Mixture of Weak and Strong Experts on Graphs

Hanqing Zeng*
Meta AI
zengh@meta.com

Hanjia Lyu*
University of Rochester
hlyu5@ur.rochester.edu

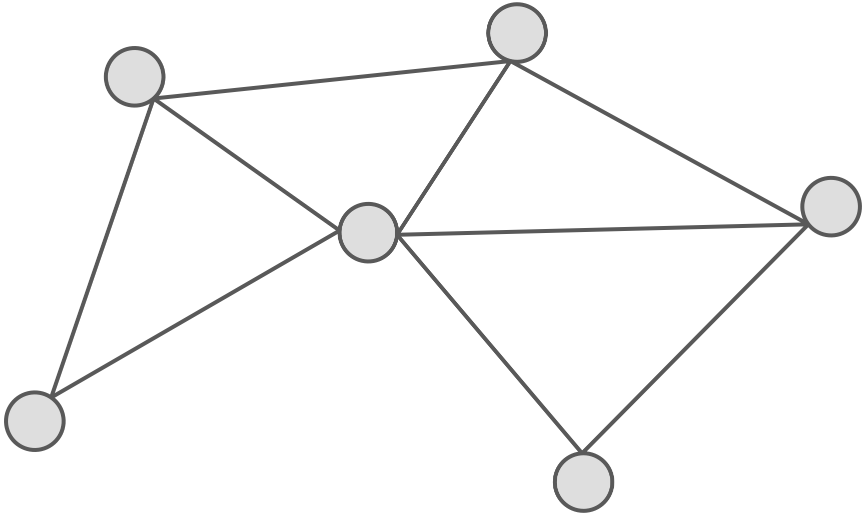
Diyi Hu
University of Southern California
diyihu@usc.edu

Yinglong Xia
Meta AI
yxia@meta.com

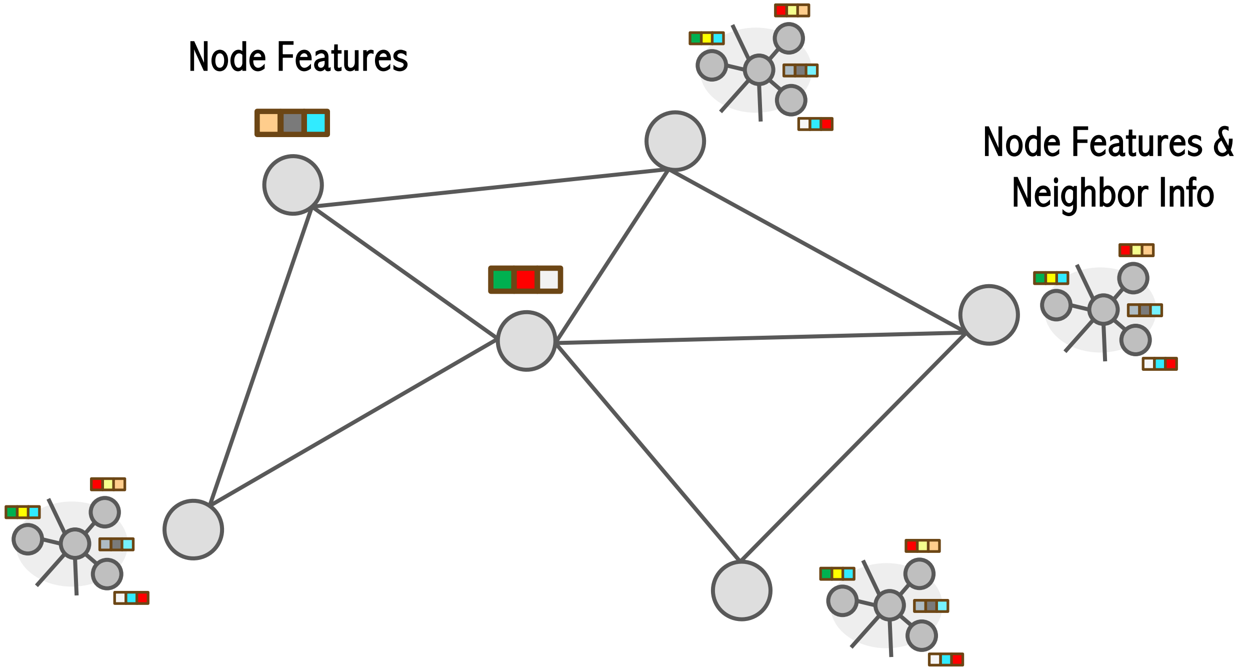
Jiebo Luo
University of Rochester
jluo@cs.rochester.edu

* Equal contribution.

