X) \leq C \exp\left(-d\|\mu - \nu\|\Gamma(p, q)\right)$$

- Loss increases with inter-class edge probability (noisy graph)



(a) $\|\mu_d - \nu_d\| = 2\sigma$



(b) $\|\mu_d - \nu_d\| = 16\sigma$

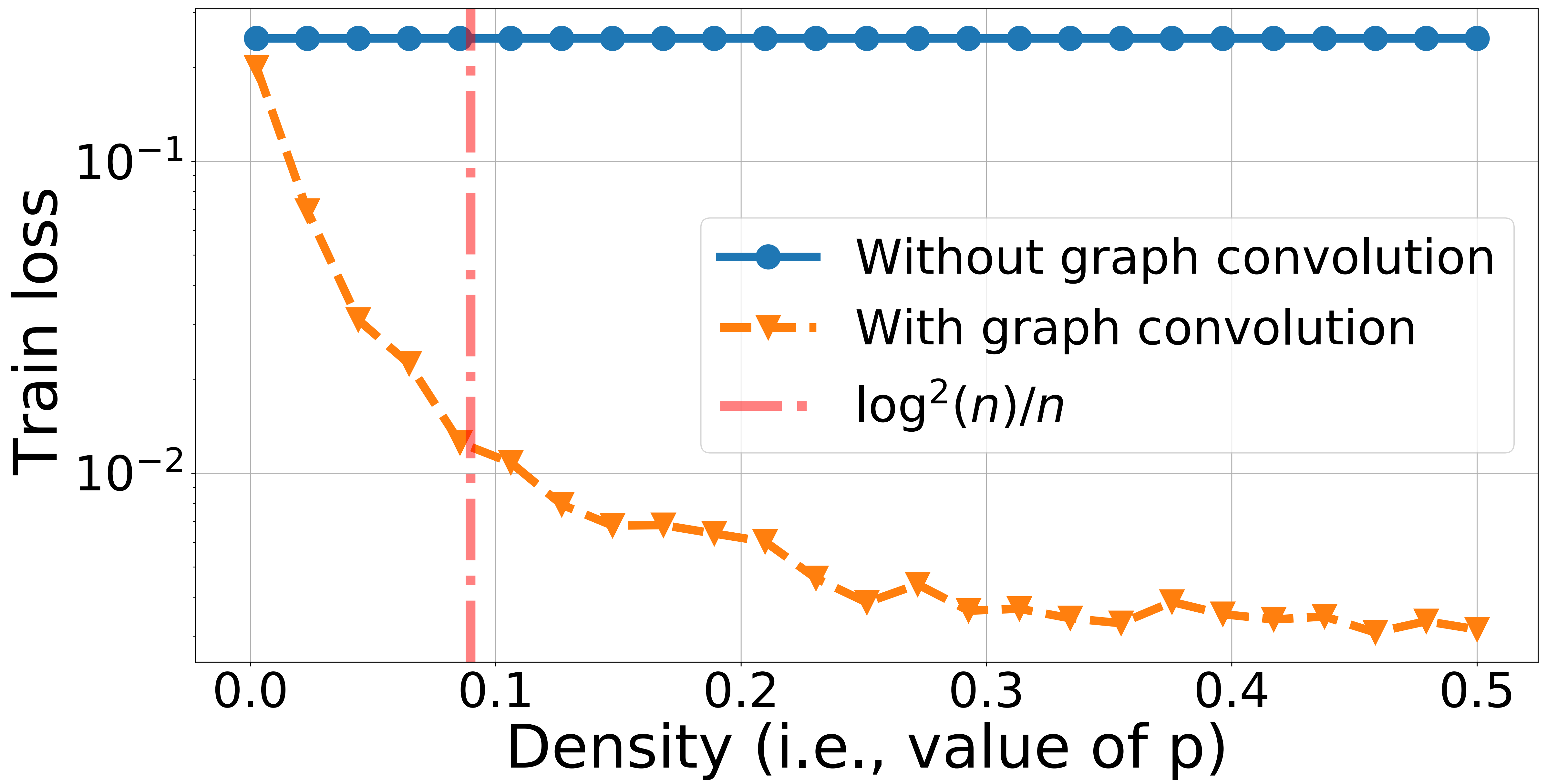
Proof sketch

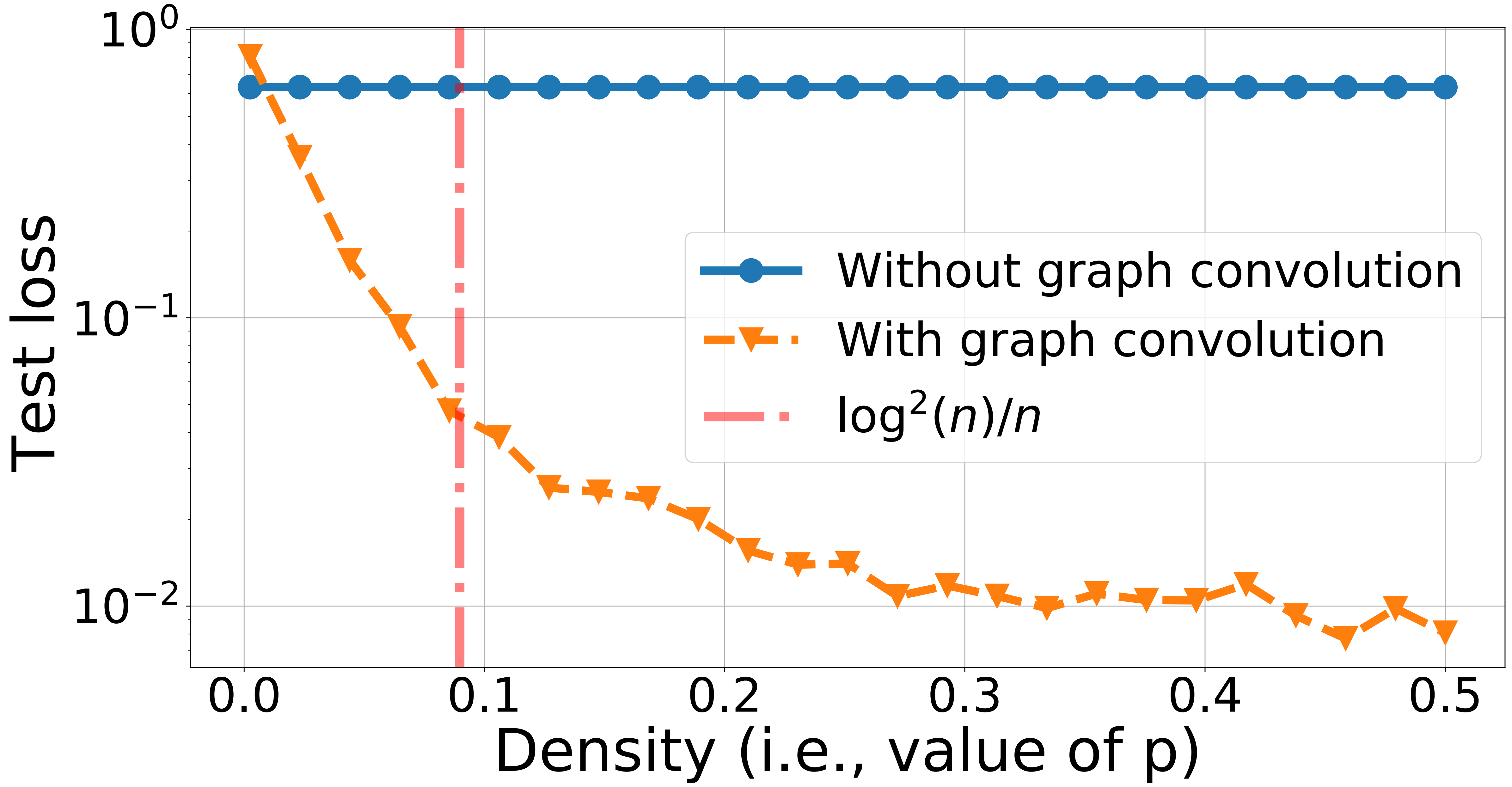
- Take the vector and the intercept that defines the hyperplane that passes through the midpoint of the two means: $\tilde{w} \propto \mu - \nu$ and $\tilde{b} = \langle \mu + \nu, \tilde{w} \rangle / 2$

- We show that $\tilde{w}, \tilde{b}$ is an approximately good minimizer for the cross entropy training loss as n increases, and the test loss goes exponentially to zero.

- Then using the fact that $\tilde{w}, \tilde{b}$ do not depend on p, q, implies that out-of-distribution generalization for the graph.

What about sparse graphs?





- Degrees can be $\mathcal{O}(1)$, thus no variance reduction!
- In message passing terminology, this simply means that messages do not propagate to a lot of neighbors within each class.

What about powers of $D^{-1}A$?

- Our analysis extends to $\left(D^{-1}A\right)^k$
- The distance between the means reduces by a factor $\Gamma(p,q)^k$
- The variance reduces by a factor $1/(\mathbb{E}[D])^k$
- For powers of convolution to be useful we need, for example, $\Gamma(p,q) = \Omega(1)$

What’s next?

- Is graph convolution helpful in the following settings:
  - Multiple classes — how does the distance between means matter here?
  - If the data is non-linearly separable — what does the optimal classifier look like?
- What about generalization in the above scenarios?
- Analysis of deeper GCNs

Thank you!