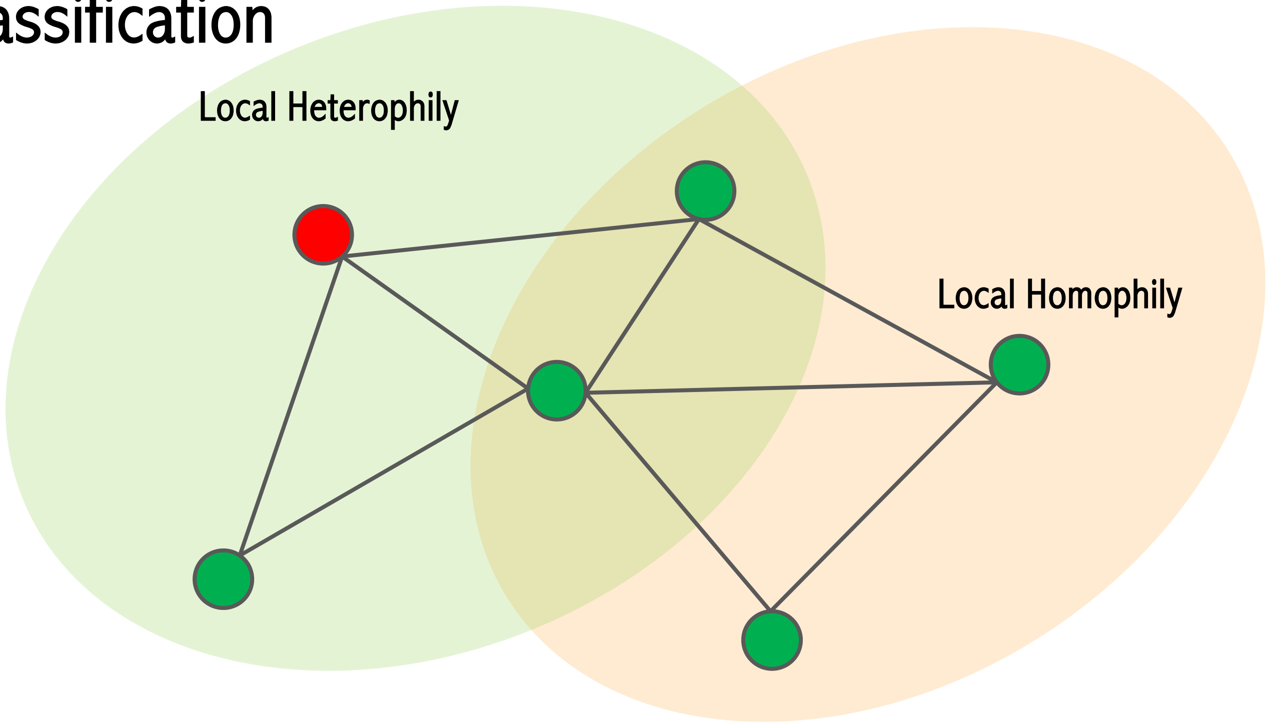
Node Classification

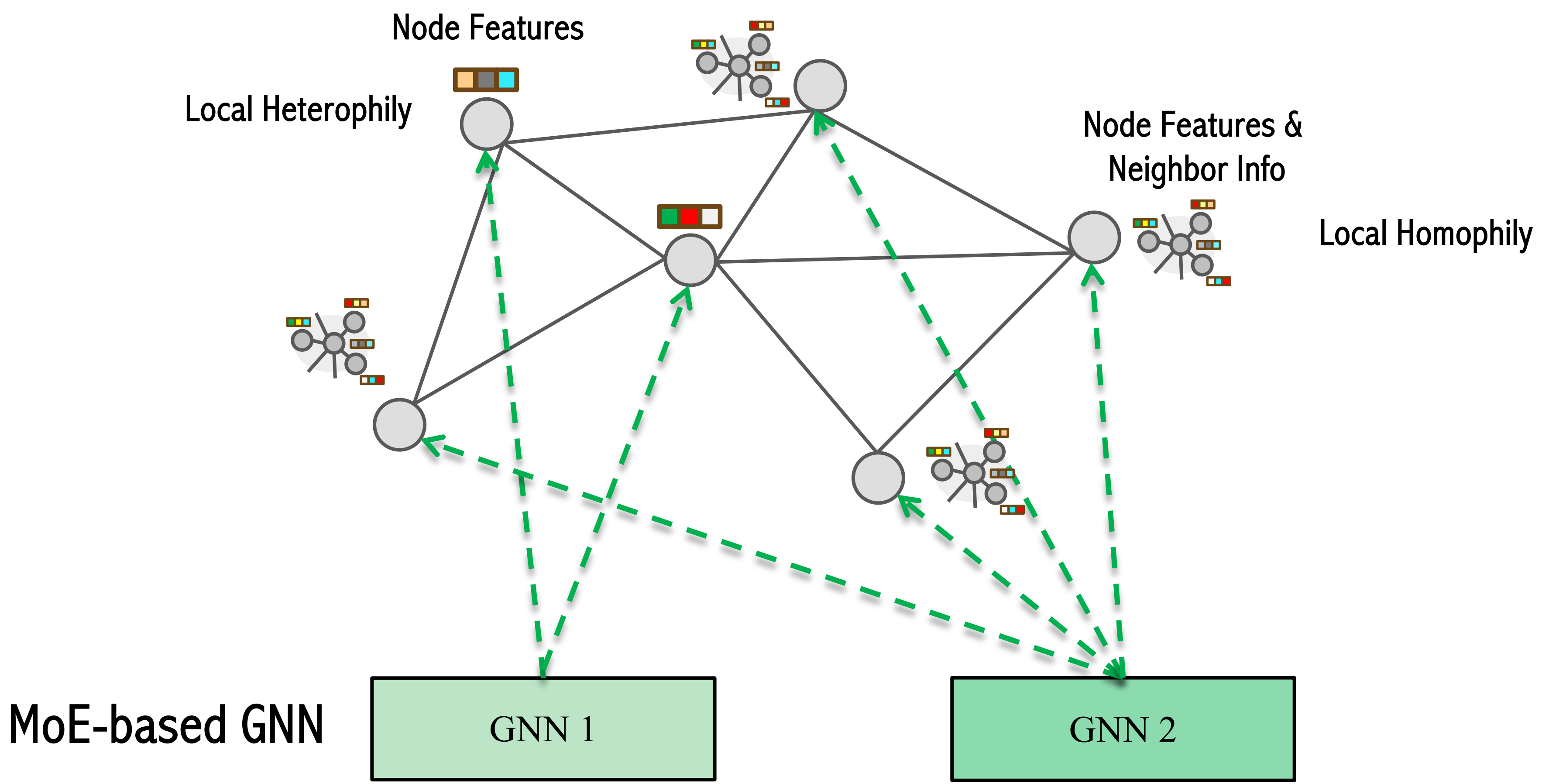


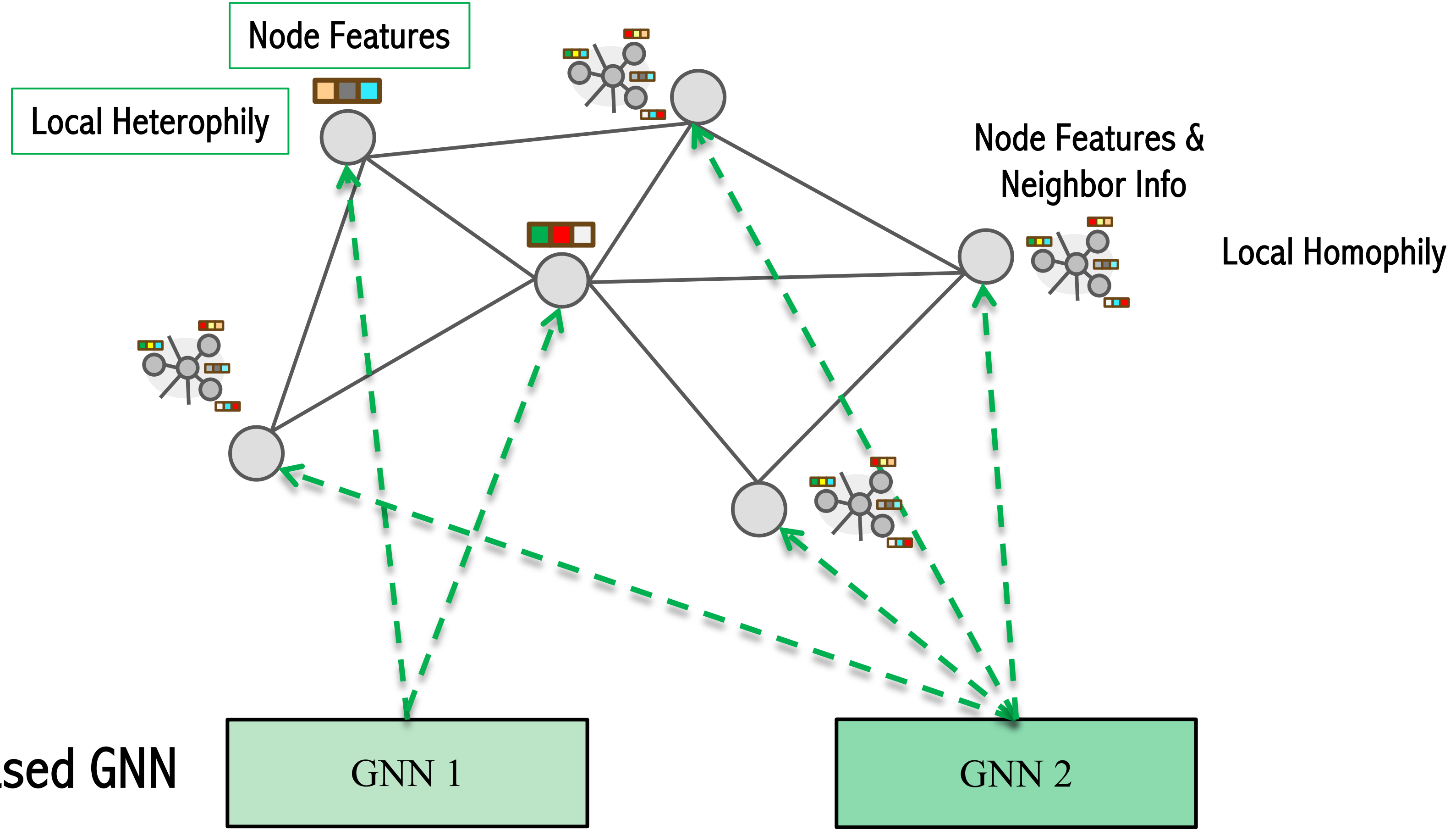
Node Classification



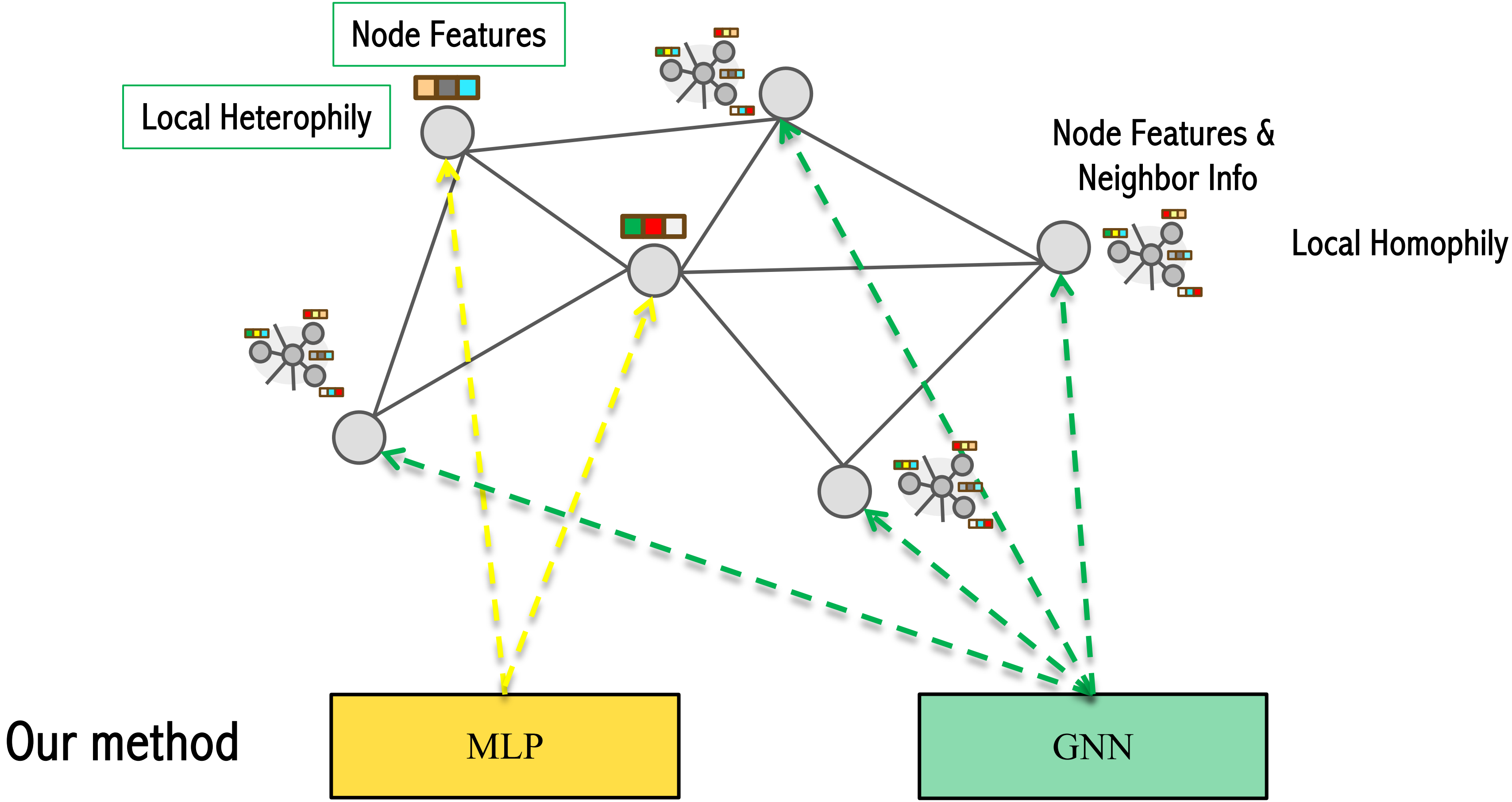
Node Classification

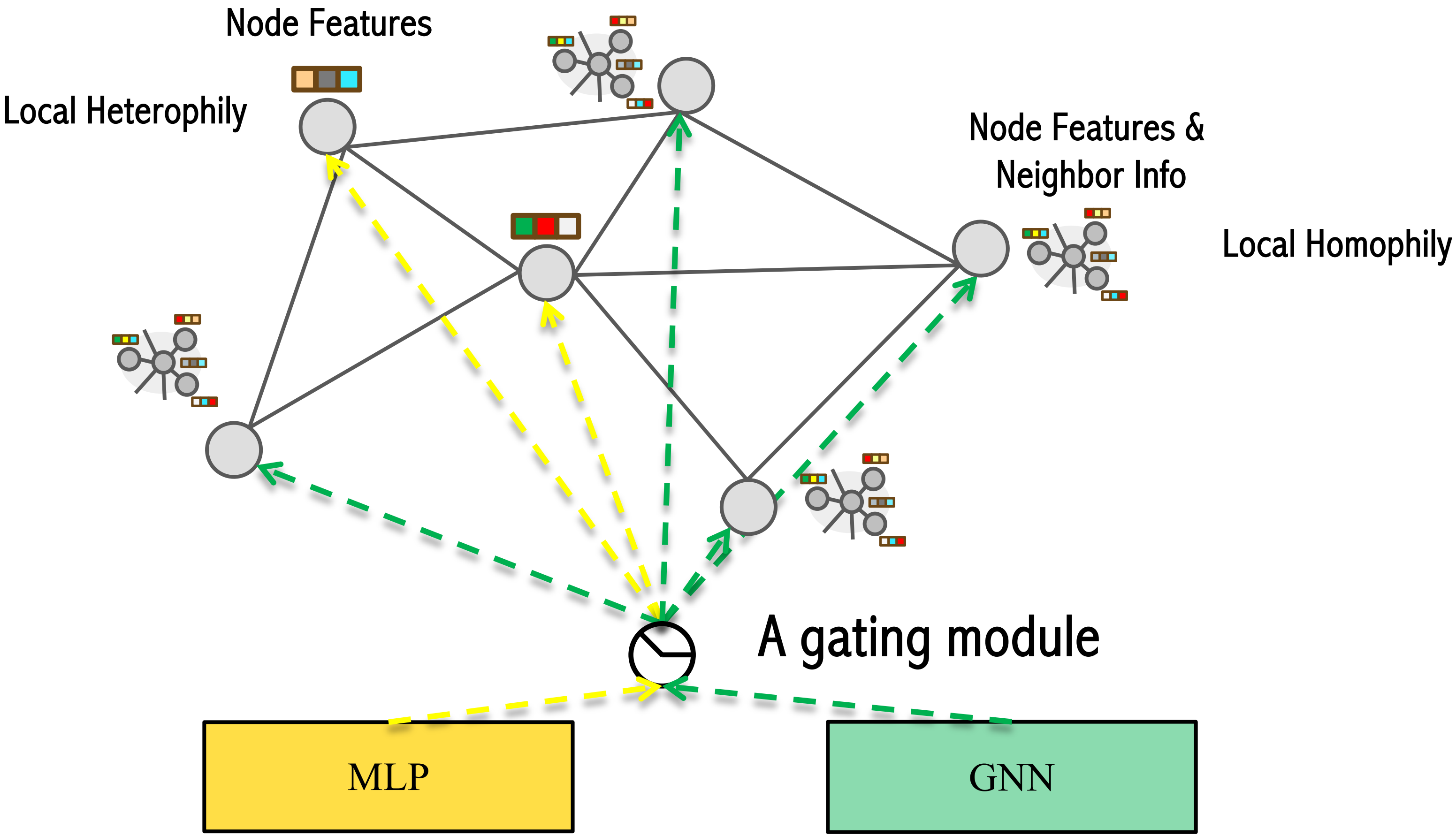


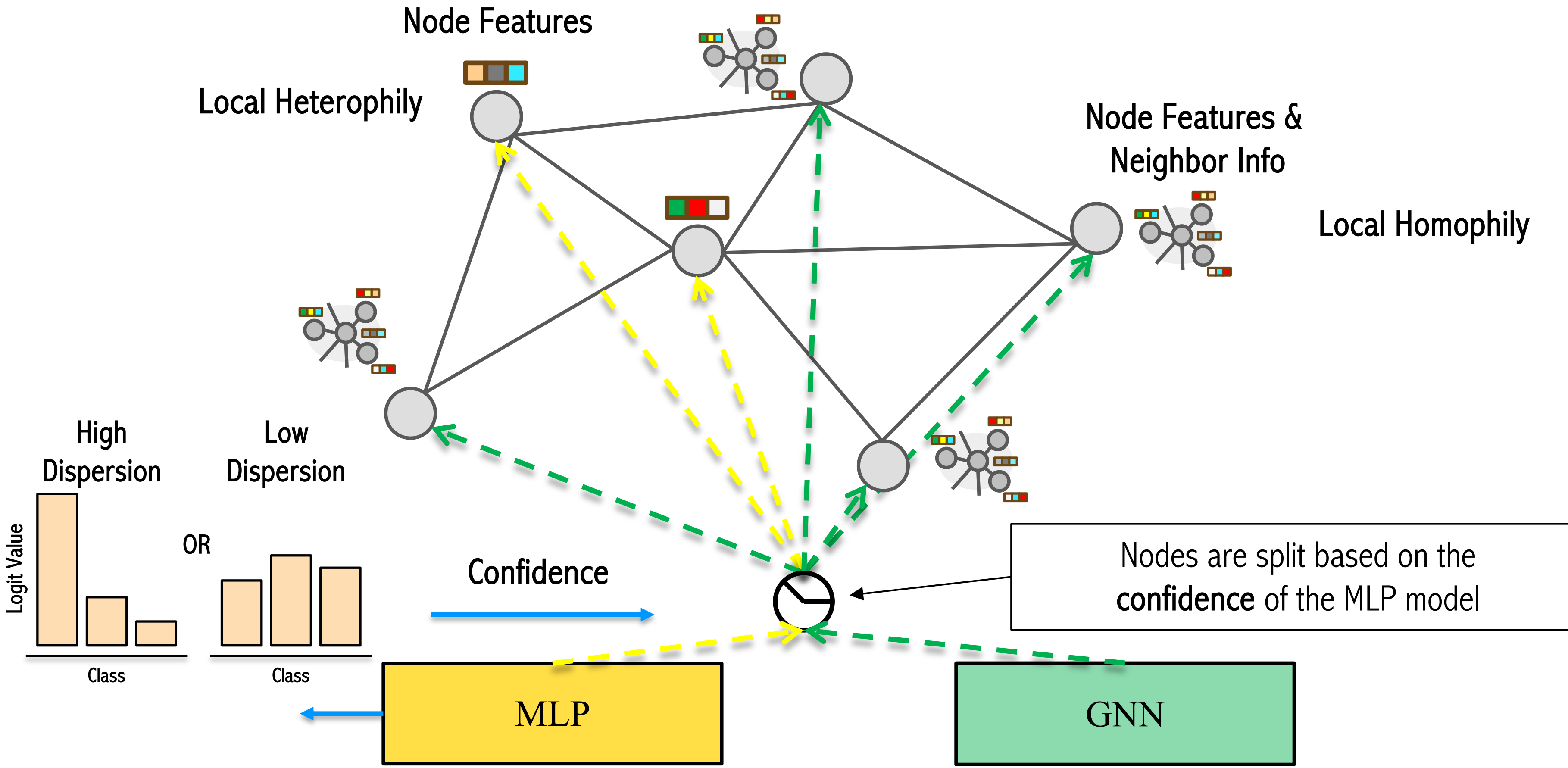

Node Features
Local Heterophily
Node Features &
Neighbor Info
Local Homophily
MoE-based GNN
GNN 1
GNN 2



MoE-based GNN

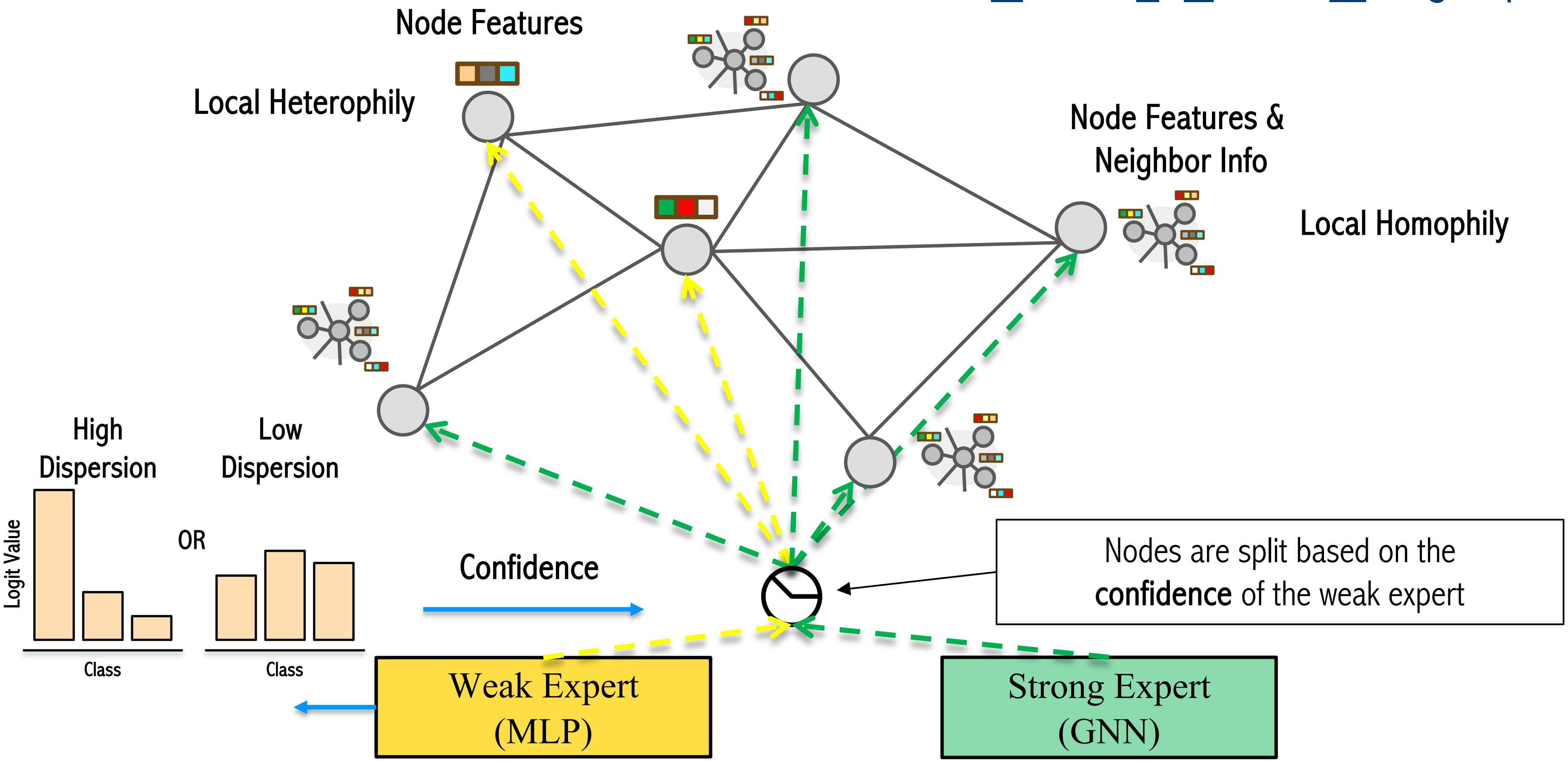

Node Features
Local Heterophily
Node Features &
Neighbor Info
Local Homophily
Our method
MLP
GNN


Node Features
Local Heterophily
Node Features &
Neighbor Info
Local Homophily
A gating module
MLP
GNN



Nodes are split based on the **confidence** of the MLP model

Mowst: Mixture of Weak & Strong Experts



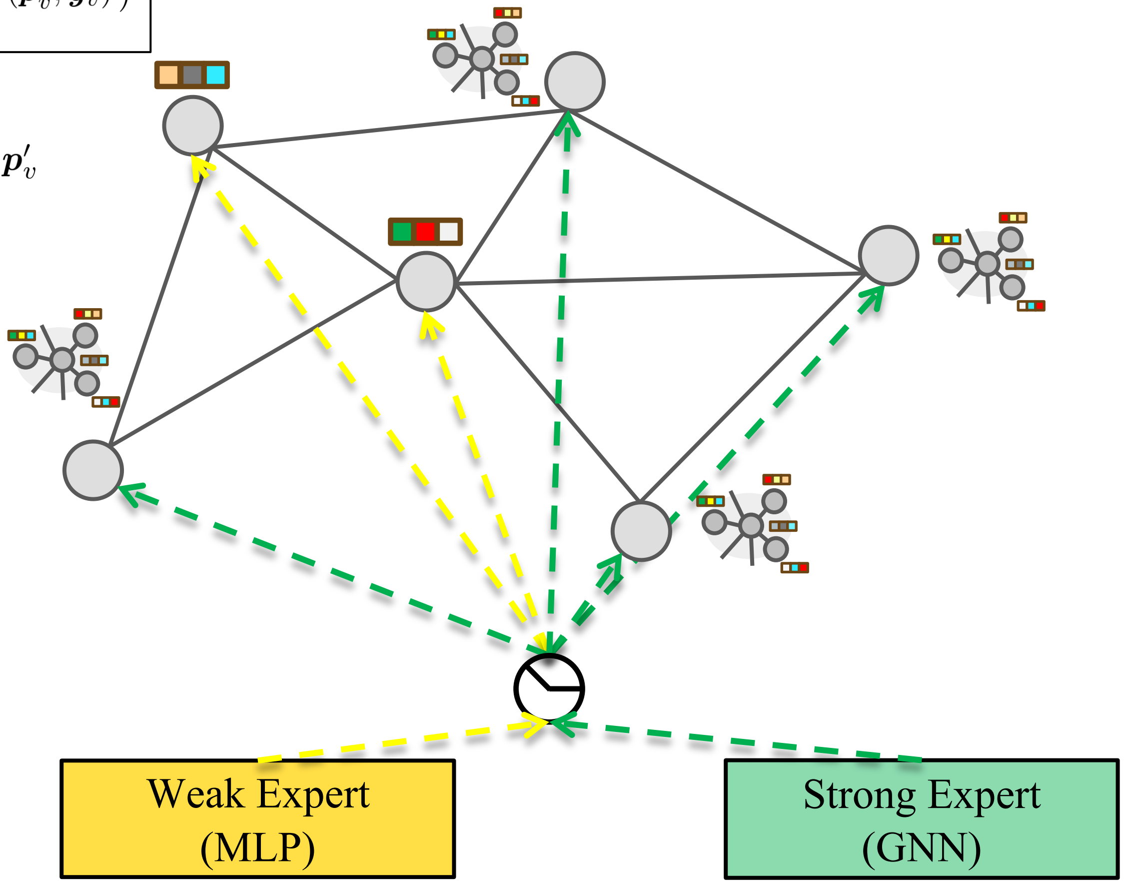
Mowst: Mixture of Weak & Strong Experts

$$L_{\texttt{Mowst}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left(C\left(\boldsymbol{p}_v\right) \cdot L\left(\boldsymbol{p}_v, \boldsymbol{y}_v\right) + \left(1 - C\left(\boldsymbol{p}_v\right)\right) \cdot L\left(\boldsymbol{p}'_v, \boldsymbol{y}_v\right) \right)$$

Target node: v

MLP's prediction: $\boldsymbol{p}_v$ GNN's prediction: $\boldsymbol{p}'_v$

How confident is MLP: $C\left(\boldsymbol{p}_v\right)$



Mowst: Mixture of Weak & Strong Experts

$$L_{\texttt{Mowst}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left(C\left(\boldsymbol{p}_v\right) \cdot L\left(\boldsymbol{p}_v, \boldsymbol{y}_v\right) + \left(1 - C\left(\boldsymbol{p}_v\right)\right) \cdot L\left(\boldsymbol{p}'_v, \boldsymbol{y}_v\right) \right)$$

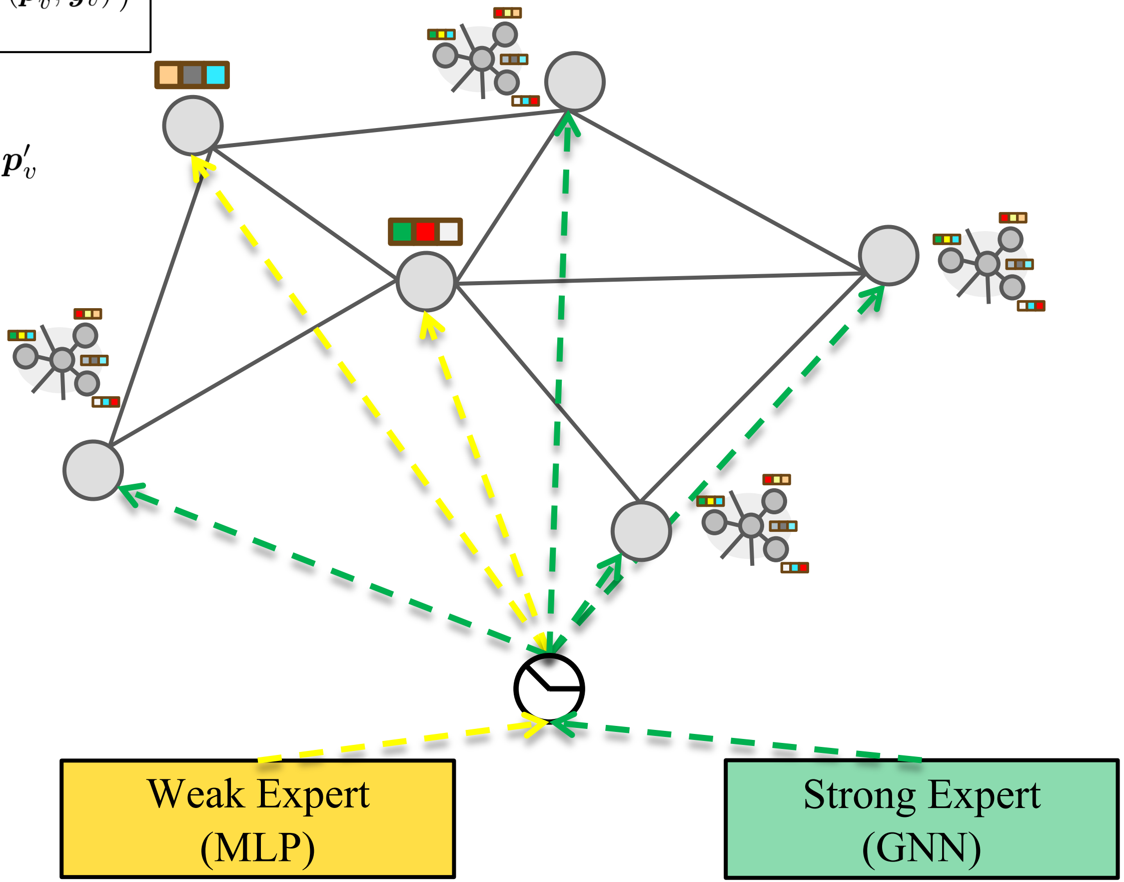
Target node: v

MLP's prediction: $\boldsymbol{p}_v$ GNN's prediction: $\boldsymbol{p}'_v$

How confident is MLP: $C\left(\boldsymbol{p}_v\right)$

Case 1:

If the node's self features are sufficient



$$L_{\texttt{Mowst}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left(C(\boldsymbol{p}_v) \cdot L(\boldsymbol{p}_v, \boldsymbol{y}_v) + (1 - C(\boldsymbol{p}_v)) \cdot L(\boldsymbol{p}'_v, \boldsymbol{y}_v) \right)$$

Mowst: Mixture of Weak & Strong Experts

Target node: v

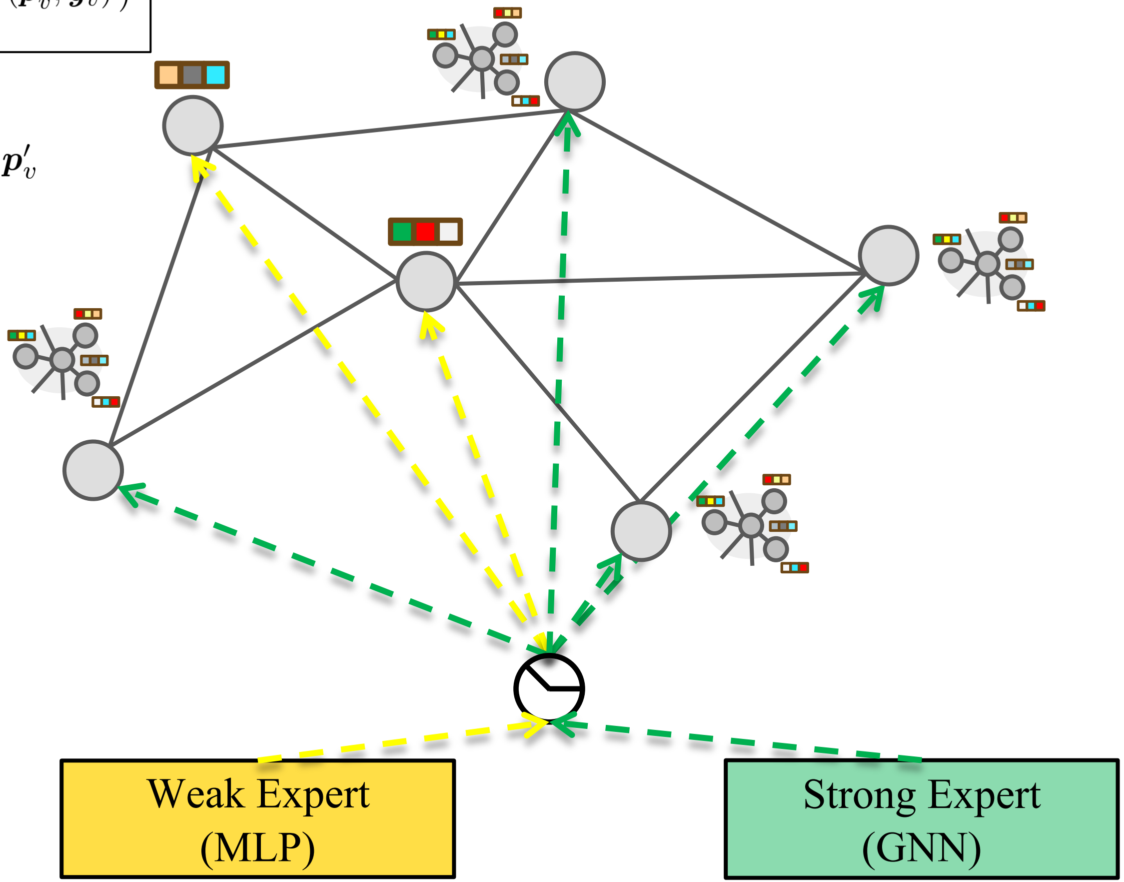
MLP's prediction: $\boldsymbol{p}_v$ GNN's prediction: $\boldsymbol{p}'_v$

How confident is MLP: $C(\boldsymbol{p}_v)$

Case 2:

If the node's self features are insufficient

- MLP is certain
- MLP is not certain



Mowst: Mixture of Weak & Strong Experts

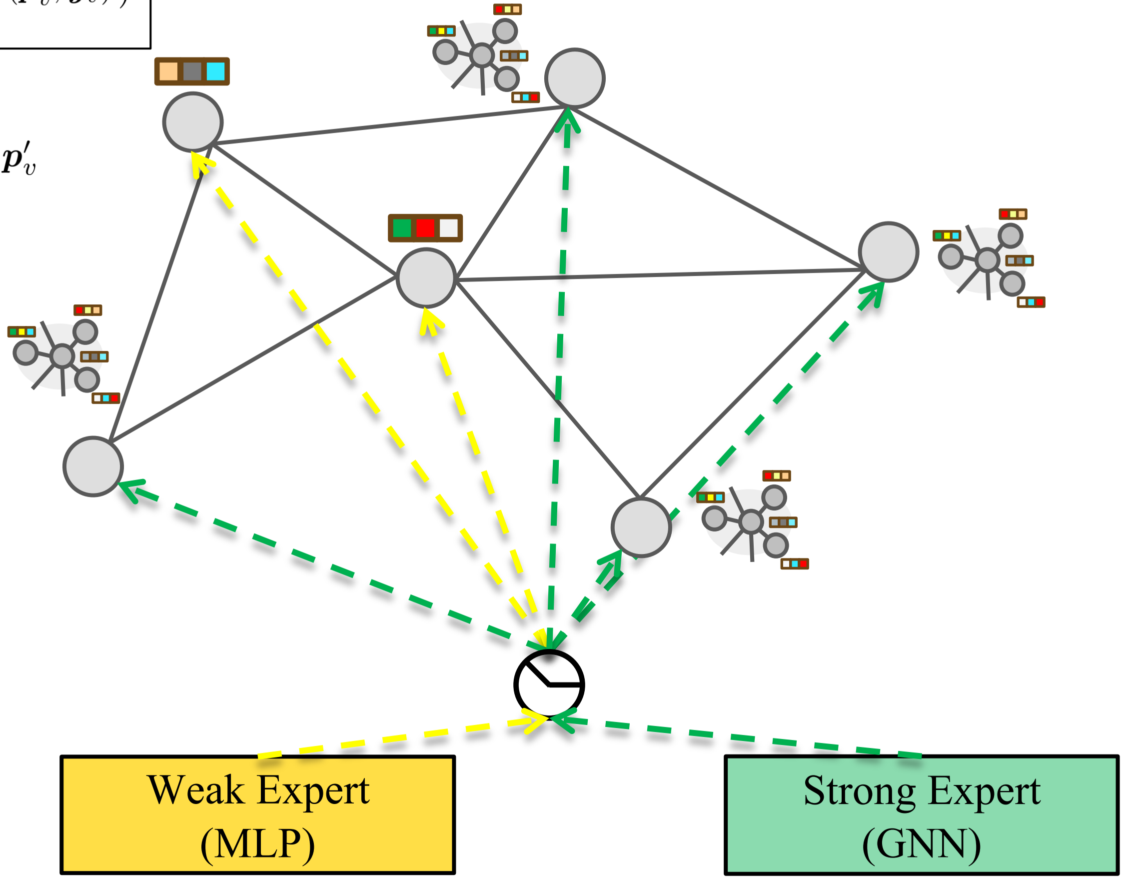
$$L_{\texttt{Mowst}} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left(C\left(\boldsymbol{p}_v\right)\cdot L\left(\boldsymbol{p}_v, \boldsymbol{y}_v\right) + \left(1 - C\left(\boldsymbol{p}_v\right)\right)\cdot L\left(\boldsymbol{p}'_v, \boldsymbol{y}_v\right)\right)$$

Target node: v

MLP's prediction: $\boldsymbol{p}_v$ GNN's prediction: $\boldsymbol{p}'_v$

How confident is MLP: $C\left(\boldsymbol{p}_v\right)$

- Mowst is at least as expressive as the MLP or GNN expert alone



Mowst: Mixture of Weak & Strong Experts

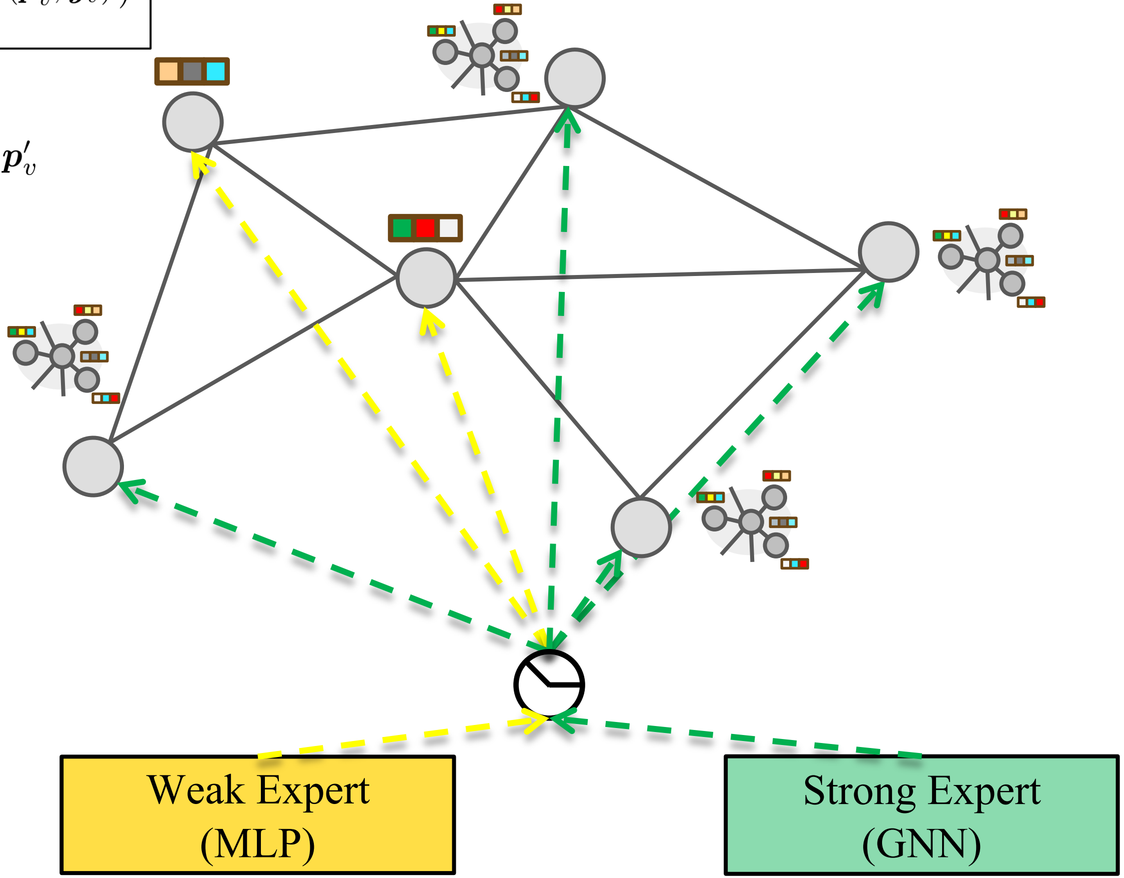
$$L_{\texttt{Mowst}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left(C\left(\boldsymbol{p}_v\right) \cdot L\left(\boldsymbol{p}_v, \boldsymbol{y}_v\right) + \left(1 - C\left(\boldsymbol{p}_v\right)\right) \cdot L\left(\boldsymbol{p}'_v, \boldsymbol{y}_v\right) \right)$$

Target node: v

MLP's prediction: $\boldsymbol{p}_v$ GNN's prediction: $\boldsymbol{p}'_v$

How confident is MLP: $C\left(\boldsymbol{p}_v\right)$

- Mowst is at least as expressive as the MLP or GNN expert alone
- Mowst-GCN is more expressive than the GCN expert alone



Mowst: Mixture of Weak & Strong Experts

$$L_{\texttt{Mowst}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left(C\left(\boldsymbol{p}_v\right) \cdot L\left(\boldsymbol{p}_v, \boldsymbol{y}_v\right) + \left(1 - C\left(\boldsymbol{p}_v\right)\right) \cdot L\left(\boldsymbol{p}'_v, \boldsymbol{y}_v\right) \right)$$

Target node: v

MLP's prediction: $\boldsymbol{p}_v$ GNN's prediction: $\boldsymbol{p}'_v$

How confident is MLP: $C\left(\boldsymbol{p}_v\right)$

- Mowst is at least as expressive as the MLP or GNN expert alone
- Mowst-GCN is more expressive than the GCN expert alone
- The worst-case cost of Mowst-GCN is similar to that of a vanilla GCN

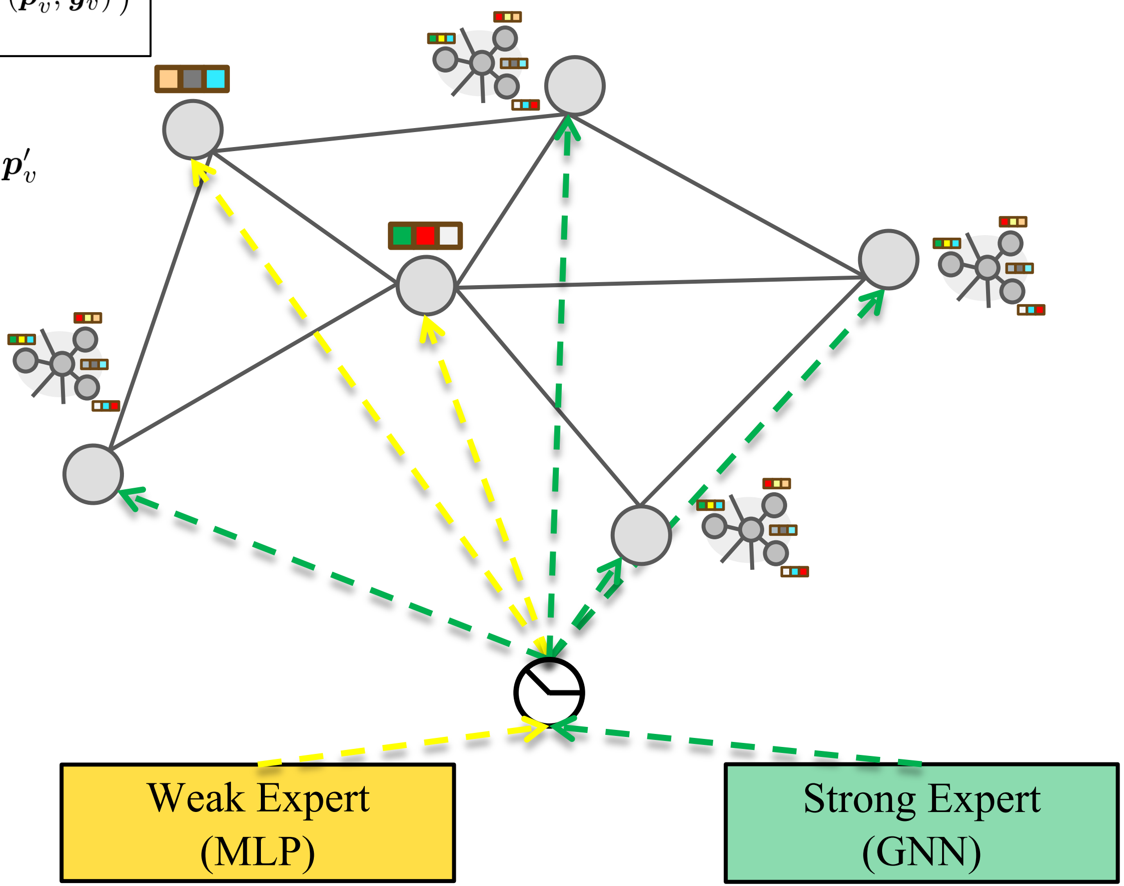
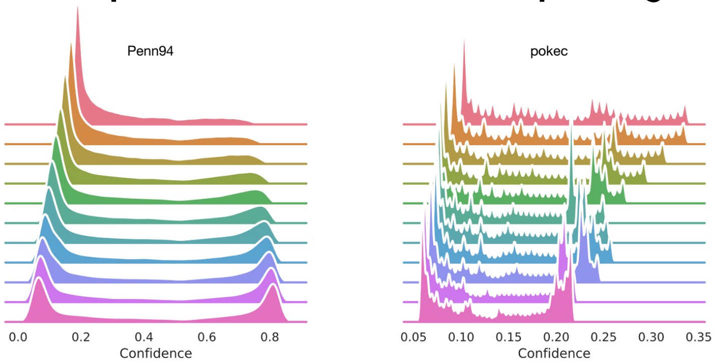


Table 1: Mowst outperforms baselines under the same number of layers and hidden dimension. Values with '†', '‡' and '††' are from Hu et al. (2020), Lim et al. (2021), and Wang et al. (2023). For each graph, we show the **best** and <u>second best</u> results, and **absolute gains** against the GNN counterparts (*e.g.*, Mowst(*)-GCN *vs.* GCN and GraphMoE-GCN). All results are averaged over 10 runs.

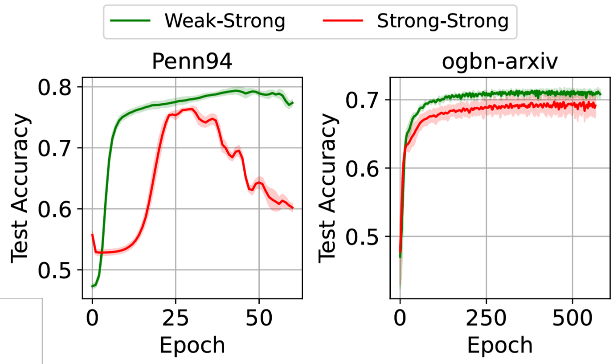
	Flickr	ogbn-products	ogbn-arxiv	Penn94	pokec	twitch-gamer
MLP	46.93 ±0.00	61.06† ±0.08	55.50† ±0.23	73.61‡ ±0.40	62.37‡ ±0.02	60.92‡ ±0.07
GAT	52.47 ±0.14	OOM	71.58 ±0.17	81.53‡ ±0.55	71.77‡ ±6.18	59.89‡ ±4.12
GPR-GNN	53.23 ±0.14	72.41 ±0.04	71.10 ±0.22	81.38‡ ±0.16	<u>78.83</u>‡ ±0.05	61.89‡ ±0.29
AdaGCN	48.96 ±0.06	69.06 ±0.04	58.45 ±0.50	74.42 ±0.58	55.92 ±0.35	61.02 ±0.14
GCN	53.86 ±0.37	75.64† ±0.21	71.74† ±0.29	82.17 ±0.04	76.01 ±0.49	62.42 ±0.53
Mowst(*)-GCN	<u>54.62</u> ±0.23 (+0.76)	76.49 ±0.22 (+0.85)	**72.52** ±0.07 (+0.64)	<u>83.19</u> ±0.43 (+1.02)	77.28 ±0.08 (+0.29)	63.74 ±0.23 (+0.83)
GIN	53.71 ±0.35	-	69.39 ±0.56	82.68 ±0.32	53.37 ±2.15	61.76 ±0.60
Mowst(*)-GIN	**55.48** ±0.32 (+1.77)	-	71.43 ±0.26 (+2.04)	**84.56** ±0.31 (+1.88)	76.11 ±0.39 (+22.74)	64.32 ±0.34 (+2.56)
GIN-skip	52.70 ±0.00	-	71.28 ±0.00	80.32 ±0.43	76.29 ±0.51	64.27 ±0.25
Mowst(*)-GIN-skip	53.19 ±0.31 (+0.49)	-	71.79 ±0.23 (+0.51)	81.20 ±0.55 (+0.88)	**79.70** ±0.23 (+3.41)	**64.91** ±0.22 (+0.64)
GraphSAGE	53.51 ±0.05	<u>78.50</u>† ±0.14	71.49† ±0.27	76.75 ±0.52	75.76 ±0.04	61.99 ±0.30
GraphMoE-SAGE	52.16 ±0.13	77.79 ±0.00	71.19 ±0.15	77.04 ±0.55	76.67 ±0.08	63.42 ±0.23
Mowst(*)-SAGE	53.90 ±0.18 (+0.39)	**79.38** ±0.44 (+0.88)	<u>72.04</u> ±0.24 (+0.55)	79.07 ±0.43 (+2.03)	77.84 ±0.04 (+1.33)	<u>64.38</u> ±0.14 (+1.05)

More details in our paper

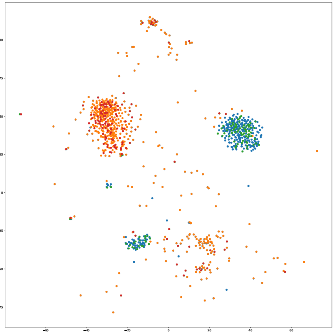
Specialization via data splitting



Weak-strong vs. Strong-strong

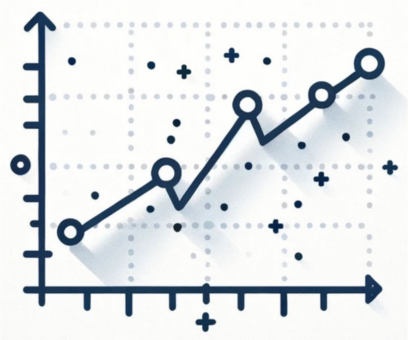


Denoised Fine-tuning



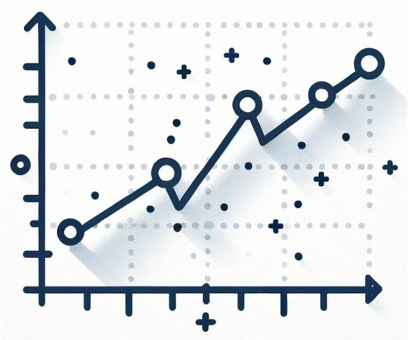
Future work

- More experts
- Other non-graph domains



Future work

- More experts
- Other non-graph domains



Thank you!